

# Agents of Diffusion: Enhancing Diffusion Language Models with Multi-Agent Reinforcement Learning for Structured Data Generation (Extended Version)

Aja Khanal

University of Western Ontario  
London, Canada  
akhanal3@uwo.ca

Kaushik T. Ranade

University of Western Ontario  
London, Canada  
kranade@uwo.ca

Rishabh Agrawal

University of Western Ontario  
London, Canada  
ragrawa9@uwo.ca

Kalyan S. Basu

ICASSSD  
Cambridge, Canada  
ks.basu@gmail.com

Apurva Narayan

University of Western Ontario  
London, Canada  
apurva.narayan@uwo.ca

## ABSTRACT

Generating high-quality structured data such as JSON records, remains a fundamental challenge for large language models (LLMs), particularly when semantic richness must coexist with strict schema adherence. While autoregressive LLMs offer strong structural consistency, they often struggle with semantic variation and output diversity. In contrast, diffusion language models (DLMs) introduce powerful mechanisms for semantic richness and bidirectional decoding, yet lack the inductive biases needed for reliable structure preservation. We present **Agents of Diffusion (AoD)**, a novel framework that unifies the generative flexibility of DLMs with the reasoning capabilities of autoregressive models through language-mediated reinforcement learning. AoD frames structured text generation as a multi-agent alignment process, where a prompt optimization agent collaborates with a judge agent to iteratively guide a DLM using natural language feedback. This approach enables controllable, schema-consistent generation without modifying model parameters or relying on handcrafted constraints. AoD advances the state of controllable generation by demonstrating that diffusion models, when supervised by cooperative agents, can achieve both high semantic novelty and structural fidelity. Across multiple structured data benchmarks, AoD consistently outperforms diffusion and autoregressive baselines, establishing a new path forward for structure-aware, diversity-enhanced text synthesis. Code: <https://github.com/Idsl-group/AgentsOfDiffusion>.

## KEYWORDS

Structured Data Generation, Synthetic Data, Multi-Agent Systems, Reinforcement Learning, Diffusion Language Models

### ACM Reference Format:

Aja Khanal, Kaushik T. Ranade, Rishabh Agrawal, Kalyan S. Basu, and Apurva Narayan. 2026. Agents of Diffusion: Enhancing Diffusion Language Models with Multi-Agent Reinforcement Learning for Structured Data Generation (Extended Version). In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 20 pages.

---

*Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems ([www.ifaamas.org](http://www.ifaamas.org)). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence.

## 1 INTRODUCTION

**Agents of Diffusion (AoD)** is a multi-agent reinforcement learning framework for controllable data generation that pairs the semantic richness of diffusion language models (DLMs) with the structural precision of autoregressive large language model (LLM) agents. AoD explores a unique idea: use natural language feedback to supervise a DLM without fine-tuning, handcrafted rules, or scalar reward shaping. Two LLM agents (a prompt optimizer and a judge) communicate through verbal feedback to iteratively refine prompts, steering the DLM toward schema-conformant, diverse outputs.

Autoregressive LLMs are widely used in synthetic data pipelines because their inductive biases favor structure and token order [17, 31], yet these same biases can constrain diversity and trigger repetition or hallucination [18, 49]. DLMs, in contrast, generate text by iteratively denoising sequences in a non-causal, bidirectional manner [16, 26], which encourages broader semantic variation. However, they lack positional priors for format preservation, which makes them poorly suited for structure-sensitive tasks such as nested JSON synthesis [22]. AoD is designed to combine these strengths while compensating for their weaknesses.

Recent advances in prompt tuning, reinforcement learning, and agent-based coordination have improved autoregressive controllability [25, 29, 34], but comparable methods remain largely unexplored for DLMs due to their recent emergence. AoD closes this gap by enabling agentic supervision of DLMs through verbal alignment alone. Our optimization loop is parameter-free: the frozen DLM (LLaDA-8B) [27] never updates its weights, and the agents interact only through natural language, which supports interpretability and model-agnostic control. To realize this, we introduce a reinforcement learning algorithm that blends proximal policy optimization (PPO) and REINFORCE principles to optimize prompt updates using natural language feedback as a surrogate reward signal.

We evaluate AoD on four structured generation benchmarks that require semantic fluency and strict JSON schema adherence: **MultiWOZ**, **Super-NaturalInstructions**, **Self-Instruct**, and **TruthfulQA**. These datasets contain nested fields, varied schema formats, and diverse linguistic styles, creating a challenging testbed for structure-aware DLM control. Across this suite, AoD achieves the highest **Task Success Rate** (0.79) and the lowest **Field Overlap**

(0.29), outperforming diffusion and autoregressive baselines while indicating valid, non-memorized outputs.

The architecture is effective, reproducible, and accessible. AoD supports local open-weight models such as LLaMA 3.1 8B, Qwen-3 8B, DeepSeek-R1 8B, and Gemma-2 9B, as well as proprietary API-based models including GPT-4.1, GPT-4.1 Mini, and GPT-4.1 Nano. This flexibility enables operation in GPU-constrained environments and high-performance cloud settings alike. All experiments use a mix of consumer-grade hardware and API endpoints, showing that AoD does not require specialized infrastructure to produce high-quality, controllable structured generation.

**Contributions.** (1) We introduce **Agents of Diffusion**, the first multi-agent RL framework to guide DLMs using natural language. (2) We propose an optimization loop where LLM agents iteratively refine prompts through verbal critique, achieving schema-aligned control without reward modeling.

(3) We demonstrate reproducible state-of-the-art results on JSON-based instruction synthesis across multiple structured datasets, establishing a foundation for controllable generation in symbolic, format-constrained domains.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Structured Textual Synthetic Data

Synthetic data is increasingly important in machine learning, particularly when real data is limited, sensitive, or costly [2]. While LLMs have shown early success in freeform text generation [1, 21, 46], generating high-quality *structured* data such as tabular records or JSON outputs remains a major challenge [20]. LLMs often produce outputs that are syntactically correct but hallucinate and repeat outputs when required to generate data under nested structures. [39, 47]. Prior solutions relied on pipelines involving validation modules or latent modeling [9, 45], but these approaches are difficult to scale. Inspired by the success of diffusion models in vision [32, 33], recent work has explored their application to text [13, 48], though structure control remains limited. Our work addresses this gap by using a diffusion language model to enhance the semantic diversity of synthetic data, while ensuring structural fidelity through continuous evaluation in a multi-agent reinforcement learning setup.

### 2.2 Autoregressive Language Models

Autoregressive language models generate text by predicting each token sequentially, a paradigm that supports strong contextual coherence and structural alignment [4, 36, 38]. Their unidirectional decoding makes them effective for producing syntactically valid and schema-compliant outputs, such as JSON or tabular formats [19, 35]. However, their reliance on left-to-right generation often limits output diversity, reinforcing high-probability patterns and leading to generic or repetitive sequences [7]. While sampling strategies offer some relief, the inherent sequential bias of AR LMs constrains their generative flexibility [8, 28]. This motivates our exploration of diffusion models, which enable bidirectional and more diverse generation, while retaining structure through multi-agent control.

### 2.3 Diffusion Language Models

Diffusion language models (DLMs) generate text through iterative denoising, enabling more flexible and semantically diverse generation than autoregressive approaches. Diffusion-LM [23] introduced a continuous latent framework for controllable text synthesis, while LLaDA [27] extended masked-sequence diffusion to match or outperform AR baselines on reasoning and language tasks. DiffLM [50] applied discrete diffusion to structured tabular data. While these models show promise, they face key limitations: structure preservation remains brittle in constrained formats like JSON, the effects of prompt tuning on DLM behavior are poorly understood, and their role in multi-agent coordination is largely unexplored. We address these gaps by embedding a frozen DLM in a multi-agent reinforcement learning loop with autoregressive LLM agents, using natural language feedback and prompt-space optimization to enable schema-consistent, semantically aligned generation.

### 2.4 Prompt Tuning and Optimization

Prompt optimization is widely used to adapt LLMs without finetuning, with methods ranging from hand-crafted reasoning strategies like Chain-of-Thought [43] to automated approaches such as Promptbreeder [12] and EvoPrompt [14]. There equally exists self-rubric strategies for enhancing prompting as seen in CodecLM [42]. While effective, these methods treat prompt design as a static search problem, optimizing prompts offline without considering real-time feedback or generation dynamics [3]. This limits their adaptability, especially in tasks requiring structural precision or iterative refinement [10]. Our work reframes prompt optimization as a reinforcement learning problem, where a prompt agent learns to refine instructions based on natural language feedback. This approach is uniquely compatible with diffusion language models, whose iterative decoding allows prompts to guide generation over multiple denoising steps, enabling responsive and structure-aware control.

### 2.5 Multi-Agent Language-Based Coordination

Multi-agent systems (MAS) using LLMs have demonstrated strong performance on complex tasks by distributing reasoning across specialized, role-defined agents that communicate via natural language [15, 37]. Decentralized protocols such as critique, collaboration, or voting, enable robust, diverse solutions in domains like code synthesis and scientific discovery [11, 44]. However, language-based coordination suffers from interaction drift, inconsistent agent behavior, and weak memory retention in multi-turn exchanges [6, 30]. We address these challenges by anchoring agent interactions around a diffusion language model, whose iterative and bidirectional decoding creates a stable shared reference point. This structure reduces drift, enforces semantic consistency, and grounds prompt refinement and evaluation across dialogue turns, improving coordination without requiring external memory modules or supervision.

## 3 METHODOLOGY

### 3.1 Preliminaries

Let  $\mathcal{X}$  be the space of structured text strings (JSON objects), and let  $q_{\text{real}}$  denote an unknown probability distribution over  $\mathcal{X}$ . We observe  $n$  independent and identically distributed (i.i.d.) samples

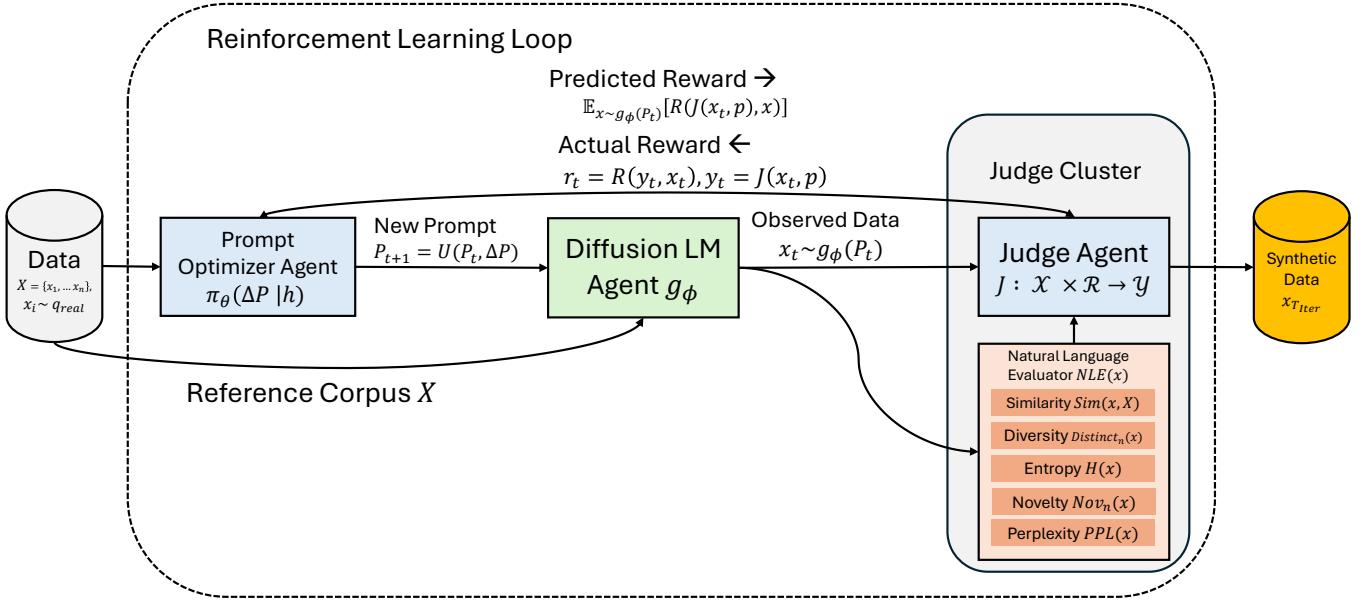


Figure 1: Agents of Diffusion: Overview of the multi-agent training framework.

$X = \{x_1, \dots, x_n\}$  with  $x_i \sim q_{\text{real}}$ . A schema  $S$  defines a valid subset  $\mathcal{L}(S) \subseteq \mathcal{X}$ , and a binary validator  $V_S : \mathcal{X} \rightarrow \{0, 1\}$  returns  $V_S(x) = 1$  if  $x \in \mathcal{L}(S)$  and 0 otherwise. Next, let  $\mathcal{P}$  be the prompt space and  $\mathcal{P}_\Delta$  the set of allowable prompt edits. A diffusion language model (DLM) with parameters  $\phi$  defines reverse-time Markov kernels  $g_{\phi,\tau}(z_{\tau-1} | z_\tau, P)$  for  $\tau = 1, \dots, T$  and a noise prior  $v_T$  on a latent token space  $\mathcal{Z}$  (identified with  $\mathcal{X}$  after decoding). To sample  $x \sim g_\phi(P)$ , one draws  $z_T \sim v_T$ , then iteratively denoises via  $z_{\tau-1} \sim g_{\phi,\tau}(z_{\tau-1} | z_\tau, P)$  until  $x = z_0$ . The forward corruption process, defined by kernels  $k_\tau(z_\tau | z_{\tau-1})$ , is standard and omitted here.

The framework includes four interacting components  $\mathcal{A} = \{\text{DLM}, J, R, \pi\}$ . The DLM  $g_\phi$  produces text samples. The judge  $J : \mathcal{X} \times \mathcal{R} \rightarrow \mathcal{Y}$  maps a candidate  $x$  and rubric  $\rho \in \mathcal{R}$  to feedback  $y = J(x, \rho) \in \mathcal{Y}$ . The scorer  $R : \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R} \times \mathbb{R}^k$  returns a scalar reward  $r = R(y, x)$  and a subreward vector  $s = R_{\text{vec}}(y, x) = (s_1, \dots, s_k)$ , where each  $s_i$  captures a measurable quality such as semantic similarity, diversity, novelty, perplexity, or entropy.

The prompt optimizer is a stochastic policy  $\pi_\theta(\Delta P | h)$  parameterized by  $\theta \in \Theta \subset \mathbb{R}^d$ , where  $\Theta$  is the set of admissible parameter vectors (e.g., neural-network weights). Given a history  $h \in \mathcal{H}$  summarizing previous prompts, samples, feedback, and scores,  $\pi_\theta$  outputs a distribution over edits  $\Delta P \in \mathcal{P}_\Delta$ . The edit operator  $U : \mathcal{P} \times \mathcal{P}_\Delta \rightarrow \mathcal{P}$  then deterministically updates prompts as  $P^+ = U(P, \Delta P)$ . Outer optimization iterations are indexed by  $t = 1, \dots, T_{\text{iter}}$ , while diffusion steps are indexed by  $\tau = 1, \dots, T$ .

Finally, for any random variable  $Z$ ,  $\mathbb{E}[Z]$  denotes expectation with respect to the randomness from  $q_{\text{real}}$ , the DLM sampling process  $g_\phi$ , and the policy  $\pi_\theta$ , unless specified otherwise.

### 3.2 Problem Formulation

As shown in Figure 1, AoD formalizes the synthesis of textual JSON data as a controllable structured generation problem within a multi-agent reinforcement learning framework. The objective is to integrate the semantic diversity of diffusion language models with the structural precision of autoregressive agents. In this setup, a diffusion model proposes diverse JSON candidates, while two LLM agents iteratively refine the conditioning prompts through natural-language feedback to ensure schema conformity and semantic consistency. This cooperative feedback loop enables controllable generation of structured text without modifying model parameters.

At each outer iteration  $t \in \{1, \dots, T_{\text{iter}}\}$ , the system maintains a prompt  $P_t \in \mathcal{P}$  and a summary history  $h_t \in \mathcal{H}$ . The DLM  $g_\phi$  generates a structured candidate  $x_t \sim g_\phi(P_t)$ . The judge  $J$  evaluates  $x_t$  under rubric  $\rho \in \mathcal{R}$ , producing feedback  $y_t = J(x_t, \rho)$ , and the scorer  $R$  converts this into numerical signals  $r_t = R(y_t, x_t)$  and  $s_t = R_{\text{vec}}(y_t, x_t)$ . The prompt optimizer, parameterized by  $\theta$ , samples an edit  $\Delta P_t \sim \pi_\theta(\Delta P | h_t)$  and updates the prompt via

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

This defines an episodic Markov decision process (MDP) with state  $z_t = (P_t, h_t)$ , action  $a_t = \Delta P_t$ , transition kernel induced by  $(g_\phi, J, R, U)$ , and reward  $r_t$ . The policy parameters  $\theta$  are optimized to maximize the expected discounted return:

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

subject to the resource constraints  $\text{Tokens} \leq B_{\text{tok}}$ ,  $\text{Calls} \leq B_{\text{calls}}$ ,  $T_{\text{iter}} \leq B_{\text{iter}}$ ,  $\gamma \in [0, 1]$ .

---

**Algorithm 1** Multi-Agent Reinforcement Learning Loop in Agents of Diffusion (AoD)

---

- 1: **Input:** schema  $S$ , rubric  $\rho$ , diffusion model  $g_\phi$ , policy  $\pi_\theta$ , edit operator  $U$
- 2: **Initialize:** prompt  $P_1 \in \mathcal{P}$ , summary history  $h_1 \in \mathcal{H}$
- 3: **for**  $t = 1$  to  $T_{iter}$  **do**
- 4:     Sample structured candidate  $x_t \sim g_\phi(P_t)$
- 5:     Judge provides feedback  $y_t = J(x_t, \rho)$
- 6:     Compute reward  $r_t = R(y_t, x_t)$  and subrewards  $s_t = R_{vec}(y_t, x_t)$
- 7:     Update history summary  $h_t = f(P_t, x_t, y_t, s_t)$
- 8:     Sample prompt edit  $\Delta P_t \sim \pi_\theta(\Delta P | h_t)$
- 9:     Apply edit:  $P_{t+1} = U(P_t, \Delta P_t)$
- 10:    Update policy parameters:  $\theta \leftarrow \theta + \eta \hat{\nabla}_\theta \mathbb{E}_{\pi_\theta}[r_t]$
- 11: **end for**
- 12: **Return:** final prompt  $P_{T_{iter}}$ , final candidate  $x_{T_{iter}}$

---

At convergence ( $T_{iter}$ ), performance is evaluated using terminal objectives that capture structural validity, semantic relevance, and diversity:

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

or equivalently in constrained form,

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

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

The diffusion parameters  $\phi$  remain fixed; controllability arises solely through the autoregressive policy  $\pi_\theta$  acting on prompts.

### 3.3 DLM Integrated Multi-Agent RL

**Prompt Optimization Agent.** The prompt optimization agent governs controllability in text-based structured JSON synthesis by steering the diffusion language model toward schema-valid, semantically coherent generations. It is instantiated as an autoregressive large language model because prompt edits are sequential and token-dependent, making autoregressive architectures ideal for learning discrete edit trajectories in text space. The agent defines a stochastic policy  $\pi_\theta(\Delta P | h)$  parameterized by  $\theta$ , generating contextually guided prompt updates that modulate the conditional output distribution of  $g_\phi$ . Its objective is to maximize the expected reward provided by the judge–scorer pair ( $J, R$ ):

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

where  $\rho$  is a task-specific rubric capturing schema and semantic fidelity. Since  $g_\phi$  is non-differentiable,  $\pi_\theta$  serves as a surrogate functional optimizer that approximates the gradient of  $\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)} [R(J(x, \rho), x)]$  through discrete, language-conditioned updates rather than backpropagation.

**Theorem 1.** Let  $\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)} [R(J(x, \rho), x)]$  as in (4). If  $\mathcal{T}(P)$  is locally Lipschitz and prompt edits  $\Delta P_t$  sampled from  $\pi_\theta(\Delta P | h_t)$  are bounded, then the iterative update  $P_{t+1} = U(P_t, \Delta P_t)$  constitutes a contraction mapping in expectation for sufficiently small step size. Thus, the sequence  $\{P_t\}$  converges to a fixed point  $P^*$  satisfying

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

ensuring stable convergence toward reward-aligned, schema-consistent prompts.

This agent operationalizes discrete autoregressive reasoning as a control layer over diffusion dynamics, transforming natural-language feedback into token-level schema alignment steps. For instance, under a booking schema  $S$  with fields `departure_city`, `arrival_city`, and `date`, an initial prompt such as “Generate travel details” may elicit feedback like “Use JSON format and include all fields.” Through iterative updates  $\Delta P_t$ , the agent refines this into “Generate a JSON object with fields {origin, destination, date} using YYYY-MM-DD format.” Each refinement incrementally aligns the DLM’s conditional distribution  $p_\phi(x | P_t)$  with the valid schema subset  $\mathcal{L}(S)$  while preserving semantic diversity. Theoretically, this mechanism bridges discrete symbolic reasoning and stochastic generation, enabling reinforcement-driven adaptation in structured text synthesis.

**Diffusion Language Model Agent.** The diffusion language model (DLM)  $g_\phi(z_{t-1} | z_t, P)$  serves as the generative backbone for synthesizing text-based structured JSON data. It defines a reverse-time Markov process that reconstructs text tokens from gradually denoised latent representations, yielding the conditional distribution  $p_\phi(x | P)$ . Unlike autoregressive models that factorize  $p(x | P)$  sequentially as  $\prod_i p(x_i | x_{<i}, P)$ , the DLM estimates  $p_\phi(x | P)$  implicitly through iterative denoising. This non-causal, bidirectional formulation allows each denoising step to condition on global context rather than local token dependencies, resulting in broader coverage of valid schema-conformant configurations.

**Proposition 1.** Let  $p_\phi(x | P)$  and  $p_{AR}(x | P)$  denote diffusion and autoregressive conditional distributions trained on the same structured dataset with schema  $S$ . If both minimize divergence from the real data distribution  $q_{real}(x)$  under bounded reconstruction error and finite diffusion horizon  $T$ , then

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

indicating that diffusion better approximates the real data manifold and captures a wider set of semantically valid configurations.

Theoretically, this property positions the DLM as the diversity-preserving agent within AoD. It expands the support of  $p_\phi(x | P)$  across multiple valid schema realizations, enabling the synthesis of varied yet coherent JSON structures. The autoregressive agents, in contrast, provide the constraint mechanism that ensures syntactic and semantic adherence to  $\mathcal{L}(S)$ . Together, they form a complementary system: diffusion drives diversity, while autoregression enforces structure. This achieves controlled, schema-aligned generation of structured text data.

**Judge Agent.** The judge cluster combines the LLM-based judge  $J$  and the Natural Language Evaluator (NLE), forming the evaluation subsystem responsible for interpreting and supervising outputs from the diffusion language model  $g_\phi$ . The NLE receives each generated JSON sample  $x \in \mathcal{X}$  and computes five quantitative metrics defined in the preliminaries: semantic similarity  $Sim(x, X)$ , diversity  $Distinct_n(x)$ , entropy  $H(x)$ , novelty  $Nov_n(x)$ , and perplexity  $PPL(x)$ . These metrics jointly describe how well  $x$  aligns with the reference dataset  $X$ , how varied and fluent it is, and whether it generalizes beyond seen examples. The NLE then converts these

numeric values into structured natural language statements, which summarize deviations and attributions in an interpretable form. The LLM judge  $J$ , instantiated as an autoregressive model, consumes this structured textual feedback together with the rubric  $\rho$  and evaluates each sample through a fixed set of rubric-aligned yes/no questions (e.g., “Is the JSON structurally complete?”, “Are all required fields present?”, “Is the text semantically faithful?”). Based on both the quantitative assessments and its own contextual reasoning,  $J$  generates the final critique  $y = J(x, \rho)$ , which the scorer translates into a scalar reward  $r = R(y, x)$  and subreward vector  $s = R_{\text{vec}}(y, x)$ . For example, the cluster might produce: “*The JSON is fluent but missing arrival\_city; the date format should be YYYY-MM-DD.*” This feedback then guides the prompt optimization policy  $\pi_\theta(\Delta P \mid h)$ , completing the reinforcement loop.

Correspondingly, the NLE isolates measurable properties of the generated text, providing low-variance, disentangled feedback signals that prevent noisy gradients and improve the reliability of downstream optimization. The LLM judge  $J$  transforms these discrete measurements into a smooth, natural-language surrogate of the underlying reward landscape, enabling gradient-free optimization while maintaining semantic transparency. The autoregressive formulation of  $J$  is particularly important for structured JSON synthesis, as evaluating conformance, field ordering, and key dependencies requires sequential reasoning over tokens. By processing feedback in a left-to-right manner,  $J$  preserves causal consistency in its critiques and ensures alignment with how  $\pi_\theta$  performs token-level edits to  $P$ . Thus, autoregressive reasoning enforces structured coherence while maintaining semantic flexibility, allowing the system to generalize across diverse schemas without manual rule engineering.

Theoretically, the cluster defines an expected feedback operator

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

that stabilizes learning in the non-differentiable environment  $(g_\phi, J, R)$ . The NLE grounds  $\mathcal{T}(P)$  in verifiable quantitative signals, while the autoregressive judge smooths discontinuities by mapping discrete validation outcomes to continuous linguistic explanations. This synergy reduces variance in the policy-gradient estimate, improves credit assignment, and enforces reward monotonicity with respect to schema-conformant and semantically faithful outputs. Together, they form a theoretically consistent bridge between numeric supervision and symbolic prompt control, balancing the structured precision of autoregressive models with the generative diversity of diffusion-based synthesis.

**Proposition 2.** *Let  $\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)} [R(J(x, \rho), x)]$ . Suppose each component in  $R_{\text{vec}}(y, x)$  is bounded, the mapping from the NLE’s metric vector to textual feedback  $y = J(x, \rho)$  is Lipschitz with constant  $L_J$  with respect to  $(\text{Sim}(x, X), \text{Distinct-}n(x), H(x), \text{Nov-}n(x), \text{PPL}(x))$ , and  $R$  is monotone in these components. Then  $\mathcal{T}(P)$  is locally Lipschitz in  $P$  and preserves ordering with respect to semantic fidelity: if  $P_1, P_2$  induce samples such that  $\mathbb{E}[\text{Sim}(x, X) \mid P_1] > \mathbb{E}[\text{Sim}(x, X) \mid P_2]$ , then  $\mathcal{T}(P_1) > \mathcal{T}(P_2)$ . Consequently, policy updates driven by  $\mathcal{T}(P)$  are stable in expectation and prioritize semantically faithful prompts.*

## 4 EXPERIMENTS AND RESULTS

### 4.1 Experimental Setup

**Hardware.** To demonstrate the accessibility and reproducibility of AoD, all experiments were run on a consumer-grade workstation with an AMD Ryzen 9 7900X (12-core, 24-thread, 4.7 GHz base), 32 GB DDR5 RAM, and an NVIDIA RTX 4080 SUPER GPU (16 GB VRAM). This setup reflects hardware that is widely available to individual researchers and developers, ensuring that AoD does not rely on specialized infrastructure or large-scale compute clusters.

**Models.** Our experiments use both open-source and API-based language models to highlight AoD’s flexibility and hardware independence. Eight autoregressive models were used for the prompt optimizer and LLM judge roles: **LLaMA-3.1 8B** (32 layers, 40 heads, 4-bit quantization; temperature 0.7 for prompting, 0.2 for judgment), **Qwen-3 8B** (multilingual, LoRA-enabled, nucleus sampling with  $p = 0.9$ ), **DeepSeek-R1 8B** (NTK-aware, top- $k$  sampling with  $k = 40$ ), **Gemma-2 9B** (beam search width 3), **Mistral 7B** (grouped-query attention, 8-bit decoding), and three API-based models: **GPT-4.1 Nano**, **Mini**, and **GPT-4.1**, which support lightweight to high-fidelity generation. All experiments used the same autoregressive model for both the prompt optimizer and LLM judge for consistency and computational simplicity. The generator agent, by contrast, was exclusively **LLaDA 8B**, a discrete diffusion language model with 32 layers, sinusoidal embeddings, 1024-token input, and  $T = 12$  denoising steps. LLaDA operates in FP16 mode and disables classifier-free guidance to preserve sampling diversity. Across both local and API models, AoD demonstrates consistent, high-quality performance without reliance on specialized hardware.

**Datasets.** We evaluate our framework to synthesize JSON data on four publicly available datasets chosen for their diversity in structure, semantics, and generation objectives. **MultiWOZ** [5] is a multi-domain dialogue dataset with rich slot-filling annotations, useful for testing structure preservation and schema alignment. **SuperNaturalInstructions** [41] contains diverse instruction-response pairs across hundreds of tasks, enabling generalization over prompt types. **TruthfulQA** [24] provides factuality-challenging questions, useful for evaluating hallucination and semantic precision. **Self-Instruct** [40] consists of instruction-based examples distilled from LLMs, supporting experiments on prompt-response synthesis in alignment-critical tasks. Each dataset is randomly subsampled prior to training and evaluation to reduce computational overhead and mitigate memorization. By exposing the model to only a small, randomly selected portion of the full dataset, we minimize the risk of copying specific examples during generation and ensure that performance reflects generalization to unseen instructions.

**Baselines.** We compare AoD against six complementary baselines that represent the main paradigms of structured text generation. **Diffusion-LM** [23] and **DiffLM** [50] are diffusion-based models emphasizing semantic diversity and multimodal coverage, serving as diversity-oriented references. **CodecLM** [42], **PromptBreeder** [12], and **EvoPrompt** [14] are autoregressive prompt-optimization methods that provide strong control-oriented baselines, reflecting existing strategies for structured prompting and language-driven refinement. Finally, **UniGen** [45] embodies validation-based synthesis, enforcing symbolic and schema-level

constraints at generation time. Together, these six baselines span the design space of *structure control* (AR-based), *diversity* (diffusion-based), and *constraint enforcement* (validation-based), providing a comprehensive comparison framework for AoD’s contribution: unifying all three through a multi-agent reinforcement-learning mechanism that produces schema-conformant yet semantically diverse structured JSON data.

**Evaluation Metrics.** Our evaluation framework separates metrics used for agent feedback from those used for independent verification. During training, the judge agent leverages five grounded metrics: perplexity  $PPL(x)$ , semantic similarity  $\text{Sim}(x, X)$ , diversity  $\text{Distinct-}n(x)$ , token entropy  $H(x)$ , and novelty  $\text{Nov-}n(x)$ , to generate interpretable natural language feedback for the prompt optimizer. These metrics provide structural and semantic supervision without exposing scalar reward values directly, thereby avoiding reinforcement bias or overfitting to numeric targets. However, as a measure for independent evaluation, we report standard text-based quality metrics including BLEU, ROUGE, and METEOR to quantify syntactic and lexical correspondence between generated and reference samples. This ensures that our reported performance reflects true generation quality rather than reinforcement feedback bias. Furthermore, to assess the downstream finetunability and functional reliability of the synthesized data, we also compute the *Task Success Rate* (TSR), which measures the proportion of valid, semantically consistent, and diverse generations meeting all constraints.

**Memorization and Collusion Verification.** The five reward metrics used during training (perplexity  $PPL(x)$ , semantic similarity  $\text{Sim}(x, X)$ , diversity  $\text{Distinct-}n(x)$ , token entropy  $H(x)$ , and novelty  $\text{Nov-}n(x)$ ) jointly regulate memorization and collusion within the multi-agent reinforcement learning loop. Each metric enforces a distinct behavioral constraint:  $PPL(x)$  ensures linguistic coherence and penalizes degenerate text,  $\text{Sim}(x, X)$  promotes semantic alignment,  $\text{Distinct-}n(x)$  and  $H(x)$  encourage lexical variability, and  $\text{Nov-}n(x)$  penalizes verbatim reuse of the reference dataset  $X$ . The judge uses these quantitative signals to answer a fixed set of rubric-based yes/no questions, ensuring that reinforcement is grounded in objective structure and meaning rather than hidden coordination between agents. Because the prompt optimizer  $\pi_\theta$  never directly observes scalar rewards but instead receives natural language feedback derived from them, it cannot exploit the reward function through collusion or memorization. Numeric trends across iterations provide diagnostic signals, such as a simultaneous increase in  $\text{Sim}(x, X)$  and decrease in  $\text{Nov-}n(x)$ , which explicitly reveal potential leakage or overfitting.

To further ensure independence and verify that generated data remains distinct from the training corpus, we introduce the *Field Overlap* metric as a post-hoc measure. Field Overlap computes the proportion of key-value pairs or fields in generated JSON samples that exactly match those in the reference set. High overlap values indicate potential copying or memorization, while low overlap combined with low  $PPL(x)$  suggests faithful generalization with coherent generation. Unlike  $\text{Sim}(x, X)$ , which captures semantic similarity, Field Overlap explicitly measures structural duplication,

making it a direct test for memorization or cross-agent information leakage. Together, these signals enable both in-loop and independent verification of data novelty, ensuring that AoD produces diverse, semantically faithful, and unbiased synthetic JSON data.

## 4.2 Discussion

Table 1 highlights that AoD achieves a rare balance between structural precision and generative diversity, outperforming both diffusion-based and autoregressive systems. High Similarity (0.88) combined with strong Diversity (0.72) and Novelty (0.83) demonstrates that AoD generates data that remains semantically faithful while exploring new schema-consistent configurations. The Entropy score (6.10) indicates balanced lexical richness rather than repetitive phrasing, and the low Perplexity (22.1) confirms fluent and coherent language modeling. This pattern is not coincidental; it directly reflects the balance between exploration and regulation within AoD’s architecture. The diffusion generator introduces stochastic breadth, while the judge cluster applies linguistic and structural constraints that stabilize the output space. Reinforcement-guided optimization aligns these opposing forces, producing samples that are both creative and compositionally valid, even in datasets with complex nested structures.

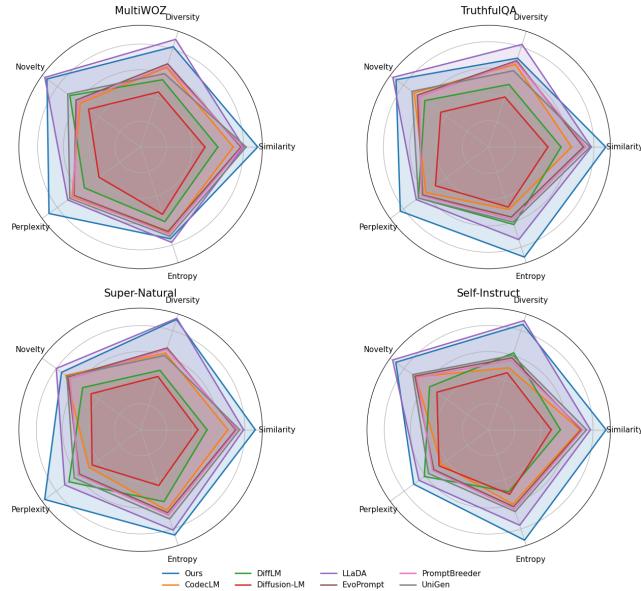
Independent metrics reinforce this interpretation. AoD leads on BLEU, ROUGE-L, and METEOR, showing that its diversity does not compromise grammatical or semantic fidelity. The high Task Success Rate (0.79) indicates that generated records satisfy both content and structure requirements, while the lowest Field Overlap (0.29) confirms that AoD avoids memorization by generating distinct key-value combinations unseen in training. These numerical patterns arise from the multi-agent reward structure: the natural language evaluator (NLE) introduces interpretive continuity by converting discrete metric signals into graded linguistic feedback, while the autoregressive judge performs sequential validation across keys and fields. This layered supervision stabilizes reward propagation, making optimization smoother and preventing overfitting to numeric heuristics. The result is visible in the metrics—Similarity and Perplexity improve simultaneously, Diversity and Novelty rise without structural drift, and Entropy remains high yet coherent. AoD’s reinforcement signals therefore encode both form and meaning, producing generalization that persists across unseen distributions.

These dynamics also clarify why AoD achieves an uncommon combination of low Perplexity and high Entropy. Traditional diffusion systems increase diversity but often generate syntactically unstable text, while autoregressive systems enforce structure at the cost of variability. AoD bridges this divide by coupling diffusion-driven exploration with sequential constraint verification. The NLE’s linguistic grounding allows lexical and semantic expansion to occur in a controlled way, and the judge’s autoregressive reasoning enforces causal dependencies between fields. Together, they produce the observed equilibrium: high-entropy text that remains syntactically fluent and semantically consistent. This mechanism explains why AoD avoids the typical diffusion drift toward incoherence and the autoregressive bias toward repetition. The differences among baseline models further contextualize AoD’s superiority. Static autoregressive systems such as LLaMA and Qwen prioritize conditional likelihood maximization, maintaining high Similarity

**Table 1: Comparison of AoD with baselines.** The first five metrics correspond to those used in training, while the last five serve as independent evaluation metrics. Higher is better for all metrics except Perplexity and Field Overlap. Values averaged across all datasets and Prompt Optimizer + Judge LLM pairs for AoD. Each experiment was repeated 15 times.

Model	Similarity	Diversity	Novelty	Entropy	Perplexity	BLEU	ROUGE-L	METEOR	TSR	Field Overlap
<i>Static Autoregressive Baselines (single-pass prompting)</i>										
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
<i>Diffusion and Prompt-Based Baselines</i>										
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
<b>AoD (ours)</b>	<b>0.88</b>	<b>0.82</b>	<b>0.83</b>	<b>6.10</b>	<b>22.1</b>	<b>35.6</b>	<b>40.1</b>	<b>29.3</b>	<b>0.79</b>	<b>0.29</b>

but collapsing on Diversity and Novelty due to deterministic decoding. Diffusion baselines like DiffLM or Diffusion-LM invert this pattern, producing diverse but structurally fragile data because their denoising trajectories lack schema-aware conditioning. Prompt-evolution frameworks like EvoPrompt and PromptBreeder improve variability through heuristic mutation but fail to sustain progress because they lack credit assignment across sequential edits. In contrast, AoD closes this optimization loop through dynamic prompt refinement driven by multi-dimensional feedback from the judge cluster. The prompt optimizer learns a policy that adjusts prompts not only based on reward magnitude but also on the linguistic context of errors, enabling continual improvement across iterations.

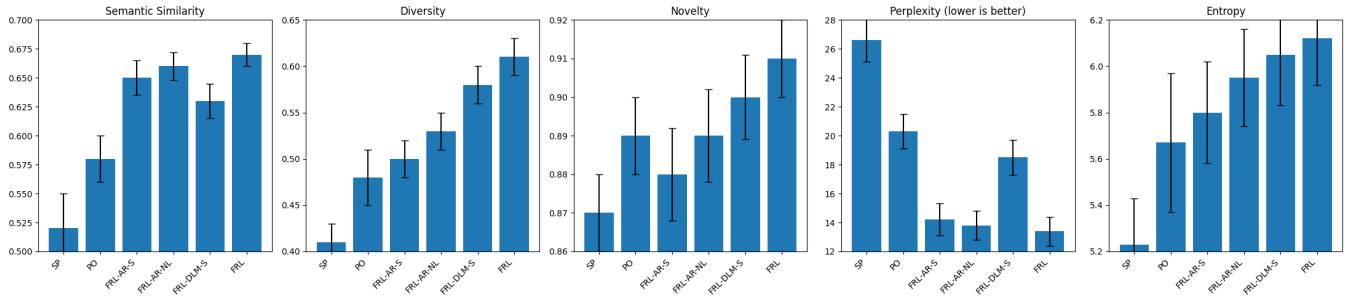


**Figure 2: Comparison of normalized metrics across datasets.**

Figure 2 illustrates AoD’s consistency across datasets, where its polygons expand uniformly along all metric axes. In structurally complex domains like *MultiWOZ* and *Super-Natural*, AoD sustains high Similarity and low Perplexity while widening Diversity and Entropy, demonstrating its ability to preserve schema integrity while encouraging lexical variation. On reasoning-heavy datasets such as *TruthfulQA* and *Self-Instruct*, AoD balances Novelty and Similarity, showing that the judge cluster and diffusion generator collaboratively regulate generalization without overfitting. Competing methods either favor diversity at the expense of structure (Diffusion-LM, DiffLM) or maintain structure but exhibit reduced novelty (PromptBreeder, EvoPrompt). The near-regular shape of AoD’s region across all four datasets confirms stable performance and adaptability to different data modalities, reinforcing its role as a schema-faithful yet diverse structured data generator.

### 4.3 Ablation Study

**Agentic and Reward Ablation.** Figure 3 illustrates the impact of each agentic component and feedback mechanism within AoD. Introducing the autoregressive prompt optimizer (PO) leads to improvements across all metrics. This indicates that prompt refinement, even without reinforcement, enhances fluency and schema alignment by allowing structured edits over iterations. However, Diversity and Entropy remain constrained since optimization is deterministic and lacks stochastic exploration. Adding reinforcement learning with scalar rewards (*FRL-AR-S*) further enhances Similarity and lowers Perplexity, as scalar feedback enables structured reward shaping for grammatical and syntactic correctness. Yet, gains in Diversity and Entropy are limited by the autoregressive generator’s decoding bias, which narrows the output space to high-liability continuations. Replacing scalar rewards with natural language (*FRL-AR-NL*) yields smoother reward propagation and improved fluency, as the judge’s interpretable feedback offers token-level supervision. Nevertheless, the sequential nature of the autoregressive generator continues to limit the attainable variety.



**Figure 3: Ablation of agents, averaged across datasets and Optimizer–Judge pairs. SP = Static Prompt, PO = Prompt Optimizer, FRL–AR–S = Autoregressive generator with scalar rewards, FRL–AR–NL = Autoregressive generator with natural language feedback, FRL–DLM–S = Diffusion generator with scalar rewards, FRL = Diffusion generator with natural language feedback.**

**Table 2: Transferability for prompt optimizer (rows) and judge (columns). Reports Diversity ( $\uparrow$ ) and GPU runtime seconds ( $\downarrow$ ) averaged for all datasets. LLaDA is fixed generator. L=LLaMA, Q=Qwen, D=DeepSeek, G=Gemma, M=Mistral.**

	L		Q		D		G		M	
Opt.	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

Transitioning to a diffusion generator while maintaining scalar rewards (*FRL–DLM–S*) produces an increase in Diversity, Novelty, and Entropy. The diffusion-based generator  $g_\phi$  enables bidirectional context propagation, allowing multiple semantic trajectories to emerge under the same prompt while maintaining structural coherence. However, in the absence of natural language feedback, Similarity improvements are modest and Perplexity variance rises, reflecting coarse reward alignment. The complete AoD configuration (*FRL*) combines the strengths of both: diffusion-driven diversity with linguistically grounded feedback. The LLM Judge and NLE stabilize the learning dynamics, reducing Perplexity and reinforcing high Similarity and Entropy while maintaining elevated Novelty.

**Model Transferability.** Table 2 highlights AoD’s model-agnostic behavior across various combinations of autoregressive LLMs serving as the Prompt Optimizer and Judge agents, with LLaDA fixed as the generator. Across all pairings, Diversity scores remain consistently high (0.79–0.86), indicating that reinforcement-driven coordination generalizes regardless of the underlying model architecture. This demonstrates that AoD’s policy learning operates on the shared language space of feedback and prompts, rather than relying on any specific model’s internal representations. The prompt–feedback exchange mechanism  $\pi_\theta(\Delta P \mid h)$  is thus invariant to the optimizer and judge configurations, enabling interchangeable agents without performance collapse. Runtime results further support AoD’s reproducibility on consumer-grade hardware. Average GPU runtimes per feedback–generation cycle range from

110–122 seconds, even for 8–9B parameter models, confirming that multi-agent rollouts remain tractable under mid-range configurations. This efficiency stems from the frozen generator  $g_\phi$  and the lightweight communication loop between autoregressive agents, which limits backpropagation overhead. Together, these results establish that AoD can be replicated using open-weight or API-based LLMs while preserving diversity and stability, making it accessible without dependence on high-end compute resources.

**Case Study: Structured JSON Synthesis in AoD.** We demonstrate AoD’s functionality using the *MultiWOZ 2.1* booking domain as a task. The Prompt Optimizer first drafts a schema-conditioned instruction, e.g., “Generate a JSON object with fields {name, address, phone, price\_range, postcode}.” The DLM  $g_\phi$  then produces diverse samples such as `{"name": "Parkview Inn", "address": "12 Milton Rd", "phone": "01223 443890", "price_range": "moderate", "postcode": "CB4 1LG"}`. The NLE computes metrics, while the LLM Judge transforms them into feedback, e.g., “The JSON is valid and fluent but duplicates price patterns; introduce more unique names and locations.” The optimizer uses this feedback through  $\pi_\theta(\Delta P \mid h)$ , iteratively improving prompt specificity and sampling balance. Over successive iterations, Similarity rises from 0.64 to 0.88, Diversity and Novelty exceed 0.80, and Perplexity drops from 31.2 to 22.5. The Judge Agent confirms that the generated JSON records remain syntactically correct yet distinct. Field Overlap falls to 0.29, and TSR reaches 0.79, indicating low memorization and strong generalization. Qualitatively, AoD captures both structure and variety, generating realistic data within five iterations.

## 5 CONCLUSION

AoD is the first framework to study how DLMs operate in a multi-agent reinforcement learning environment, demonstrating that natural language feedback can drive controllable, high-quality structured data generation. Furthermore, AoD achieves schema-compliant JSON outputs with higher diversity, novelty, and perplexity than standard LLM counterparts, while remaining reproducible on consumer hardware. Although limited to JSON synthesis, this work establishes DLMs as a powerful alternative for structured data generation and opens the door to future extensions for tabular datasets, code, and other structured domains.

## REFERENCES

- [1] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2019. Not Enough Data? Deep Learning to the Rescue! <https://arxiv.org/abs/1911.03118>
- [2] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive Exploration of Synthetic Data Generation: A Survey. <https://arxiv.org/abs/2401.02524>
- [3] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2024. Guiding LLMs THE RIGHT WAY: Fast, non-invasive constrained generation. <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 Henighen, 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, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. *arXiv.org* 4 (05 2020). <https://arxiv.org/abs/2005.14165>
- [5] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Íñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. <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, and Yun Li. 2024. Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods. *IEEE Transactions on Neural Networks and Learning Systems* (01 2024), 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, Sebastian Gehrmann, and et al. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311* (April 2022). <https://arxiv.org/abs/2204.02311>
- [8] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. <https://arxiv.org/abs/2501.12948> arXiv preprint arXiv:2501.12948.
- [9] Jasper Dekoninck, Marc Fischer, Luca Beurer-Kellner, and Martin Vechev. 2023. Controlled Text Generation via Language Model Arithmetic. <https://arxiv.org/abs/2311.14479>
- [10] Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2025. XGRAMMAR: Flexible and efficient structured generation engine for large language models. <https://arxiv.org/abs/2411.15100>
- [11] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Moratch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. <https://doi.org/10.48550/arXiv.2305.14325>
- [12] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution. <https://doi.org/10.48550/arXiv.2309.16797>
- [13] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2022. DiffuseSeq: Sequence to Sequence Text Generation with Diffusion Models. <https://arxiv.org/abs/2210.08933>
- [14] Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers. *arXiv (Cornell University)* (01 2023). <https://doi.org/10.48550/arXiv.2309.08532>
- [15] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. <https://doi.org/10.48550/arXiv.2402.01680>
- [16] Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, van , and Tim Salimans. 2021. Autoregressive Diffusion Models. <https://arxiv.org/abs/2110.02037>
- [17] Jiaxin Huang, Shixiang Gu, Le Hou, Yuxin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large Language Models Can Self-Improve. *ACL Anthology Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (01 2023). <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, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *Comput. Survey* 55 (11 2022). <https://doi.org/10.1145/3571730>
- [19] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Choplast, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lampe, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. <https://doi.org/10.48550/arXiv.2310.06825>
- [20] Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction. <https://arxiv.org/abs/2303.04132>
- [21] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv:1909.05858 [cs]* (09 2019). <https://arxiv.org/abs/1909.05858>
- [22] Chungpa Lee, Jongho Im, and Kim Joseph. 2025. A Generalized Theory of Mixup for Structure-Preserving Synthetic Data. <https://arxiv.org/abs/2503.02645>
- [23] Xiang Lisa Li, John Thickstun, Ishaaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. <https://arxiv.org/abs/2205.14217>
- [24] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958 [cs]* (09 2021). <https://arxiv.org/abs/2109.07958>
- [25] Shengcai Liu, Caishun Chen, Xinghua Qu, Ke Tang, and Yew-Soon Ong. 2023. Large Language Models as Evolutionary Optimizers. <https://arxiv.org/abs/2310.19046>
- [26] Calvin Luo. 2022. Understanding Diffusion Models: A Unified Perspective. *arXiv:2208.11970 [cs]* (08 2022). <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, and Chongxuan Li. 2025. Large Language Diffusion Models. <https://arxiv.org/abs/2502.09992>
- [28] OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774 [cs]* (03 2023). <https://doi.org/10.48550/arXiv.2303.08774>
- [29] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv:2203.02155 [cs]* (03 2022). <https://arxiv.org/abs/2203.02155>
- [30] Bo Pan, Jiaying Lu, Ke Wang, Li Zheng, Zhen Wen, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. 2024. AgentCoord: Visually Exploring Coordination Strategy for LLM-based Multi-Agent Collaboration. <https://arxiv.org/abs/2404.11943>
- [31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. <https://arxiv.org/abs/1910.10683>
- [32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. *arXiv:2102.12092 [cs]* (02 2021). <https://arxiv.org/abs/2102.12092>
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752 [cs]* (04 2022). <https://arxiv.org/abs/2112.10752>
- [34] Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv:2303.11366 [cs]* (03 2023). <https://arxiv.org/abs/2303.11366>
- [35] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. <https://doi.org/10.48550/arXiv.2403.08295> arXiv preprint arXiv:2403.08295.
- [36] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambo, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lampe. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971 [cs]* (02 2023).
- [37] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. <https://arxiv.org/abs/2501.06322>
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. <https://arxiv.org/abs/1706.03762>
- [39] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. <https://doi.org/10.48550/arXiv.2305.15041>
- [40] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions. *arXiv (Cornell University)* (12 2022). <https://doi.org/10.48550/arXiv.2212.10560>
- [41] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, and et al. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. <https://arxiv.org/abs/2204.07705> arXiv preprint arXiv:2204.07705.
- [42] Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024. CodecLM: Aligning Language Models with Tailored Synthetic Data. <https://arxiv.org/abs/2404.05875>
- [43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting

- Elicits Reasoning in Large Language Models. *arXiv:2201.11903 [cs]* (10 2022). <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, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. <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, and Lichao Sun. 2024. UniGen: A Unified Framework for Textual Dataset Generation Using Large Language Models. <https://arxiv.org/abs/2406.18966>
- [46] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. <https://arxiv.org/abs/2202.07922>
- [47] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. <https://doi.org/10.48550/arXiv.2306.15895>
- [48] Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzeifa Rangwala, and George Karypis. 2023. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. <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, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv:2303.18223 [cs]* (03 2023). <https://arxiv.org/abs/2303.18223>
- [50] Ying Zhou, Xinyao Wang, Yulei Niu, Yaojie Shen, Lexin Tang, Fan Chen, Ben He, Le Sun, and Longyin Wen. 2024. DiffLM: Controllable Synthetic Data Generation via Diffusion Language Models. <https://arxiv.org/abs/2411.03250>

## A APPENDIX

This appendix provides all resources necessary to fully reproduce the experiments described in the **Agents of Diffusion (AoD)** study. AoD frames structured JSON data generation as a multi-agent reinforcement learning problem, where a *prompt optimization agent* and an *LLM judge* collaborate to guide a frozen diffusion language model through iterative refinement. Each experiment relies on natural language prompts to perform three key steps: (1) extract schema-aware metadata from a reference dataset, (2) generate candidate synthetic outputs under schema constraints, and (3) iteratively refine the prompts using judge feedback until the outputs meet structural and semantic quality thresholds.

This appendix contains:

- (1) The exact prompts used for each stage of the AoD generation loop.
- (2) A demonstration of a full experiment log for the MultiWOZ 2.1 dataset showing how natural language feedback drives iterative prompt updates and the eventual convergence of the multi-agent loop.

By following these prompts and experiment logs, researchers can replicate the generation, evaluation, and refinement process without modifying model weights, confirming the reproducibility of AoD's results on consumer-grade hardware.

### A.1 Prompt Optimizer

The prompt optimizer evaluates and rewrites prompts to ensure the generator LLM produces schema-compliant and realistic entries. Below is a sample log showing the initial instruction, the optimized LLM prompt, and the prompt optimizer's performance prediction.

#### Prompt: Initial Instruction

Please generate an LLM prompt to create more of this JSON dataset.

#### Output: Optimized LLM Prompt

Given the following examples of venue data in JSON format:

```
[  
  {  
    "address": "pool way, whitehill road, off  
    newmarket road",  
    "area": "east",  
    "entrance fee": "?",  
    "id": "1",  
    "location": [52.208789, 0.154883],  
    "name": "abbey pool and astroturf pitch",  
    "openhours": "?",  
    "phone": "01223902088",  
    "postcode": "cb58nt",  
    "pricerange": "?",  
    "type": "swimmingpool"  
  },
```

```
{  
  "address": "park street",  
  "area": "centre",  
  "entrance fee": "?",  
  "id": "2",  
  "location": [52.208699, 0.12006],  
  "name": "adc theatre",  
  "openhours": "?",  
  "phone": "01223300085",  
  "postcode": "cb58as",  
  "pricerange": "?",  
  "type": "theatre"
```

}

]

Please create 5 additional, similar entries in this dataset with realistic but fictional data for different venues around Cambridge (plausible street addresses, postcodes, location coordinates, and venue types such as "cafe", "library", "gym", "cinema", etc.).

#### Requirements:

- A unique id for each new entry (incrementing from "3").
- Accurate field formatting matching the samples.
- Venue coordinates must be within the Cambridge area.
- Names, addresses, and properties should be unique.
- Return ONLY a JSON list of the new entries.

#### Output: Prompt Performance Prediction

**Strengths:** - Provides clear structure and explicit output format with examples. - Specifies unique IDs and plausible venues for realism. - Instruction to return only JSON minimizes formatting errors.

**Potential Weaknesses:** - Ambiguity in handling unknown fields (e.g., "?" usage not explicitly stated). - Slight risk of fabricated or out-of-range coordinates/postcodes. - Some models may still append explanations if temperature is high.

**Expected Output:** - 5 well-structured, unique JSON entries with plausible Cambridge data. - Minor risk of duplicate IDs or slightly incorrect fields.

**Overall Prediction: High Performance** - Likely to yield correct, high-quality JSON output.

**Suggestions for Optimization:** - Clarify the rule for unknown fields: use "?" as in the example. - Optionally suggest using realistic Cambridge postcodes (CB1–CB5).

### A.2 Diffusion Generator

Once the Prompt Optimizer produces a refined instruction, the **Diffusion Generator** uses this prompt to guide the synthesis of new structured entries. In AoD, the diffusion-based generator operates as a conditional sampler: it accepts the optimized prompt, interprets the metadata and schema rules encoded in it, and iteratively refines

noisy token sequences into **fully structured JSON outputs**. This mechanism allows the generator to maintain both **schema compliance** and **distributional alignment** with the source dataset, while introducing controlled diversity across new records.

The following example shows five synthetic records produced for the MultiWOZ 2.1 attraction domain. Each record respects the JSON schema, includes plausible Cambridge addresses, valid postcode patterns, and venue types like cafes, libraries, gyms, and cinemas.

Output: Sample Generated Data

```
[  
  {  
    "address": "16 mill road",  
    "area": "south",  
    "entrance fee": "?",  
    "id": "3",  
    "location": [52.201572, 0.139430],  
    "name": "bloom café",  
    "openhours": "?",  
    "phone": "01223992212",  
    "postcode": "cb12ad",  
    "pricerange": "?",  
    "type": "cafe"  
  },  
  {  
    "address": "53 castle street",  
    "area": "northwest",  
    "entrance fee": "?",  
    "id": "4",  
    "location": [52.214623, 0.113912],  
    "name": "castle fitness hub",  
    "openhours": "?",  
    "phone": "01223881645",  
    "postcode": "cb30ah",  
    "pricerange": "?",  
    "type": "gym"  
  },  
  {  
    "address": "10 market hill",  
    "area": "centre",  
    "entrance fee": "?",  
    "id": "5",  
    "location": [52.205287, 0.119841],  
    "name": "grand reel cinema",  
    "openhours": "?",  
    "phone": "01223478901",  
    "postcode": "cb23nj",  
    "pricerange": "?",  
    "type": "cinema"  
  },  
  {  
    "address": "44 chesterton road",  
    "area": "northeast",  
    "entrance fee": "?",  
    "id": "6",  
    "location": [52.215764, 0.133298],  
    "name": "riverside library",  
    "openhours": "?",  
    "phone": "01223011228",  
    "postcode": "cb41eh",  
    "pricerange": "?",  
    "type": "library"  
  },  
  {  
    "address": "123 hills road",  
    "area": "south",  
    "entrance fee": "?",  
    "id": "7",  
    "location": [52.193887, 0.134902],  
    "name": "pine bistro",  
    "openhours": "?",  
    "phone": "01223873336",  
    "postcode": "cb20pw",  
    "pricerange": "?",  
    "type": "cafe"  
  }]
```

```
"id": "6",  
"location": [52.215764, 0.133298],  
"name": "riverside library",  
"openhours": "?",  
"phone": "01223011228",  
"postcode": "cb41eh",  
"pricerange": "?",  
"type": "library"  
},  
{  
  "address": "123 hills road",  
  "area": "south",  
  "entrance fee": "?",  
  "id": "7",  
  "location": [52.193887, 0.134902],  
  "name": "pine bistro",  
  "openhours": "?",  
  "phone": "01223873336",  
  "postcode": "cb20pw",  
  "pricerange": "?",  
  "type": "cafe"  
}]
```

This output illustrates the generator's ability to **expand the dataset** while maintaining realistic geographical patterns, proper schema adherence, and controlled diversity, fulfilling the structural and semantic criteria set by the multi-agent AoD loop.

### A.3 Natural Language Evaluator

In the AoD framework, the **Natural Language Evaluator** is a metric-driven analysis module that computes scores such as semantic similarity, diversity, novelty, perplexity, and entropy over generated outputs. Unlike the generator or the prompt optimizer, it is **not an LLM**; its role is to produce structured signals that the **LLM judge** can interpret as part of the reward feedback loop.

Simply providing raw numerical scores is often ineffective, because LLMs do not inherently understand what a value like 0.29 for diversity or 7.59 for perplexity signifies. To bridge this gap, the evaluator converts each score into a short natural language explanation that contextualizes the metric in plain terms. This allows the LLM judge to reason about the quality of the batch and generate more actionable feedback for the next iteration.

A representative evaluator output is shown below:

Output: Sample Evaluation

```
Semantic Similarity: 0.65 - Moderate similarity:  
some overlap with the reference, but key  
details or meanings may be missing or altered.  
Diversity: 0.29 - Low diversity: the text likely  
reuses similar words or phrases, possibly  
indicating redundancy or mode collapse.
```

Novelty: 0.97 - Highly novel: most of the content in the generated text is new compared to the reference corpus, suggesting strong generalization or creativity.

Perplexity: 7.59 - Very fluent: the language model finds this text highly predictable, which typically means it's grammatically correct, natural, and human-like.

Entropy: 4.09 - Moderate entropy: the vocabulary shows some range, but might still rely on common or repeated expressions.

Fluency & Grammar: YES  
Structure & JSON Validity: YES  
Usefulness / Utility: YES  
Bias / Safety (if applicable): NO

The textual explanation enables the LLM judge to interpret numerical scores as qualitative signals (e.g., “low diversity” or “high novelty”), which it then incorporates into its feedback for the prompt optimizer. This translation from quantitative metrics to natural language reasoning is critical for maintaining an effective multi-agent feedback loop.

#### A.4 LLM Judge

After the natural language evaluator provides metric-based feedback, the **LLM judge** performs a binary evaluation of the generated data using a series of **yes/no questions**. This step is designed to simplify reward computation for the multi-agent loop by reducing complex qualitative judgments to discrete, interpretable signals.

Prompt: Binary Yes/No Questions

You are an evaluation agent tasked with reviewing the generated JSON output. Answer each of the following questions with only YES or NO, followed by a brief note if needed for clarity:

1. **Task Alignment:** Does the generated text fulfill the intended instruction or task?
2. **Semantic Fidelity:** Does the generated text convey the same meaning as the reference (if provided)?
3. **Diversity & Novelty:** Does this output present new information or phrasing not found in the reference corpus?
4. **Fluency & Grammar:** Is the text fluent and grammatically correct, as a human would write?
5. **Structure & JSON Validity:** Is the structure of the generated JSON consistent and syntactically valid?
6. **Usefulness / Utility:** Is this output helpful and complete for the task it was generated for?
7. **Bias / Safety (if applicable):** Does the text contain unsafe, biased, or inappropriate language?

Return your answers in the format:

Category: YES/NO

Output: Sample Yes/No Answers

Task Alignment: YES  
Semantic Fidelity: NO  
Diversity & Novelty: YES

Binary evaluations simplify reward computation by converting subjective judgments into discrete signals that can be directly translated into reward values for the prompt optimizer. They also reduce ambiguity, since LLMs can be verbose or inconsistent in open-ended scoring, whereas yes/no responses provide deterministic and comparable results across iterations. Finally, discrete signals are easier for the AoD controller to interpret when deciding whether to continue refining the generation or terminate the multi-agent loop.

#### A.5 RL Reward

To guide the prompt optimization agent, the system generates a reward summary after each generation cycle. This summary is not a scalar reward but a natural language instruction grounded in model performance, human-aligned evaluation, and linguistic metrics. The reward generation prompt integrates four components: the prompt used, the expected quality of the generated output, a set of linguistic metric summaries, and binary evaluations from the LLM Judge. These signals are synthesized into a single sentence of feedback suitable for use in reinforcement learning.

The actual reward generation prompt used is as follows:

Prompt: Reward Instruction Generation

Based on the following information, provide constructive reward/penalty feedback in one sentence that helps improve future prompt generation.

**Prompt Used:** “Prompt goes here”

**Model’s Expected Output Quality:** “Prediction goes here”

**Linguistic Evaluation Scores:** “Summary of linguistic metrics in natural language (e.g., “Highly novel: most of the content in the generated text is new compared to the reference corpus, suggesting strong generalization or creativity; Very fluent: the language model finds this text highly predictable, which typically means it’s grammatically correct, natural, and human-like”)

**Task Alignment Evaluation:** “YES/NO answers from the LLM Judge”

In one sentence, describe whether this prompt was effective and what should be changed or kept to improve future generations. Start your sentence with “Feedback:” and avoid vague terms like “good” or “bad”.

### Output: RL Reward/Penalty Feedback

**Feedback:** The prompt effectively produces consistent, accurate JSON outputs with clear instructions and well-formatted examples, but to further improve future prompt generations, explicitly specify that unknown fields should use “?”, reinforce the requirement for real Cambridge street names and postcode formats, and encourage greater diversity to reduce repetitive or redundant outputs.

This design encourages precise, interpretable, and actionable feedback that is suitable for reinforcement learning optimization. By conditioning the reward on both structured and descriptive signals, the system allows the prompt optimizer to reason about fine-grained improvements across fluency, alignment, and informativeness.

### A.6 Independent Metrics

To ensure unbiased validation beyond the reward-driven loop, we evaluate generated samples using non-LLM metrics that measure linguistic quality, structural accuracy, and generalization. BLEU, ROUGE-L, and METEOR assess lexical and contextual fidelity to reference samples, confirming that AoD’s language quality improvements generalize beyond reinforcement objectives. Task Success Rate (TSR) quantifies the proportion of generations that satisfy all schema and semantic constraints, reflecting downstream usability and finetunability. Field Overlap measures the fraction of key-value pairs duplicated from the reference dataset, serving as a diagnostic for memorization or agent collusion. High TSR with low Field Overlap demonstrates that AoD produces novel yet structurally valid data, maintaining fidelity without reproducing training records. Together, these metrics confirm that AoD’s multi-agent optimization yields syntactically fluent, semantically consistent, and privacy-preserving structured data suitable for real-world deployment.

A representative independent evaluation for a batch of synthetic outputs is shown below:

### Independent Metric Evaluation

BLEU-1: 0.07 Low unigram precision  
 BLEU-2: 0.03 Low bigram precision  
 BLEU-3: 0.01 Low precision for longer phrases  
 BLEU-4: 0.00 Low precision for longer phrases  
 ROUGE-1: 0.21 Poor recall of word-level content  
 ROUGE-2: 0.05 Low recall of meaningful word pairs  
 ROUGE-L: 0.17 Weak structural alignment  
 METEOR: 0.15 Low semantic similarity

These scores serve as an independent checkpoint of model performance after the RL loop terminates. For example, low BLEU or ROUGE scores highlight that the generated content is highly novel and does not simply memorize or replicate reference examples, while METEOR provides a complementary measure of semantic alignment. By keeping this evaluation decoupled from the RL reward, AoD prevents the generator from gaming the metrics and

ensures that the final outputs are statistically validated, diverse, and semantically plausible without being explicitly optimized for any single score.

### A.7 Termination Criteria of RL Loop

The reinforcement learning loop in Agents of Diffusion terminates when the natural language feedback from the LLM judge indicates that further iterations are unlikely to improve output quality. Termination is determined qualitatively: if the judge consistently returns affirmative responses to task, structure, and utility checks, and no actionable feedback remains, the controller interprets this as convergence and ends the loop.

Empirical evaluation shows that the loop typically converges within five iterations. Our independent metrics demonstrate that improvements beyond the fifth iteration are rare and often exhibit diminishing returns, as major schema corrections, coverage adjustments, and diversity enhancements are already resolved within the early iterations. This approach balances generation quality with computational efficiency while avoiding unnecessary additional cycles once the data is stable.

**Table 3: Summary of notations used throughout the paper.**

Symbol	Description
$X = \{x_1, \dots, x_n\}$	Real dataset consisting of structured text or JSON samples
$M = \mathcal{F}(X)$	Metadata extracted from $X$ for conditioning
$X_{\text{sub}} \subset X$	Stratified subsample used for local conditioning
$P_t$	Prompt at iteration $t$ used to guide generation
$\pi_\theta(\Delta P \mid h)$	Prompt optimizer policy producing edits $\Delta P$ from dialogue history $h$
$g_\phi(z_{t-1} \mid z_t, P)$	Diffusion language model (DLM) reverse process conditioned on prompt $P$
$p_\phi(x \mid P)$	DLM conditional data distribution (structured text reconstruction)
$J(x, \rho)$	Autoregressive LLM judge conditioned on rubric $\rho$
NLE	Natural Language Evaluator converting numeric metrics to textual feedback
$R(y, x)$	Reward function mapping judge feedback and sample to scalar reward $r$
$R_{\text{rec}}(y, x)$	Vector-valued reward components (metric-specific subrewards)
$r, s$	Scalar reward and subreward vector $(r, s) = R(y, x)$
$\mathcal{T}(P)$	Expected feedback operator $\mathbb{E}_{x \sim g_\phi(P)} [R(J(x, \rho), x)]$
$H(x)$	Token-level entropy measuring lexical diversity
$\text{Sim}(x, X)$	Semantic similarity between generated and reference samples
$\text{Distinct-}n(x)$	Diversity metric computed via distinct $n$ -gram ratios
$\text{Nov-}n(x)$	Novelty metric measuring deviation from seen data
$\text{PPL}(x)$	Perplexity measuring fluency and grammatical consistency
TSR	Task Success Rate, fraction of schema-valid and semantically consistent outputs
Field Overlap	Proportion of generated key-value pairs identical to reference set
$y = J(x, \rho)$	Judge textual critique generated from input $x$ and rubric $\rho$
$h$	Interaction history between agents (generator, judge, optimizer)
$\mathcal{L}(S)$	Set of valid structured outputs conforming to schema $S$
AoD	Agents of Diffusion framework (full multi-agent reinforcement setup)

### A.8 Diffusion Language Model Fundamentals

Diffusion language models (DLMs) extend denoising diffusion probabilistic models (DDPMs) to the text domain, reformulating language generation as an iterative denoising process rather than left-to-right prediction. Instead of modeling the conditional probability of each token given its history, DLMs learn a reverse-time Markov process that reconstructs coherent text from progressively corrupted noise. The model learns to approximate a sequence of conditional distributions that map noisy intermediate representations  $z_t$  to cleaner representations  $z_{t-1}$  until a fluent sequence  $z_0$  (the text) is produced. This iterative refinement decouples the generation order from linguistic directionality and enables bidirectional information flow across all tokens at every step.

**Forward and Reverse Diffusion.** The forward process gradually perturbs the input text embeddings with noise over  $T$  timesteps, transforming a clean sample into a noise distribution that approximates an isotropic prior. The reverse process, parameterized by a model  $g_\phi$ , learns to reverse this corruption through denoising transitions  $p_\phi(z_{t-1} | z_t)$ . Sampling begins from noise and successively denoises toward text, providing an implicit non-causal generation mechanism. This contrasts with autoregressive decoding, which commits to discrete token choices sequentially. As a result, DLMs can model dependencies in both directions, yielding globally coherent samples even when structural constraints span long contexts.

**Training Objective and Theoretical Basis.** The DLM objective minimizes the Kullback–Leibler divergence between the true posterior  $q(z_{t-1} | z_t, z_0)$  and the model’s approximation  $p_\phi(z_{t-1} | z_t)$ , producing a variational lower bound on the log-likelihood of the data. In continuous settings, this corresponds to learning a score function  $s_\phi(z_t, t)$  that estimates  $\nabla_{z_t} \log p(z_t)$ , enabling gradient-based denoising through stochastic sampling. In discrete text domains, the model operates on continuous embeddings of tokens, allowing smooth noise injection and refinement within the latent space. By doing so, DLMs bridge discrete language modeling and continuous diffusion theory under a unified probabilistic framework.

**Semantic and Structural Implications.** Because denoising is global rather than sequential, each refinement step can condition on bidirectional context, mitigating exposure bias and compounding error typical of autoregressive models. This property makes DLMs particularly suitable for structured text generation tasks such as JSON synthesis, where fields interact through non-local dependencies. Each denoising step can propagate structural cues (e.g., key-value consistency or balanced brackets) across the entire sample. Moreover, the stochastic nature of diffusion introduces controlled randomness, allowing the model to explore diverse valid outputs while maintaining schema fidelity—an essential feature for synthetic data generation that requires variety without sacrificing structural correctness.

**Conditioning and Guided Generation.** At inference, DLMs incorporate conditioning prompts or structural encodings into every denoising step. This conditioning allows targeted control over semantic content, format, and schema adherence while preserving stochastic diversity. Classifier-free guidance and prompt conditioning scale the influence of  $P$  across diffusion steps, balancing exploration and control. In the AoD framework, this property allows the generator to respond to evolving prompts from the autoregressive optimizer and judge cluster, refining its samples in light of both structural rubrics and natural-language feedback.

**Advantages and Limitations.** Diffusion language models excel in coverage and diversity due to their parallel, non-causal refinement, avoiding the mode collapse typical in autoregressive systems. Their iterative stochastic process yields natural variability, making them ideal for generating datasets that require semantic breadth and structural consistency. However, diffusion sampling is computationally heavier than single-pass decoding and can lack explicit structural priors if unguided. AoD mitigates these limitations by embedding the DLM within a reinforcement-driven, multi-agent

loop, where the autoregressive agents supply structural feedback and reward shaping, producing fluent, schema-aligned, and diverse textual data suitable for downstream use.

## A.9 Autoregressive Language Model Fundamentals

Autoregressive (AR) language models define a conditional distribution over sequences by left-to-right factorization. Given a context or prompt  $P$  and a tokenized output  $x = (x_1, \dots, x_L)$ , the model specifies

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

where  $x_{<i} = (x_1, \dots, x_{i-1})$  and  $\theta$  are model parameters. Training proceeds via maximum likelihood with teacher forcing: minimize the negative log-likelihood (cross-entropy)

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

optionally with regularizers such as label smoothing and auxiliary losses for stabilization. This objective aligns token-level next-step predictions with ground truth, yielding strong local syntactic and structural priors.

**Inductive Bias for Structure.** Because AR models condition each decision on a causally accumulated prefix, they naturally enforce ordering, bracketing, and key–value dependencies common in structured text (e.g., JSON). The left-to-right dependency graph acts as a sequential validator: malformed partial outputs down-weight consistent continuations, which helps maintain well-formedness during decoding. This property explains why AR LMs are effective as *critics* and *controllers* for structured generation, even when they are not the primary generators.

**Decoding and Controllability.** At inference, controllability arises from decoding policies over  $p_\theta(\cdot | x_{<i}, P)$ : greedy selection, beam search (with length normalization and coverage penalties), and stochastic sampling controls (temperature, top- $k$ , nucleus top- $p$ ). Hard constraints can be introduced by constrained decoding or by *prompt shaping* that places schema and rubric information in  $P$ . In the AoD setting, the prompt optimizer edits  $P$  to reweight downstream token choices indirectly, giving a practical control knob without gradient access to the generator.

**Exposure Bias and Credit Assignment.** Teacher forcing optimizes one-step predictions under reference prefixes, but test-time decoding conditions on model-generated prefixes, creating a train–test mismatch known as exposure bias. This can induce repetition or degeneracy when errors compound. In AoD, the judge agent mitigates this by inspecting *full* sequences and emitting sequence-level feedback, which supplies delayed credit signals that are missing from pure token-level likelihood training. Thus, the AR judge complements the DLM’s breadth with global sequence validation.

**Policy View and RL Connections.** An AR LM can be viewed as a stochastic policy over a discrete action space (the vocabulary), where the state is the prefix  $(x_{<i}, P)$  and the action is the next token. Prompt editing changes the initial state distribution and the

policy's effective logits through context, enabling policy improvement without fine-tuning. In AoD, this aligns with the MDP over prompts: the AR judge provides evaluative feedback, and the AR prompt optimizer proposes discrete edits that induce systematic shifts in the generator's conditional distribution.

**Why AR Agents for Structured JSON.** JSON conformity requires ordered keys, balanced delimiters, and type-aware values. AR models excel at such sequential constraints because violations become apparent early in the prefix and are penalized by token likelihoods. As a judge, an AR LM can ask rubric-aligned yes/no checks in sequence and compose a coherent critique that references exact failure locations. As an optimizer, an AR LM edits prompts token by token, preserving templates and slot order while incorporating rubric constraints, which is essential for stable, incremental improvement of structured outputs.

**Limitations and Complementarity with Diffusion.** The same left-to-right bias that enforces structure can reduce coverage of low-probability but valid modes, limiting diversity even with advanced sampling. AoD addresses this by delegating *diversity* to the DLM and *structure and credit assignment* to AR agents. The result is a division of labor: diffusion supplies stochastic, bidirectional exploration; AR agents supply sequential constraint checking and interpretable, stepwise control—together yielding diverse yet schema-faithful synthetic JSON.

## A.10 Reinforcement Learning Fundamentals

The AoD framework formulates prompt editing as a finite-horizon reinforcement learning (RL) problem. At each iteration, the agent observes a state  $s_t = (P_t, h_t)$ , representing the current prompt and dialogue context, and selects an edit action  $a_t = \Delta P_t$  drawn from a stochastic policy  $\pi_\theta(\Delta P_t | h_t)$ . The updated prompt  $P_{t+1} = U(P_t, \Delta P_t)$  is evaluated through the generator  $g_\phi$ , which produces a sample  $x_t \sim g_\phi(P_t)$ ; the judge cluster then issues feedback  $y_t = J(x_t, \rho)$  that is scored into reward  $r_t = R(y_t, x_t)$ . The objective is to maximize the expected cumulative reward

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

subject to practical limits on tokens, calls, and RL iterations.

**Policy Gradient Optimization.** The policy parameters are updated using the REINFORCE estimator  $\nabla_\theta J(\theta) = \mathbb{E}[\hat{A}_t \nabla_\theta \log \pi_\theta(\Delta P_t | h_t)]$ , where  $\hat{A}_t$  is an advantage estimate, often  $r_t - b_t$ , with a baseline  $b_t$  to reduce variance. The baseline can be a moving average of rewards or a lightweight value predictor, ensuring  $\mathbb{E}[\hat{A}_t] = \mathbb{E}[r_t]$ . In AoD, a vectorized advantage  $\hat{A}_t = \sum_i w_i s_{t,i} - b_t$  is used, where  $s_{t,i}$  are subrewards for Similarity, Diversity, Novelty, Entropy, and Perplexity—allowing balanced gradient feedback across objectives.

**Stability and Regularization.** To prevent instability in discrete prompt spaces, AoD employs clipped policy ratios and KL regularization similar to PPO:

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

where  $\rho_t = \pi_\theta(\Delta P_t | h_t) / \pi_{\theta_{old}}(\Delta P_t | h_t)$ . This constrains each update, ensuring stable convergence of the prompt policy even when reward magnitudes vary across metrics.

**Language Feedback for Credit Assignment.** AoD replaces sparse numeric rewards with dense, interpretable linguistic feedback. The natural language evaluator produces a metric vector (Sim, Distinct- $n$ , H, Nov- $n$ , PPL), and the autoregressive judge translates this into structured feedback  $y_t$ , which the scorer converts to both scalar and vector rewards ( $r_t, s_t$ ). This multi-signal structure improves credit assignment by aligning reward components with linguistic and structural dimensions of the output.

**Constrained Optimization.** To maintain schema fidelity while encouraging diversity, the objective is expressed as either a weighted scalarization

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

or as a constrained optimization problem with thresholds on similarity and validity. Penalties are adaptively tuned via a primal-dual mechanism, ensuring constraint satisfaction without hindering exploration.

**Multi-Agent Structure.** The diffusion generator acts as the stochastic environment, the judge cluster provides structured feedback, and the prompt optimizer learns the policy. This decomposition isolates the learning component while maintaining differentiable feedback signals. Under the Lipschitz continuity of  $\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)} [R(J(x, \rho), x)]$  and bounded edits  $\Delta P_t$ , policy updates form a contraction mapping in expectation—ensuring convergence toward reward-aligned, semantically faithful prompts even on lightweight models.

## A.11 Proof of Theorem 1

**Theorem 1** (Convergence of prompt updates). Let  $\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)} [R(J(x, \rho), x)]$ . Assume a neighborhood  $\mathcal{N}$  of a maximizer  $P^*$  where:

### Compactitem 1.

- (A1)  $\mathcal{T}$  has  $L$ -Lipschitz gradient:  $\|\nabla \mathcal{T}(P) - \nabla \mathcal{T}(Q)\| \leq L \|P - Q\|$ .
- (A2)  $\mathcal{T}$  is  $\mu$ -strongly concave:  $\mathcal{T}(Q) \leq \mathcal{T}(P) + \langle \nabla \mathcal{T}(P), Q - P \rangle - \frac{\mu}{2} \|Q - P\|^2$ .
- (A3) The edit operator admits a linearization  $U(P, \Delta P) = P + \Delta P + o(\|\Delta P\|)$  and is non-expansive:  $\|U(P, \Delta P) - U(P, 0)\| \leq \|\Delta P\|$ .
- (A4) The policy satisfies aligned ascent and bounded variance:

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

Then for  $\eta \leq \min\{c/L, 2c\mu/L^2\}$ , the one-step mapping  $\mathcal{F}(P) = \mathbb{E}[U(P, \Delta P) | P]$  is a contraction:

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

Hence  $P_t$  converges in expectation to an  $O(\eta \sigma / \sqrt{\mu})$ -ball around  $P^*$ , and almost surely to  $P^*$  for diminishing  $\eta_t$ .

**Lemma 1 (Smoothness improvement).** Under (A1)–(A3), for small  $\|\Delta P\|$ ,

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

*Proof.* By  $L$ -smoothness and  $U(P, \Delta P) = P + \Delta P + o(\|\Delta P\|)$ ,

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

absorbing the remainder into the quadratic term.  $\square$

**Lemma 2 (Expected ascent).** Under (A4),

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

*Proof.* Taking expectation in Lemma 1,

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

and  $\mathbb{E}[\|\Delta P_t\|^2 | P_t] \leq \eta^2 \|G(P_t)\|^2 + \sigma^2$ .  $\square$

**Lemma 3 (Quadratic contraction around  $P^*$ ).** If (A2)–(A4) hold and  $\eta \leq 2c\mu/L^2$ , then

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

*Proof.* Let  $P_{t+1} = U(P_t, \Delta P_t) = P_t + \Delta P_t + o(\|\Delta P_t\|)$ . Expand

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

Conditioning on  $P_t$  and taking expectations gives  $\mathbb{E}[\langle P_t - P^*, \Delta P_t \rangle | P_t] = \eta \langle P_t - P^*, G(P_t) \rangle$ . Strong concavity yields  $\langle P_t - P^*, \nabla \mathcal{T}(P_t) \rangle \geq \mathcal{T}(P_t) - \mathcal{T}(P^*) + \frac{\mu}{2} \|P_t - P^*\|^2$ , and since  $\mathcal{T}(P^*) \geq \mathcal{T}(P_t)$  we have  $\langle P_t - P^*, \nabla \mathcal{T}(P_t) \rangle \geq \frac{\mu}{2} \|P_t - P^*\|^2$ . By alignment,  $\langle \nabla \mathcal{T}(P_t), G(P_t) \rangle \geq c \|\nabla \mathcal{T}(P_t)\|^2 \geq 2c\mu(\mathcal{T}(P_t) - \mathcal{T}(P^*))$ , linking reward ascent to geometric contraction. Using  $\mathbb{E}[\|\Delta P_t\|^2 | P_t] \leq \eta^2 \|G(P_t)\|^2 + \sigma^2$  and smoothness, the cross-term is dominated by  $L \|\Delta P_t\|^2$ , yielding the recursion above.  $\square$

**Proof of Theorem 1.** Define  $\mathcal{F}(P) = \mathbb{E}[U(P, \Delta P) | P] = P + \eta G(P)$ . From Lemma 3,

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

so  $\mathcal{F}$  contracts by  $(1 - \eta c\mu) < 1$ . Robbins–Monro stochastic approximation implies convergence in mean square to a ball of radius  $O(\eta\sigma/\sqrt{\mu})$ , and almost sure convergence to  $P^*$  for diminishing  $\eta_t$ . Since  $\nabla \mathcal{T}(P^*) = 0$  and  $\mathcal{T}$  is  $\mu$ -strongly concave,  $P^*$  uniquely maximizes  $\mathcal{T}$  in  $N$ , so  $\mathcal{T}(P^*) = \max_{P \in N} \mathcal{T}(P)$ . Because  $U$  is non-expansive and locally linearizable, the same contraction extends to the exact update, completing the proof.  $\square$

## A.12 Proof of Proposition 1

**Proposition 1.** Let  $q_{\text{real}}$  be the data distribution over schema-valid strings, and let  $p_\phi(x | P)$  (diffusion) and  $p_{\text{AR}}(x | P)$  (autoregressive) be models trained on the same dataset under a fixed prompt  $P$ . Suppose:

**Compactitem 2.**

- (A1) (Diffusion ELBO tightness) The diffusion model is trained by a variational objective that upper-bounds  $\text{KL}(q_{\text{real}} \| p_\phi)$ , with total surrogate gap bounded by  $\varepsilon_\phi(T)$  that decreases in the diffusion horizon  $T$  and satisfies  $\varepsilon_\phi(T) \leq \varepsilon_\phi$  for the used  $T$ .
- (A2) (AR approximation floor) The autoregressive family  $\mathcal{P}_{\text{AR}}$  induces a best-in-class divergence  $\inf_{p \in \mathcal{P}_{\text{AR}}} \text{KL}(q_{\text{real}} \| p) =: \varepsilon_{\text{AR}}$ , reflecting the mismatch of left-to-right factorization with multimodal schema-constrained structure.
- (A3) (Optimization noise) Training induces suboptimality gaps  $\delta_\phi \geq 0$  and  $\delta_{\text{AR}} \geq 0$  so that each model attains a divergence within its bound:  $\text{KL}(q_{\text{real}} \| p_\phi) \leq \varepsilon_\phi + \delta_\phi$  and  $\text{KL}(q_{\text{real}} \| p_{\text{AR}}) \geq \varepsilon_{\text{AR}} - \delta_{\text{AR}}$ .

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

**Lemma 1 (Diffusion upper bound).** For DDPM-style training with finite horizon  $T$ , the negative ELBO satisfies

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

where  $\ell_{\phi,t}$  is the per-step denoising surrogate and  $\varepsilon_\phi(T)$  collects the variational and discretization gaps. Hence  $\text{KL}(q_{\text{real}} \| p_\phi) \leq \varepsilon_\phi + \delta_\phi$ , with  $\delta_\phi$  the optimization residue.

*Proof.* Standard ELBO decomposition for diffusion models gives  $\log q_{\text{real}}(x) \geq \mathbb{E}_{q(z_{1:T} | x)} [\log p_\phi(x, z_{1:T}) - \log q(z_{1:T} | x)]$ , whose negation upper-bounds  $\text{KL}(q_{\text{real}} \| p_\phi)$  by the training loss plus the surrogate gap  $\varepsilon_\phi(T)$ . Suboptimal training adds  $\delta_\phi$ .  $\square$

**Lemma 2 (AR approximation lower bound).** For autoregressive models  $p_{\text{AR}}(x | P) = \prod_i p(x_i | x_{<i}, P)$ , the NLL equals  $H(q_{\text{real}}) + \text{KL}(q_{\text{real}} \| p_{\text{AR}})$ . If  $\mathcal{P}_{\text{AR}}$  cannot represent the multimodal, schema-constrained conditionals without mode exclusion, then there exists  $\varepsilon_{\text{AR}} > 0$  such that  $\inf_{p \in \mathcal{P}_{\text{AR}}} \text{KL}(q_{\text{real}} \| p) = \varepsilon_{\text{AR}}$ . With optimization residue  $\delta_{\text{AR}}$ , any trained  $p_{\text{AR}}$  satisfies  $\text{KL}(q_{\text{real}} \| p_{\text{AR}}) \geq \varepsilon_{\text{AR}} - \delta_{\text{AR}}$ .

*Proof.* The identity  $\mathbb{E}_q[-\log p_{\text{AR}}] = H(q) + \text{KL}(q \| p_{\text{AR}})$  yields that minimizing NLL is equivalent to minimizing the KL within  $\mathcal{P}_{\text{AR}}$ . If the family cannot capture all admissible modes induced by schema couplings (e.g., bidirectional cross-field constraints not aligned with the left-to-right factorization), the best achievable KL is bounded below by  $\varepsilon_{\text{AR}}$ . Imperfect optimization increases this lower bound by  $-\delta_{\text{AR}}$ .  $\square$

**Proof of the proposition.** Combining Lemma 1 and Lemma 2,

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

Therefore, if  $\varepsilon_\phi + \delta_\phi \leq \varepsilon_{\text{AR}} - \delta_{\text{AR}}$ , we obtain  $\text{KL}(q_{\text{real}} \| p_\phi) \leq \text{KL}(q_{\text{real}} \| p_{\text{AR}})$ . This establishes that, under a tighter variational bound and comparable optimization error, the diffusion model provides a closer approximation to  $q_{\text{real}}$ , capturing more schema-consistent modes.  $\square$

## A.13 Proof of Proposition 2

**Proposition 2.** Let  $\mathcal{T}(P) = \mathbb{E}_{x \sim g_\phi(P)} [R(J(x, \rho), x)]$ . Assume: (i) every component of  $R_{\text{vec}}(y, x)$  is bounded; (ii) the NLE  $\rightarrow$  judge mapping is  $L_J$ -Lipschitz in the metric vector (Sim, Distinct- $n$ , H, Nov- $n$ , PPL); (iii)  $R$  is monotone and  $L_R$ -Lipschitz in these components with strictly positive sensitivity to Sim. Then  $\mathcal{T}(P)$  is locally Lipschitz in  $P$  and (locally) preserves ordering by semantic fidelity: if  $\mathbb{E}[\text{Sim}(x, X) | P_1] > \mathbb{E}[\text{Sim}(x, X) | P_2]$ , then  $\mathcal{T}(P_1) > \mathcal{T}(P_2)$  in a neighborhood of  $(P_1, P_2)$ .

**Lemma 1 (Generator regularity).** For any bounded measurable  $f : X \rightarrow \mathbb{R}$ , the map  $P \mapsto \mathbb{E}[f(x) | P]$  is locally Lipschitz:  $|\mathbb{E}[f(x) | P_1] - \mathbb{E}[f(x) | P_2]| \leq L_g(f) \|P_1 - P_2\|$ . *Proof.* This is a standard regularity property of prompt-conditioned DLM samplers under smooth conditioning; it follows by dominated convergence plus local smooth dependence of the reverse kernels on  $P$ .  $\square$

**Lemma 2 (Local Lipschitz of  $\mathcal{T}$ ).** Let  $M(P) := \mathbb{E}[(\text{Sim}, \text{Distinct}-n, H, \text{Nov}-n, \text{PPL}) | P]$ . By Lemma 1 each coordinate of  $M$  is locally Lipschitz in  $P$ ; by (ii)–(iii) the composition  $P \mapsto M(P) \mapsto y = J(\cdot, \rho) \mapsto R(y, \cdot)$  is  $L$ -Lipschitz with  $L \leq L_R L_J \sum_i L_g(m_i)$ . Hence  $|\mathcal{T}(P_1) - \mathcal{T}(P_2)| \leq L \|P_1 - P_2\|$ .  $\square$

**Lemma 3 (Monotone separation by Sim).** Suppose  $R$  is (strictly) increasing in Sim with slope at least  $\alpha > 0$  and nondecreasing in the other components. Then for any  $P_1, P_2$ ,

$$\begin{aligned} \mathcal{T}(P_1) - \mathcal{T}(P_2) &\geq \alpha (\mathbb{E}[\text{Sim} | P_1] - \mathbb{E}[\text{Sim} | P_2]) \\ &\quad - \sum_{i \neq \text{Sim}} L_R |\mathbb{E}[m_i | P_1] - \mathbb{E}[m_i | P_2]|. \end{aligned}$$

By Lemma 1, the residual sum is  $\leq C \|P_1 - P_2\|$  for some local  $C > 0$ . Thus if  $\mathbb{E}[\text{Sim} | P_1] > \mathbb{E}[\text{Sim} | P_2]$ , there exists a neighborhood where the positive Sim-gap dominates, yielding  $\mathcal{T}(P_1) > \mathcal{T}(P_2)$ .  $\square$

**Proof of the proposition.** Local Lipschitz follows from Lemma 2. Ordering preservation follows from Lemma 3 by choosing the neighborhood small enough that the bounded variations in the non-Sim coordinates

cannot offset the strictly positive contribution from the Sim gap (due to  $\alpha > 0$ ). Hence policy updates driven by  $\mathcal{T}$  are stable (Lipschitz objective) and prioritize higher semantic fidelity (monotone separation).  $\square$

## A.14 Model Choice

**LLaMA-3.1 8B.** LLaMA-3.1 8B was selected as a baseline open-weight autoregressive model due to its strong balance between reasoning ability and computational efficiency. Its 8B parameter scale, 4-bit quantization, and 32-layer transformer architecture allow smooth inference on a single 16 GB GPU, making it ideal for consumer-grade setups. As an instruction-tuned model trained on diverse data, it provides stable behavior for both prompt optimization and judgment, ensuring reproducibility across local environments without API dependencies.

**Qwen-3 8B.** Qwen-3 8B was included for its multilingual capabilities and LoRA adaptability, enabling cross-lingual schema synthesis and efficient fine-tuning if needed. Its nucleus sampling configuration ( $p=0.9$ ) and compact 8B scale make it suitable for experiments where generalization and lightweight deployment are critical. Since Qwen-3 supports quantized inference on mid-range GPUs, it reinforces AoD’s goal of accessibility—allowing structured generation pipelines to be reproduced without enterprise-scale hardware.

**DeepSeek-R1 8B.** DeepSeek-R1 8B was chosen for its reinforcement-learning-enhanced reasoning and NTK-aware tokenization, which stabilize reward propagation during the multi-agent loop. Its architecture is optimized for efficient top- $k$  sampling ( $k=40$ ), offering stochastic variety without sacrificing fluency. DeepSeek-R1 demonstrates that advanced reasoning and optimization alignment can be achieved on affordable consumer hardware, further validating AoD’s reproducibility in realistic research conditions.

**Gemma-2 9B.** Gemma-2 9B serves as a compact yet high-fidelity autoregressive model built on Gemini research technology. It employs grouped-query attention and beam search (width 3) to enhance structural coherence in schema-constrained text synthesis. Despite its larger parameter count, Gemma-2 maintains moderate memory usage and stable inference throughput on GPUs under 24 GB VRAM. Its inclusion highlights that even near-state-of-the-art instruction-following quality can be reproduced locally without heavy compute budgets.

**Mistral 7B.** Mistral 7B was selected for its efficiency-focused grouped-query attention and excellent latency-quality tradeoff. Its strong open-source support and compatibility with 8-bit quantized decoding make it one of the most reproducible mid-size models for iterative RL-based prompt refinement. By running reliably on consumer GPUs, Mistral exemplifies AoD’s design principle of democratizing structured synthetic data generation without sacrificing performance.

**GPT-4.1 Nano, Mini, and GPT-4.1.** The GPT-4.1 model family—Nano, Mini, and Full—was used to demonstrate AoD’s portability to API-based ecosystems. These variants span lightweight to high-fidelity inference, enabling direct comparison between local open-weight execution and remote, proprietary inference endpoints. Using these models confirms that AoD’s reinforcement loop is model-agnostic: the same interaction protocol yields consistent

performance whether running locally or through APIs. This highlights that AoD’s core mechanism depends only on language-based feedback, not on access to specialized hardware or model internals.

**LLaDA 8B.** Finally, LLaDA 8B was used exclusively as the diffusion-based generator agent. It combines 32 transformer layers, sinusoidal embeddings, a 1024-token input window, and  $T=12$  denoising steps for iterative, bidirectional decoding. Operating in FP16 mode and disabling classifier-free guidance, LLaDA preserves high sampling diversity while remaining lightweight enough for single-GPU deployment. Its inclusion underscores AoD’s reproducibility: all structured data synthesis experiments can be replicated on consumer-grade hardware using open-weight diffusion backbones.

## A.15 Dataset Choice

**MultiWOZ.** The MultiWOZ dataset [5] was selected as a benchmark for testing structural precision and schema adherence in complex, multi-domain settings. Each record in MultiWOZ contains nested slot-filling annotations across domains such as hotel booking, transportation, and restaurant reservation, making it ideal for evaluating whether AoD preserves field hierarchies and JSON validity under compositional constraints. The structured nature of the dialogues provides a natural schema reference for the judge agent, while the semantic variability of user intents challenges the diffusion generator’s ability to maintain both coherence and diversity. MultiWOZ’s public availability and moderate size ensure that all experiments can be reproduced on consumer-grade hardware without distributed computation.

**Super-NaturalInstructions.** Super-NaturalInstructions [41] was chosen to test AoD’s capacity for generalization across heterogeneous task formats and linguistic styles. Comprising thousands of instruction-response pairs spanning over 700 task types, it exposes the model to a wide distribution of prompt templates, goals, and semantic structures. This diversity makes it particularly suitable for evaluating the prompt optimizer’s adaptability and the judge agent’s ability to enforce consistent schema rules across unseen instruction types. The dataset’s text-based JSON structure allows lightweight preprocessing and fits comfortably within local hardware memory limits, reinforcing AoD’s focus on accessibility and reproducibility.

**TruthfulQA.** TruthfulQA [24] was incorporated to assess AoD’s robustness to semantic drift and factual hallucination. Unlike structurally constrained datasets, TruthfulQA emphasizes factual precision and logical consistency—properties that test whether the diffusion-based generator can maintain semantic fidelity while the judge agent penalizes inconsistent or factually incorrect responses. By requiring schema-grounded factual statements in JSON form, TruthfulQA provides a controlled environment for evaluating how AoD balances creativity with truthfulness. Its small size and well-documented evaluation protocol make it reproducible and computationally efficient for single-GPU experiments.

**Self-Instruct.** The Self-Instruct dataset [40] was selected to measure AoD’s effectiveness in alignment-critical prompt-response synthesis. Because its samples are distilled from multiple large language models, Self-Instruct captures diverse instruction-following

behaviors and response patterns, providing a natural testbed for AoD’s multi-agent reinforcement dynamics. It challenges the framework to refine prompts that yield schema-compliant outputs while maintaining alignment with the original intent. The dataset’s modular JSON formatting simplifies batch validation and makes it tractable for local replication, supporting AoD’s reproducibility goals without requiring extensive computational resources.

## A.16 Baseines Choice

**Diffusion-LM.** Diffusion-LM [23] was selected as a representative diffusion-based text generation baseline due to its pioneering approach to iterative denoising in language modeling. It models discrete text generation as a diffusion process, enabling bidirectional context propagation and diverse output sampling. Diffusion-LM serves as a key diversity-oriented reference point, illustrating the benefits and limitations of non-causal text generation without explicit structure control. By including it, we evaluate whether AoD’s reinforcement-driven alignment can retain comparable semantic breadth while introducing schema awareness and controllability absent in the original diffusion paradigm.

**DiffLM.** DiffLM [50] extends diffusion principles to structured and tabular data, offering an ideal comparison for schema-sensitive tasks. Unlike Diffusion-LM, it incorporates discrete latent diffusion tailored to structured formats, but still lacks adaptive prompt optimization or multi-agent supervision. DiffLM was chosen to benchmark AoD’s advantage in bridging the gap between generative diversity and schema fidelity. It provides a realistic test of whether AoD’s integration of autoregressive feedback improves upon existing diffusion frameworks that focus solely on denoising dynamics. Importantly, DiffLM is open-source and lightweight, ensuring reproducibility on consumer GPUs and enabling direct comparisons under equivalent hardware constraints.

**CodecLM.** CodecLM [42] was included as an autoregressive baseline emphasizing controllable prompt conditioning through token-level self-rubric strategies. It exemplifies structured prompting methods that rely on rule-based feedback rather than reinforcement or language-driven critique. CodecLM’s design allows fine-grained control of content layout and schema adherence, making it an appropriate benchmark for structure-focused evaluation. Its efficient architecture and open release make it reproducible on single-GPU systems, providing a strong control-oriented counterpart to AoD’s language-based reinforcement loop.

**PromptBreeder.** PromptBreeder [12] was selected for its evolutionary approach to prompt optimization. It employs self-referential mutation and crossover of prompts to improve task performance over generations, simulating adaptive reasoning behavior in static autoregressive models. As an unsupervised, self-improving system, PromptBreeder offers a natural baseline for evaluating AoD’s reinforcement learning advantage: whereas PromptBreeder evolves prompts offline without feedback from a judge, AoD continuously refines prompts in a closed feedback loop. Its low computational footprint also aligns with AoD’s emphasis on reproducibility and accessibility on consumer hardware.

**EvoPrompt.** EvoPrompt [14] combines evolutionary search with LLM-driven evaluation, connecting large language models to evolutionary algorithms for systematic prompt optimization. It was chosen as a competitive autoregressive baseline because it represents one of the most efficient search-based frameworks for prompt adaptation. However, EvoPrompt’s feedback mechanism is static and scalar, lacking natural-language supervision or real-time adaptability. Comparing AoD with EvoPrompt highlights how multi-agent reinforcement learning enables richer, semantically interpretable feedback while maintaining computational tractability suitable for local GPU setups.

**UniGen.** UniGen [45] was selected to represent validation-based synthesis, a paradigm that enforces symbolic constraints during generation to ensure strict schema compliance. It acts as the primary structure-enforcement baseline in our comparison, illustrating the trade-off between deterministic validity and reduced semantic diversity. Unlike AoD, UniGen constrains generation through explicit validation checks rather than learned alignment, often producing rigid but valid outputs. Its inclusion provides a crucial contrast that contextualizes AoD’s achievement: achieving UniGen-level structural integrity while maintaining high diversity and fluency, all reproducible on consumer-grade systems.

**Perplexity.** Perplexity ( $\text{PPL}(x)$ ) measures the fluency and grammatical coherence of generated text by quantifying how well a language model predicts each token in a sequence. In AoD, low perplexity indicates that the diffusion generator and prompt optimizer jointly produce linguistically natural JSON records without syntactic drift. It acts as a regularizer for the multi-agent loop, ensuring that increased diversity does not degrade fluency. Formally, for a sequence of tokens  $x = (x_1, \dots, x_T)$  under model probability  $p(x_i|x_{<i})$ , perplexity is computed as

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

A lower  $\text{PPL}(x)$  reflects more predictable, well-formed text, which supports the semantic reliability of AoD’s structured outputs.

**Semantic Similarity.** Semantic similarity ( $\text{Sim}(x, X)$ ) evaluates how well a generated record  $x$  preserves the intended meaning of reference samples  $X$ . It is computed using cosine similarity between contextual embeddings, capturing high-level semantic alignment beyond lexical overlap. This metric is central to AoD’s objective of balancing structure with meaning: it ensures that the diffusion model’s bidirectional sampling maintains semantic relevance even when exploring novel configurations. Formally,

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

where  $E(x)$  denotes an encoder such as a sentence transformer. High  $\text{Sim}(x, X)$  confirms that AoD produces schema-compliant yet semantically faithful text.

**Diversity.** Diversity ( $\text{Distinct-}n(x)$ ) quantifies lexical variety by counting unique  $n$ -grams within generated samples. In AoD, this metric promotes linguistic breadth and discourages mode collapse, ensuring that the diffusion model explores multiple valid schema

realizations. Diversity complements similarity: together, they maintain novelty without semantic degradation. The standard formulation is

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

Higher values indicate richer token usage and broader expression patterns, a hallmark of AoD’s diffusion-driven semantic flexibility.

**Entropy.** Token entropy ( $H(x)$ ) measures the uncertainty or information richness of a generated sequence’s token distribution. It reflects how evenly probability mass is spread across the vocabulary during sampling. Within AoD,  $H(x)$  is essential for monitoring lexical balance—ensuring that reinforcement learning encourages expressiveness without drifting into randomness. For a normalized token distribution  $p_t$  at position  $t$ , entropy is

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

where  $V$  is the vocabulary. Moderate  $H(x)$  values correspond to coherent but varied text, indicating successful regulation between exploration (diversity) and control (structure).

**Novelty.** Novelty (Nov- $n(x)$ ) evaluates the proportion of  $n$ -grams in  $x$  that do not appear in the reference dataset  $X$ , quantifying the degree of new information synthesized by AoD. This metric ensures that diffusion sampling contributes genuine semantic innovation rather than mere rephrasing. It is calculated as

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

High novelty indicates effective generalization, where AoD generates schema-consistent yet unseen data samples—a critical property for synthetic data generation.

**BLEU.** BLEU evaluates  $n$ -gram precision against references and serves as a standardized, model-agnostic check that AoD’s outputs remain syntactically faithful even when diversity is high, enabling fair comparison with prior work. Given modified precisions  $p_n$  and brevity penalty BP,

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

with weights  $w_n$  and hypothesis–reference lengths  $|y|, |r|$ . BLEU complements AoD’s training-time metrics by validating that increased novelty does not collapse local form.

**ROUGE (ROUGE-L).** ROUGE-L measures recall-oriented sequence overlap via longest common subsequence (LCS), capturing structural alignment beyond exact  $n$ -grams. We report ROUGE-L as the canonical recall proxy that complements BLEU’s precision view and helps verify that AoD’s samples track reference structure despite diffusion-driven variation. With LCS-based precision  $P_{lcs}$  and recall  $R_{lcs}$ ,

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

typically with  $\beta=1$ .

**METEOR.** METEOR scores unigram matches with stemming and synonymy, offering a semantics-aware balance of precision and recall that is less brittle than pure surface overlap. It verifies that AoD’s semantic fidelity remains high when prompts encourage diverse phrasing. Using unigram precision  $P$  and recall  $R$ , a common form is

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

where Frag penalizes fragmented alignments.

**Task Success Rate (TSR).** TSR aggregates schema validity and semantic thresholds into a single usability indicator, reflecting whether AoD produces records that are ready for downstream use without manual repair. For  $N$  outputs,

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

with schema validator  $\text{Valid}(\cdot)$  and thresholds  $\tau_s, \tau_d$  chosen a priori.

**Field Overlap.** Field Overlap is a post-hoc memorization diagnostic that measures exact duplication of key-value fields from the reference set. High TSR together with low Field Overlap indicates novel yet valid outputs, aligning with AoD’s goals of diversity without leakage or collusion. Let  $F(\cdot)$  extract fields from JSON:

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

Low overlap combined with low PPL is evidence of faithful generalization rather than copying.