Knowing but Not Doing: Discourse-level Constraints Adherence of Large Language Models

Anonymous ACL submission

Abstract

Instruction following in large language models (LLMs) is commonly evaluated at the sentence or surface level, leaving discourse-level constraints underexplored. We study discourse-level instruction following by formalizing it as a structured constraint adherence problem and constructing both explicit and implicit benchmarks for evaluation. Experiments are conducted across models of different scales, revealing a systematic recognition–adherence gap: models can identify discourse constraints but frequently fail to adhere to them during generation with specific failure modes. This failure stems from a fundamental misalignment between standard training objectives and global structural integrity. Importantly, models show substantially improved adherence when discourse-level structural objectives are explicitly optimized.

1 Introduction

Instruction following in large language models (LLMs) is typically framed as the ability to generate content that adheres to a given set of constraints (Yin et al., 2023; Wen et al., 2024). Prior work has primarily investigated two broad types of constraints: those over *what to say*, such as factual correctness or task completion, and those over *how to say it*, including style, tone, format, or persona (Jiang et al., 2024; Li et al., 2024; Harada et al., 2025). Across these settings, constraints are implicitly assumed to be local (Dubois et al., 2024; Bai et al., 2024a), in the sense that adherence can be assessed without requiring global structural verification.

However, many real-world writing tasks impose constraints beyond sentence-level semantics or style, operating at the discourse level to govern how sentences are organized, ordered, and related across an entire text. For example, a scientific explanation typically progresses from background and motivation to method and implications. These requirements define *discourse-level constraints*: non-local structural conditions on the generation trajectory that cannot be reduced to, or verified at, the level of individual sentences (Figure 1).

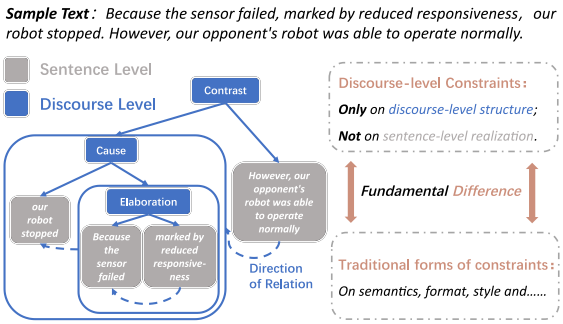

Figure 1: This shows that based on Rhetorical Structure Theory, natural language texts can be separately analyzed into the discourse level and the sentence level. Discourse-level constraints target the discourse-level structure, differing fundamentally from traditional forms of constraints.

Despite the importance of discourse-level constraints, current instruction-following research pays little explicit attention to them. This is partly because treating sentence-level semantics and discourse structure as separable targets is empirically counterintuitive, although linguistics has long established them as distinct levels (Grosz and Sidner, 1986; Mann and Thompson, 1988); as a result, existing datasets often entangle discourse structure with semantic content (Wen et al., 2024; Bai et al., 2024b), making it unclear whether apparent success reflects genuine discourse-structural adherence or merely faithful generation (Figure 1). Moreover, because evaluation signals are predominantly semantic and local (Dubois et al., 2024; Jiang et al., 2024), violations of discourse-level constraints can easily go undetected even when the generated text appears coherent and high-quality. For this reason, the effectiveness of large language

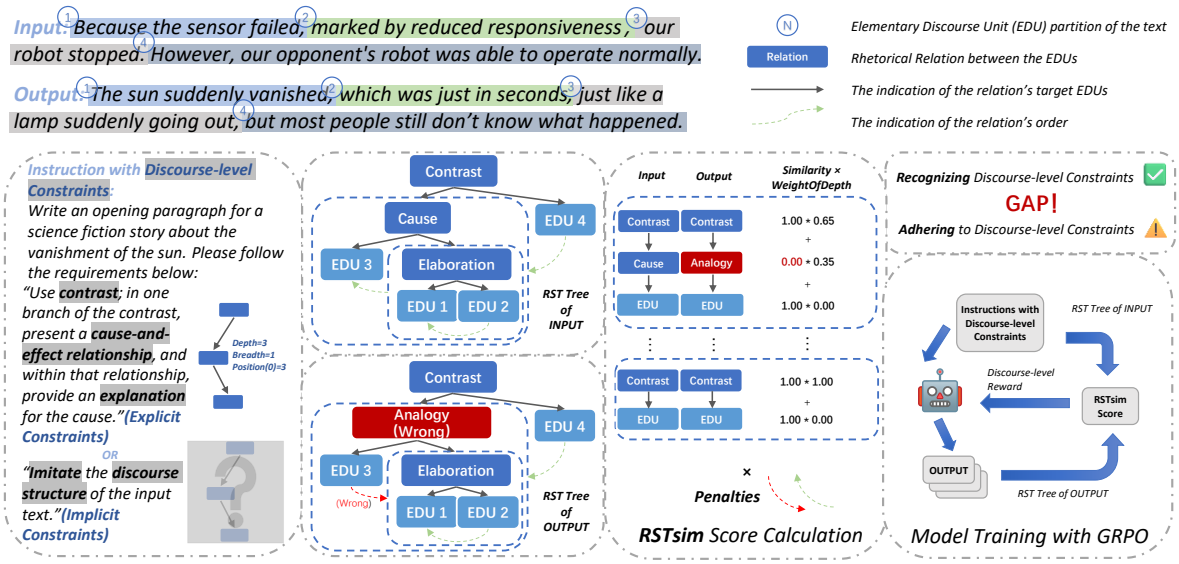

Figure 2: The framework of our study. We define the discourse-level constraints and design methods to use them to build instruction datasets.Then, we design an automatic method RSTsim as both an evaluation metric and the reward. Finally, we use GRPO to train a model.

models in following discourse-level constraints is difficult to assess.

To fill these gaps, inspired by Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Carlson et al., 2001), we study instruction following under *discourse-level constraints*. RST posits that a text can be described by discourse structure composed of hierarchically organized spans connected by rhetorical relations, which are independent of sentence realization. This aligns well with the concept of discourse-level constraints. Through extensive experiments, we show that LLMs can reliably *recognize* discourse-level constraints, yet often fail to *adhere to* them during generation, exhibiting systematic and recurring failure modes. This recognition–adherence gap is not attributable to surface fluency, decoding strategies, or prompt sensitivity, but reflects a deeper structural limitation. Specifically, autoregressive language models are trained to optimize next-token likelihood under local context (Radford and Narasimhan, 2018), without any explicit supervision on global structural correctness. Under this formulation, increasing model scale or refining prompts does not directly address the source of the problem. Within our framework, discourse structure is introduced as an explicit training signal, allowing generation behavior to be guided by structural correctness rather than local signals only (Figure 2).

Our contributions are fourfold: (i) to the best of our knowledge, our work is the **first** to introduce discourse-level instruction following as a problem setting that explicitly distinguishes constraint recognition from constraint adherence; (ii) we propose a principled dataset construction methodology that instantiates this setting under both explicit and implicit discourse constraints, enabling controlled and reproducible evaluation; (iii) we develop an automatic metric for measuring discourse-structure similarity independently of lexical realization; and (iv) we present a unified research paradigm that connects diagnosis, evaluation, and optimization, and demonstrate its effectiveness by applying reinforcement learning with discourse-level rewards to reduce the recognition–adherence gap.

2 Related Work

Instruction-following benchmarks such as FLAN (Chung et al., 2024), Super-NaturalInstructions (Wang et al., 2022), and AlpacaEval (Dubois et al., 2024) primarily evaluate semantic correctness or stylistic alignment. MT-Bench (Bai et al., 2024a), LongBench (Bai et al., 2024b) and ComplexBench (Wen et al., 2024), extend evaluation to multi-turn, long-context or complex-constraint settings. While many instructions implicitly encode discourse organization, these benchmarks neither formalize discourse structure as an explicit constraint nor assess adherence to non-local structural requirements.

In **natural language generation**, discourse structure has been explored as a means to guide text generation. Systems such as RST-based gen-

erators use predefined discourse trees to condition or guide generation, aiming to improve coherence or planning (Adewoyin et al., 2022). However, this line of work treats discourse structure as part of the generation process itself, rather than as an instruction-level constraint whose recognition and adherence can be independently evaluated.

Recent work on **constrained generation** has further exposed rule-following failures in LLMs when explicit structural directives compete with semantic priors. For example, Align to Structure (Kim et al., 2026) proposes a structural alignment method combining linguistic frameworks and fine-grained rewards to address poor hierarchical planning in long text generation.

Building on this direction, subsequent efforts have begun to isolate and evaluate **structural adherence** more directly. StructFlowBench (Li et al., 2025) introduces structural flow modeling for multi-turn instruction following and evaluates whether models preserve discourse dependencies across dialogue turns. Meanwhile, RST-LoRA (Liu and Demberg, 2024) incorporates rhetorical structure information into parameter-efficient fine-tuning. Although these approaches move toward explicit modeling of discourse structure, they remain task-specific and fragmented across evaluation and adaptation settings, leaving a unified framework for discourse-level instruction following largely unexplored.

3 Discourse-Level Constraints

3.1 Definition

Inspired by Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Carlson et al., 2001), we define *discourse-level constraints* as *structural constraints* governing the global organization of generated text.

Following the perspective of RST, discourse-level constraints specify a target discourse structure $\mathcal{T}$ that constrains how content units should be organized and related, rather than prescribing sentence-level wording. Formally, we represent $\mathcal{T}$ as a *discourse constraint tree*: a rooted tree $\mathcal{T} = (V, E, R)$ where each node $v \in V$ corresponds to a discourse span; each edge $e \in E$ encodes a parent–child hierarchical relation; and $R(e)$ denotes the rhetorical relation labels (e.g., Elaboration, Cause, and Contrast).

According to RST, relations in $\mathcal{R}$ can be binary or n-ary, with certain relations additionally imposing ordering constraints over discourse spans. Constraints are specified in two forms: **Explicit constraints**: $\mathcal{T}$ is directly specified through explicitly described constraints. **Implicit constraints**: $\mathcal{T}$ is implicitly provided by the discourse structure of a reference text Y_{ref}. Both of these essentially specify discourse structure trees.

This abstraction inherits the core structural assumptions of RST while remaining agnostic to linguistic realization details, enabling it to serve as an operational representation for discourse-level control.

3.2 Task Formulation

The core task decomposes into two subtasks:

Constraint Recognition: Given an instruction specifying explicit constraints or a reference text Y_{ref} for implicit constraints, recognize the target discourse constraint tree $\mathcal{T}$ from the instruction or the text and output it in parenthetical tree notation.

Constraint Adherence: Given input X, target constraint tree $\mathcal{T}$, and generation model P_θ, generate output Y such that its induced discourse tree T_Y satisfies $\mathcal{T}$. Adherence requires preservation of the hierarchical parent-child relations in $\mathcal{T}$, faithful realization of all rhetorical relations $R(e)$ for edges $e \in E$, and compliance with applicable ordering constraints. Crucially, constrained spans need not be surface-adjacent provided the logical discourse organization is preserved.

4 Discourse-level Constraints Recognition and Adherence in LLMs

In this section, we construct datasets and evaluation metrics for discourse-level constraint recognition and adherence, evaluate large language models under both explicit and implicit settings, analyze the underlying causes of their observed behaviors, and make corresponding optimizations to address relevant issues.

4.1 Data Construction

In practical application scenarios, there are both tasks with direct requirements and tasks involving imitation based on reference materials. Thus, we still divide the dataset into explicit constraints and implicit constraints to align with the needs of real-world applications. Our dataset contains 1,500 explicit data points and 800 implicit data points. The data examples are provided in Appendix B.

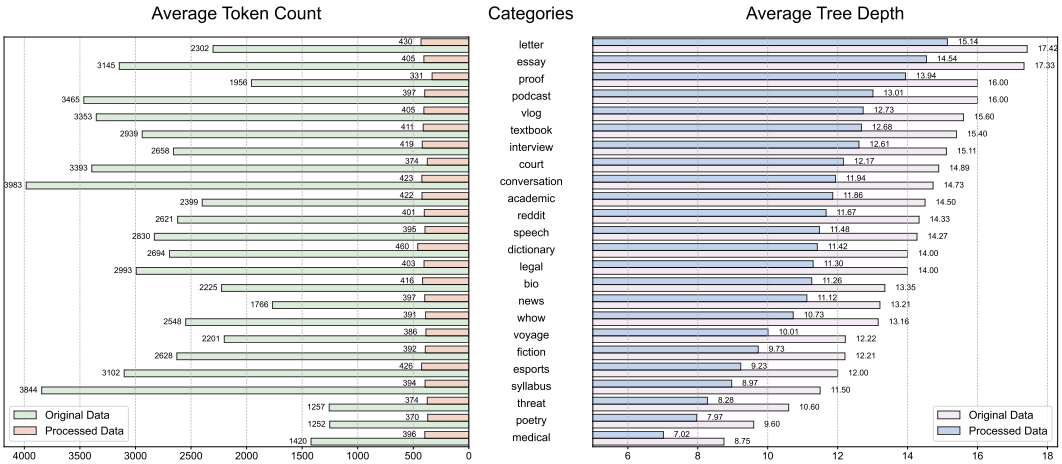

Figure 3: The average token count and RST tree depth of documents in the original GUM dataset and those after discourse-aware segmentation are grouped by categories (abbreviated).This segmentation enables the Standard-Parseval score (full) of our RST parser on the GUM dataset to jump from 55.2 to 67.0, a remarkably high value.

4.1.1 Explicit Constraints Data

Since no existing benchmark isolates discourse-level explicit constraints, we construct a synthetic dataset in which each instruction explicitly specifies a target discourse structure. Each instruction is generated by controlling three parameters that jointly define the topology of a de-rooted constraint tree:

Constraint Tree Depth (D): the maximum nesting depth of the constraint tree, capturing hierarchical complexity. **Constraint Tree Breadth (B):** the number of top-level constraint branches that capture sequential or ordering complexity. **Constraint Tree Nest Position Function ($P(i)$):** where i denotes the relative position along the breadth dimension, and $P(i)$ specifies the local nest depth at position i, determining the detailed shape of the constraint tree. The instructions are automatically generated from a repository of preset topics by enumerating combinations of $(D, B, P(i))$.

Due to the combinatorial diversity of parameter settings, the procedure yields thousands of structurally distinct instructions. Each instruction is paired with a gold discourse skeleton that represents the expected rhetorical structure of a correct response. This property enables exact-structure evaluation and also allows the dataset to be directly reused in reinforcement learning stages. Additionally, we designed five types of prompts to mitigate bias stemming from suboptimal prompt quality.

4.1.2 Implicit Constraints Data

For implicit constraint tasks, high-quality and annotated reference texts are indispensable. To evaluate implicit discourse constraints, we derive a open-source dataset from the GUM corpus (Zeldes, 2017), which provides naturally occurring English texts annotated with expert RST trees. Original documents in GUM are often long (200–5,000 tokens) and complex (tree depths 7–24), making them unsuitable for direct generation-based evaluation (Figure 3). We therefore perform discourse-aware segmentation of GUM in a top-down manner guided by RST structures. This segmentation enables the Standard-Parseval score (Black et al., 1991) (full) of our RST parser (Maekawa et al., 2024) on the GUM dataset to jump from 55.2 to 67.0. This enables our parser to accurately analyze the RST Tree.

4.1.3 Automatic Evaluation Method

To account for generation stochasticity, we adopt a sampling-based evaluation strategy. For each model–instruction pair, we perform 10 independent generations. For **explicit constraints**, evaluation is performed via exact structural matching against gold discourse skeletons. Specifically, we use Qwen3-Max (Yang et al., 2025) *only as a structure parser* to convert generated texts into discourse representations, after which all matching decisions are made deterministically. For **implicit constraints**, exact matching is infeasible due to natural variation in structure realization. We therefore evaluate using RSTsim, a discourse-structure similarity metric described below. To validate the reliability of automatic evaluation, we manually inspected 100 randomly sampled cases and observed near-perfect (99/100) agreement with human judgments.

4.1.4 RSTsim: Automatic Discourse Structure Similarity Metric

Our tasks require adherence to discourse-level structure while allowing substantial freedom at the sentential semantic level. This renders n-gram metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) ineffective, and manual RST evaluation prohibitively expensive. We therefore propose **RSTsim**, an automatic metric for measuring discourse structure similarity. RSTsim is inspired by Iruskieta's RST tree comparison framework (Iruskieta et al., 2015) and RST-Tace (Wan et al., 2019), which shows that exact constituent correspondence is unnecessary as long as central subconstituents are preserved. The core idea of RSTsim is: **A discourse structure can be defined by the set of all its "root-to-leaf" paths.** Two structures are similar if and only if their path sets are similar in terms of **labels, order, and topology**.

RSTsim is a symmetric score that combines two unidirectional similarities:

$$\text{RSTsim}(T_{ref}, T_{cand}) = \frac{S(T_{ref} \to T_{cand}) + S(T_{cand} \to T_{ref})}{2} \quad (1)$$

where $S(T_{ref} \to T_{cand})$ measures structural recall and $S(T_{cand} \to T_{ref})$ measures structural precision. Each RST tree T is represented as an ordered set of root-to-leaf paths, where p_i denotes the i-th path, $|T|$ is the total number of paths, and d_i denotes the depth of p_i. The unidirectional similarity $S(T_A \to T_B)$ is defined as a weighted best-match score over all paths in T_A:

Unidirectional Similarity $S(T_A \to T_B)$: We define this as the weighted best-match score from T_A to T_B:

$$S(T_A \to T_B) = \frac{\sum_{i=1}^{|T_A|} w(d_i) \cdot \text{Score}(p_i, p_j^*)}{\sum_{i=1}^{|T_A|} w(d_i)} \quad (2)$$

where $w(d_i) = \log(d_i + 1)$ assigns higher weights to deeper paths, reflecting their greater importance in defining global discourse structure. For each path $p_i \in T_A$, its best-matching path $p_{j^*} \in T_B$ is selected by maximizing the path match score defined as:

$$\text{Score}(p_i, p_j^*) = \underbrace{\text{sim}_{\text{path}}(p_i, p_j^*)}_{\text{1. Path Similarity}} \times \underbrace{e^{-\beta\left|\frac{i}{|T_A|} - \frac{j^*}{|T_B|}\right|}}_{\text{2. Order Penalty}} \times \underbrace{\frac{1}{1 + k_{j^*} \cdot \gamma}}_{\text{3. Reuse Penalty}} \quad (3)$$

Here, T_A and T_B are represented as ordered sets of root-to-leaf paths, where p_i denotes the i-th path in T_A and $|T_A|$ is the total number of paths.

The depth d_i of p_i corresponds to its structural embedding level, and $w(d_i) = \log(d_i + 1)$ assigns higher importance to deeper paths. For each $p_i \in T_A$, we select its best-matching path $p_{j^*} \in T_B$ by maximizing $\text{Score}(p_i, p_j)$. As defined in Eq. (3), the path match score consists of three components. $\text{sim}_{\text{path}}(p_i, p_{j^*})$ measures semantic similarity between two paths via fuzzy matching over discourse relation labels. The order penalty compares the normalized positions $\frac{i}{|T_A|}$ and $\frac{j^*}{|T_B|}$, with β controlling sensitivity to relative order deviations. The reuse penalty depends on k_{j^*}, the number of times p_{j^*} has been previously matched, and γ controls its strength, discouraging many-to-one alignments.

A new metric must be validated against human judgment. We conducted a correlation study by sampling 100 (generated text, target structure) pairs from our test set. We asked 5 annotators trained in RST to rate the *structural adherence* of the generated text on a 1–5 Likert scale. The annotators exhibit strong internal consistency, with an inter-annotator agreement of $\text{ICC}(2, k) = 0.81$ (Shrout and Fleiss, 1979), indicating reliable human judgments. We then compare RSTsim scores with the averaged human ratings. The results show that RSTsim achieves a **high Spearman rank correlation ($\rho = 0.72$)** with human judgments (Spearman, 1961). Removing any component of RSTsim consistently reduces its Spearman correlation with human judgments, and all hyperparameters are chosen via validation to maximize agreement with annotators. These results demonstrate that RSTsim not only aligns well with human perception of discourse-level structural adherence, but can also reliably replace expensive manual evaluation in large-scale experiments.

4.2 Evaluation

Experimental Setup We evaluate a representative set of open- and closed-source models spanning different scales and architectures, including Qwen2.5 (7B, 14B, 32B, all -Instruct) (Yang et al., 2024), Qwen3 (80B-A3B, 235B-A22B) (Yang et al., 2025), Gemma-4 (26B-A4B, 31B) (Farabet et al., 2026), Claude-Sonnet-4-6 (Anthropic, 2026), Gemini-3.1-Pro-Preview (DeepMind, 2026), and GPT-5.4 (OpenAI, 2026). We use a State-of-the-Art (SOTA) automatic rst parser (Maekawa et al., 2024) to parse the texts into discourse trees.

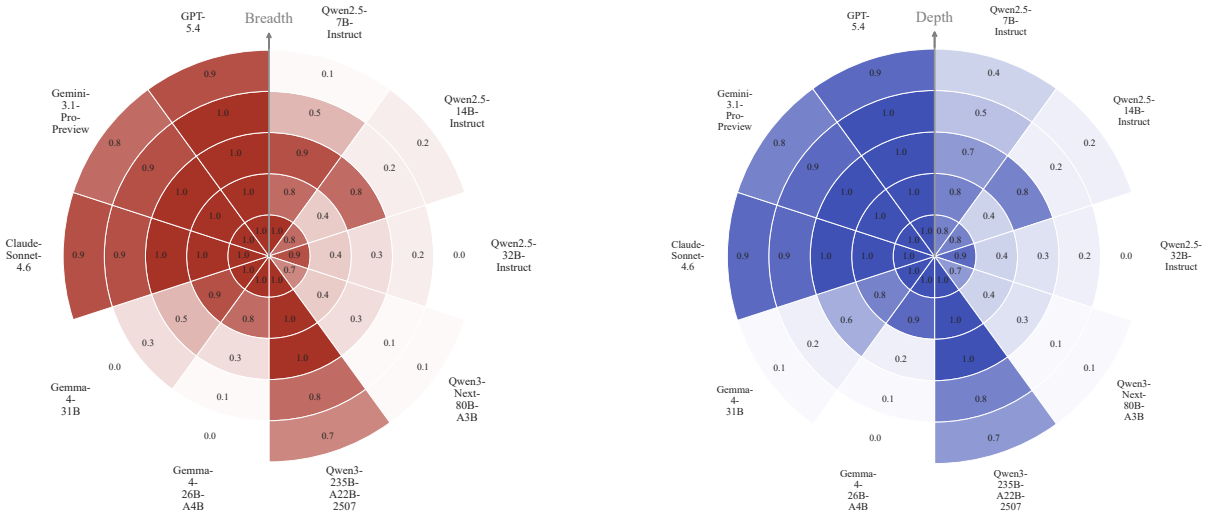

Figure 4: This presents the results of evaluating the discourse-level explicit constraint adherence capabilities of various models, with Breadth and Depth as the respective variables. The radial direction (outward from the center) corresponds to the variable increasing from 1 to 5, and **darker** colors indicate **higher scores**. It shows that as D and B of the constraint tree increase, adherence performance almost consistently degrades.

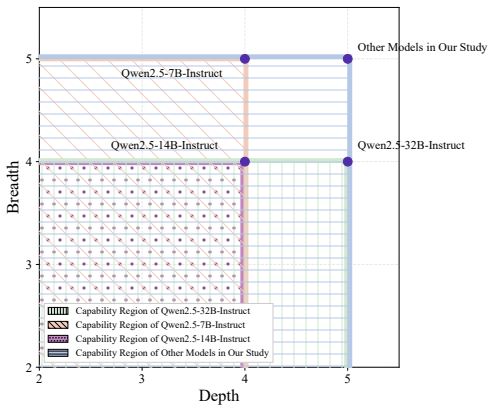

Figure 5: Capability coverage of different models under the explicit recognition task, showing the maximum (D, B) at which all 10 samples are correctly recognized. All the models in our study can recognize constraints perfectly in instructions of $D < 5$ and $B < 5$.

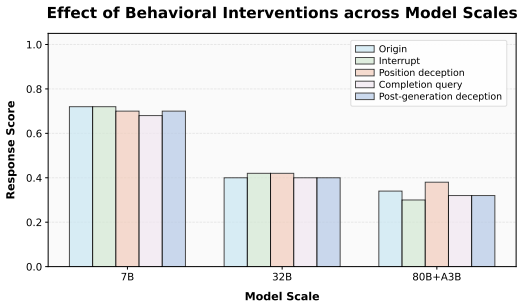

Figure 6: Model scores remain largely unchanged under different interventions, suggesting that generation decisions are insensitive to the true structural completion state.Within each sample, failure modes may occur multiple times and in multiple types; each occurrence of each type is counted as one instance in the statistics.

Discourse-level constraint adherence fails despite near-perfect recognition, revealing a systematic recognition–adherence gap. As Figure 5 and Table 1 show, all evaluated models, regardless of scale, exhibit strong capability in recognizing both explicit and implicit discourse-level constraints.Even the worst-performing model (14B) can consistently recognize constraints within parameters $D = B = 4$. However, as D and B of the constraint tree increase, adherence performance almost consistently degrades (Figure 4). Importantly, this degradation is independent of the absolute position of the deepest nesting and constraint-adherence performance exhibits clear non-monotonic behavior.We can see some mid-scale models perform the worst. Details are provided in Appendix C.5.

Generation decisions rely on surface linguistic signals rather than verified structural state. We designed a series of multi-turn behavioral interventions: (i) we interrupted generation at arbitrary positions and requested continuation, (ii) we deceived the model about its current position within the constraint structure, (iii) we queried whether specific constraints had been completed when they had not, (iv) we falsely informed the model after completion that certain constraints remained unfinished. As Figure 6 shows, these manipulations have

Model	Task (Implicit)		ρ(PPL, Score)		Prompt		
	Recognition	Adherence	Global	Local	Few-shot	CoT	Formatting
Qwen2.5-7B-Instruct	0.71	0.63	0.0137	-0.0053	0.64	0.62	0.63
Qwen2.5-14B-Instruct	0.68	0.60	0.0047	0.0026	0.61	0.62	0.59
Qwen2.5-32B-Instruct	0.69	0.63	0.0232	0.0208	0.64	0.62	0.66
Qwen3-Next-80B-A3B	0.83	0.79	0.0142	0.0317	0.80	0.78	0.80
Qwen3-235B-A22B-2507	0.85	0.81	-0.0089	0.0073	0.82	0.82	0.83
Gemma-4-26B-A4B	0.71	0.65	0.0220	0.0108	0.63	0.61	0.64
Gemma-4-31B	0.69	0.62	0.0294	-0.0028	0.60	0.62	0.63
Claude-Sonnet-4-6	0.87	0.80	0.0119	-0.0094	0.81	0.80	0.81
Gemini-3.1-Pro-Preview	0.86	0.82	0.0015	-0.0125	0.79	0.81	0.80
GPT-5.4	0.86	0.81	-0.0096	0.0144	0.80	0.81	0.82
Human	0.92	0.91	0.0152	0.0199	0.91	0.90	0.88

Table 1: Performance across implicit discourse tasks. ρ(PPL, Score) denotes Spearman correlation between global/local perplexity and discourse adherence scores (near-zero values indicate weak association). Prompt columns report results under different prompting strategies. We prompt humans in the same way as we do for LLMs.

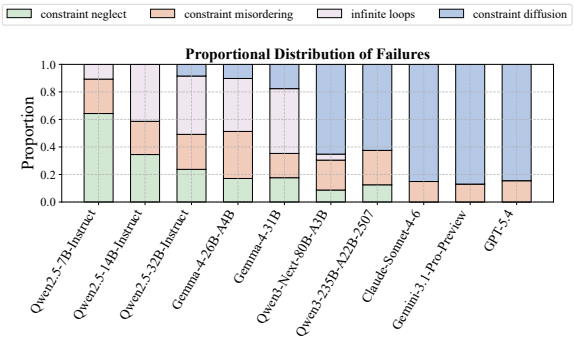

Figure 7: Frequencies of qualitative failure modes across models on the discourse-level adherence task. The distribution of failure modes is significantly associated with model size.

little effect on fluency or continuation quality: models continue seamlessly after interruptions, accept false positional cues, confidently report completion of unfinished constraints, and restart generation from arbitrarily specified locations when misled. For the 7B, 32B, and 80B+A3B models, the standard deviations are $\sigma \approx 0.014$, 0.011, and 0.027 with fluctuation amplitudes of 2%, 2.7%, and 8.1% respectively, which are all very low. None of these results can explicitly reveal the impact of these operations. These collectively demonstrate that models do not verify the actual structural completion state during generation, but instead rely on surface-level linguistic cues such as conversational confirmation or discourse markers.

Structural failures manifest as distinct, scale-dependent error modes rather than uniform degradation. Structural failures do not appear as a uniform decline in generation quality, but instead manifest as distinct and systematic error modes. Analysis reveals that discourse-level constraint violations can be categorized into four scale-related failure modes: constraint neglect, constraint misordering, infinite loops, and constraint diffusion. Increasing model size reshapes the distribution of failure modes without eliminating them. This suggests that increased capacity amplifies reliance on surface-level statistical patterns, rather than resolving the underlying cause of discourse-level structural failure (Figure 7). Furthermore, We conduct behavioral validation experiments in which models are required to explicitly state which constraint they are about to handle before continuing generation. The error patterns observed in this pre-generated plan closely mirror (93% consistency rate) those in the subsequent generation. This alignment indicates that failures occur even when the relevant constraint has already been recognized, thereby ruling out explanations based on missing understanding or insufficient planning capacity.

Failures are independent of fluency, prompting strategies, and standard supervised fine-tuning. First, controlling for output length yields nearly identical failure patterns across models. Second, few-shot prompting, chain-of-thought prompting, and explicit formatting of constraints lead to negligible improvements. Third, supervised fine-tuning improves constraint adherence only within the training distribution. When evaluated on novel constraint relations or compositions, fine-tuned models almost fail to generalize (Figure 8). Fourth, per-

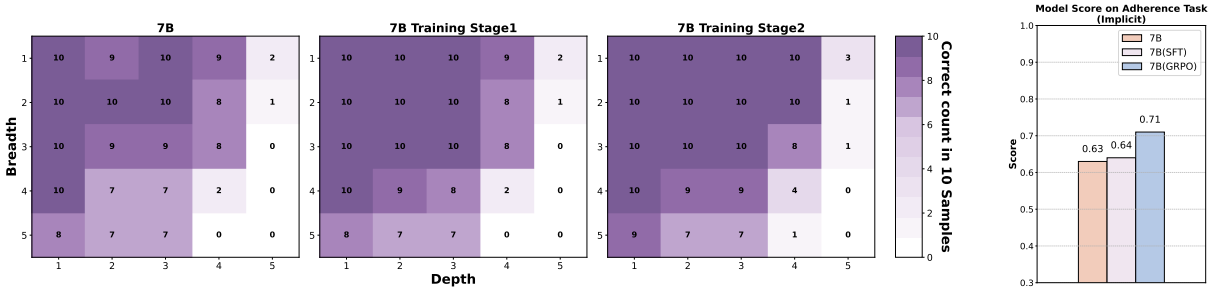

Figure 8: The left shows the model's performance on the explicit dataset during curriculum learning. The right shows the transfer performance of the trained model on the implicit task.Training is conducted on the explicit dataset with curriculum learning, using RSTsim as the reward signal. We still using correct count in 10 samples to evaluate. while the scores on explicit tasks are improved, the model's capabilities on implicit tasks are also enhanced.

plexity (Jelinek et al., 1977) measured both globally and at local violation spans shows near-zero correlation with constraint-adherence scores (Table 1). Finally, discourse-level planning exhibits no clear mitigating effect on these failures. Together, these results rule out explanations based on the generation difficulty of the text itself, fluency trade-offs or prompting deficiencies.

4.3 Training with Discourse-Level Rewards

Objective Misalignment at the Discourse Level Taken together, the preceding evidence indicates that the recognition–adherence gap does not stem from insufficient structural representation. Instead, it arises from a fundamental misalignment between the autoregressive training objective and discourse-level structural integrity. The standard objective optimizes next-token likelihood conditioned on surface context. It does not provide any learning signal for verifying whether a global discourse structure has been correctly completed. Under this objective, discourse-level constraints are inevitably reduced to soft guidance rather than enforceable conditions. Then, increasing model scale or refining prompts cannot fundamentally resolve the problem. What is needed is to provide the model with discourse-level structural coherence signals.

Discourse-Level Optimization with GRPO and RSTsim Therefore, to address the gap caused by objective misalignment, we optimize discourse-level constraint adherence via GRPO (Shao et al., 2024) with RSTsim as the reward signal. We train Qwen2.5-7B-Instruct exclusively on explicit constraints. At each step, the model generates multiple candidates whose structures are compared to gold skeletons to compute scalar rewards (RSTsim score) for GRPO. Implicit constraints are not used during training. Our training setup and capability-aware curriculum design are provided in Appendix C.6

Implications for the Recognition–Adherence Gap After only two curriculum-based GRPO iterations, the model shows substantial improvements on explicit constraints and notable transfer to implicit tasks (Figure 8). Specifically, this finding reveals two core implications: (i) GRPO with RSTsim rewards fosters general discourse planning capabilities, and the recognition-adherence gap can be alleviated by leveraging discourse-level reward signals, rather than being an issue of representational insufficiency; (ii) the discourse-level explicit and implicit constraints we designed share distinct commonalities in their requirements for model capabilities, all pointing to the model's ability to follow discourse-level instructions, which fully validates the rationality of our design.

5 Conclusion

In this paper, we study discourse-level instruction following in LLMs for the first time to the best of our knowledge. We construct explicit and implicit constraints benchmarks and introduce RSTsim, an automatic metric for structural adherence. We reveal a systematic recognition–adherence gap: models can reliably recognize discourse constraints but often fail to adhere to them during generation. This gap arises because standard autoregressive training maximizes next-token likelihood based on local context, without providing any signal for global structural correctness. Finally, we show that reinforcement learning with discourse-level rewards can effectively improve adherence, establishing a unified framework for diagnosing, evaluating, and optimizing discourse-aware generation.

Limitations

Dependence on RST Formalization. Our framework adopts Rhetorical Structure Theory (RST) as an interpretable and well-established formalization of discourse organization. While this formulation effectively captures a wide range of hierarchical discourse phenomena, natural discourse may also involve additional relational patterns beyond tree-structured representations. Extending discourse-level instruction following to richer structural paradigms remains a promising direction for future work.

Reliance on Automatic Discourse Parsing. Our evaluation and optimization pipeline relies on automatic discourse parsers to obtain structural representations. Although existing parsers already demonstrate sufficiently strong performance for our setting, further advances in discourse parsing may enable more accurate and fine-grained discourse-level supervision in the future.

Boundary Between Explicit and Implicit Constraints. This work separates discourse-level constraints into explicit and implicit settings to enable controlled analysis of different instruction-following behaviors. However, in real-world interactions, explicit and implicit structural requirements are often intertwined rather than clearly separated. Future work may further investigate mixed or gradually shifting forms of discourse-level constraints in more realistic interactive scenarios.

Cross-lingual Generalization. Our current experiments focus on English discourse structures derived from English RST resources. Since discourse organization may vary across languages and writing traditions, extending discourse-level instruction following to multilingual settings represents an important and valuable future research direction.

Structural Adherence and Other Generation Qualities. This work primarily focuses on discourse-level structural adherence. Other important dimensions of generation quality, such as creativity, persuasiveness, and long-form communicative effectiveness, are not explicitly optimized in our framework. Exploring the interaction between discourse-level control and broader generation objectives remains an interesting direction for future study.

Ethics Statement

We hereby acknowledge that all authors of this work are aware of the ACL Code of Ethics and adhere to the associated code of conduct.

This work involves limited human participation solely for evaluation purposes, including inter-annotator agreement analysis and human-based benchmarking. Human annotators are only asked to assess model outputs and reference discourse structures, and they do not interact with or influence model behavior.

Specifically, five undergraduate students were recruited as annotators. They received basic training on Rhetorical Structure Theory (RST) and the evaluation protocol before participating in the experiments. Their participation was conducted as part of a course-related academic activity, and they received course credit for their involvement. No monetary compensation was provided.

All evaluation materials are derived from the datasets used in this study and do not contain personal, sensitive, or private information. Annotator participation is voluntary, and the evaluation process poses minimal risk.

We view this human evaluation as a necessary component for validating discourse-level evaluation metrics and benchmarking model performance. We do not anticipate that the inclusion of human evaluation introduces ethical risks beyond those commonly associated with large language model research.

AI tools were used only for editorial assistance, such as improving readability and correcting grammatical errors. No content generation affecting the scientific contribution of this work was delegated to AI systems.

The GUM Corpus used in this study is a multi-source corpus whose underlying texts are aggregated from diverse public repositories, each governed by its own license terms (including CC-BY 4.0, CC-BY-SA 3.0, CC-BY-NC-SA 4.0, etc.; see dataset documentation for source-specific details). The linguistic annotations provided with the corpus are uniformly licensed under CC-BY 4.0. In accordance with the most restrictive terms among the source licenses (notably the Non-Commercial clauses), all uses in this work are strictly for non-commercial research purposes.

References

Rilwan Adewoyin, Ritabrata Dutta, and Yulan He. 2022. RSTGen: Imbuing fine-grained interpretable control into long-FormText generators. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1822–1835, Seattle, United States. Association for Computational Linguistics.

Anthropic. 2026. Introducing claude sonnet 4.6. Anthropic Research Blog.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024a. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7421–7454. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3119–3137. Association for Computational Linguistics.

E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Google DeepMind. 2026. Gemini 3.1 pro model card. Google DeepMind.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *CoRR*, abs/2404.04475.

Clement Farabet, Olivier Lacombe, and Gemma Team. 2026. Gemma 4: Byte for byte, the most capable open models. Google Blog.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Keno Harada, Yudai Yamazaki, Masachika Taniguchi, Edison Marrese-Taylor, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2025. When instructions multiply: Measuring and estimating LLM capabilities of multiple instructions following. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16506–16526, Suzhou, China. Association for Computational Linguistics.

Mikel Iruskieta, Iria Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Lang. Resour. Eval.*, 49(2):263–309.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.

Zae Myung Kim, Anand Ramachandran, Farideh Tavazoee, Joo-Kyung Kim, Oleg Rokhlenko, and Dongyeop Kang. 2026. Align to structure: Aligning large language models with structural information. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 37519–37528. AAAI Press.

Jinnan Li, Jinzhe Li, Yue Wang, Yi Chang, and Yuan Wu. 2025. Structflowbench: A structured flow benchmark for multi-turn instruction following. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 9322–9341. Association for Computational Linguistics.

Shiyang Li, Jun Yan, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, and Hongxia Jin. 2024. Instruction-following evaluation through verbalizer manipulation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3678–3692, Mexico City, Mexico. Association for Computational Linguistics.

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

Dongqi Liu and Vera Demberg. 2024. Rst-lora: A discourse-aware low-rank adaptation for long document abstractive summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2200–2220. Association for Computational Linguistics.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2803–2815, St. Julian's, Malta. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281.

OpenAI. 2026. Introducing gpt-5.4. OpenAI Blog.

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

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86 2:420–8.

Charles Spearman. 1961. The proof and measurement of association between two things.

Shujun Wan, Tino Kutschbach, Anke Lüdeling, and Manfred Stede. 2019. RST-tace a tool for automatic comparison and evaluation of RST trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 88–96, Minneapolis, MN. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxing Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. 2024. Benchmarking complex instruction-following with multiple constraints composition. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Wenpeng Yin, Qinyuan Ye, Pengfei Liu, Xiang Ren, and Hinrich Schütze. 2023. LLM-driven instruction following: Progresses and concerns. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 19–25, Singapore. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Relation	Type	Ordered
Contrast	Binary	No
Analogy	Binary	Yes
Enumeration	N-ary	No
Cause	Binary	Yes
Condition	Binary	Yes
Sequence	N-ary	Yes
Elaboration	Binary	Yes

Table 2: Discourse relation inventory $\mathcal{R}$ with type specifications, ordering constraints, and functional explanations. **Ordered** column indicates whether argument sequence is semantically significant: *Yes* means directionality matters (asymmetric), *No* means arguments are interchangeable (symmetric). Binary relations connect exactly two discourse units; N-ary relations span $\geq$2 units.

A Rhetorical Structure Theory

Rhetorical Structure Theory (RST) posits that a text is a coherent whole organized through semantic relations among multiple discourse units. These relations are characterized by two key features: (1) nuclearity, which distinguishes between a Nucleus—the essential unit conveying the primary information—and a Satellite, which serves a supportive or supplementary role; and (2) rhetorical relations (e.g., causality, contrast, elaboration), which specify the semantic function linking the units. RST typically represents textual structure as a binary tree, where leaf nodes correspond to Elementary Discourse Units (EDUs), and internal nodes encode the rhetorical relations and nuclearity configurations, thereby revealing the text's underlying coherence. The properties of some relations are summarized in Table 2.

B Data Examples

For both explicit and implicit constraints data, we both designed five types of prompts to mitigate bias stemming from suboptimal prompt quality. Experiments have shown that differences in prompts do not have an explicit impact on the performance of the LLMs in our study, so we only need to use one of them.

B.1 Explicit

We use parameters (D, B, $P(i)$) and topics to automatically construct explicit constraints data. As Figure 9 shows, Meta-constraints are combined into a constraint tree in the manner we designed, thereby generating instruction data, and meanwhile,

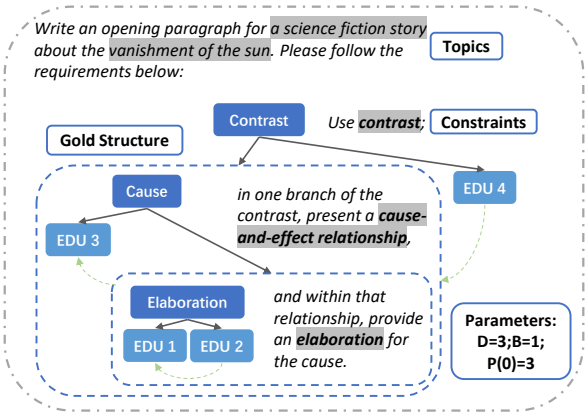

Figure 9: An illustration of the close relationship between the text and the tree in our explicit constraint data.

the discourse tree structure is obtained. Here is one example from our explicit data:

```
{
  "prompt": "Write an opening
    paragraph for a science fiction
    story about the vanishment of the
    sun. Please follow the
    requirements below:Use contrast;
    in one branch of the contrast,
    present a cause-and-effect
    relationship, and within that
    relationship, provide an
    elaboration for the cause." ,
  "constraints":"Use contrast; in one
    branch of the contrast, present a
    cause-and-effect relationship, and
     within that relationship, provide
     an elaboration for the cause." ,
  "rst_tree": "{
    ( Root
    ( Nucleus Contrast
    ( Nucleus (EDU) )
    ( Satellite Cause
      ( Nucleus (EDU) )
      ( Satellite Elaboration
        ( EDU )))))
  }" ,
  "Depth": "3" ,
  "Breadth": "1" ,
  "position": "[3,0,0,0,0]"
}
```

B.2 Implicit

```
{
  "prompt": "Write a new poetry on the
     vanishment of the sun, following
    the same overall organization and
    progression of ideas as the
    reference text. Do NOT reuse
    specific content, examples, or
    expressions from the reference
    text. Focus on maintaining a
    similar high-level structure, such
     as how ideas are introduced,
    developed, contrasted, or
```

```
    concluded.
  Reference text: I HIDE myself within
     my flower, That wearing on your
    breast, You, unsuspecting, wear me
     too - And angels know the rest. I
     hide myself within my flower,
    That, fading from your vase, You,
    unsuspecting, feel for me Almost a
     loneliness." ,
  "text": "I HIDE myself within my
    flower, That wearing on your
    breast, You, unsuspecting, wear me
     too - And angels know the rest. I
     hide myself within my flower,
    That, fading from your vase, You,
    unsuspecting, feel for me Almost a
     loneliness." ,
  "rst_tree": "{
    ( Root
    ( Nucleus (joint)
    ( Nucleus (span)
    ( Nucleus (span) )
    ( Satellite (elaboration)
    ( Nucleus (same-unit) )
    ( Nucleus (same-unit)
    ( Satellite (context) )
    ( Nucleus (span)
    ( Nucleus (same-unit) )
    ( Nucleus (same-unit)
    ( Satellite (mode-manner) )
    ( Nucleus (span) ))))))
    ( Satellite (adversative) ))
    ( Nucleus (joint-list)
    ( Nucleus (span) )
    ( Satellite (elaboration)
    ( Nucleus (same-unit) )
    ( Nucleus (same-unit)
    ( Satellite (context) )
    ( Nucleus (span)
    ( Nucleus (same-unit) )
    ( Nucleus (same-unit)
    ( Satellite (mode-manner) )
    ( Nucleus (span) ))))))))
  }",
  "category": "poetry"
}
```


Algorithm 1 RSTsim Core Algorithm

Input: Reference tree T_{ref}, Candidate tree T_{cand}
Output: Symmetric discourse similarity score
$S_1 \leftarrow \text{UnidirectionalSim}(T_{\text{ref}}, T_{\text{cand}})$
$S_2 \leftarrow \text{UnidirectionalSim}(T_{\text{cand}}, T_{\text{ref}})$
return $(S_1 + S_2)/2$

Procedure UNIDIRECTIONALSIM(T_A, T_B)
Extract ordered root-to-leaf paths $\{p_i\}_{i=1}^{|T_A|}$ from T_A
Extract ordered root-to-leaf paths $\{q_j\}_{j=1}^{|T_B|}$ from T_B
Initialize $C_j \leftarrow 0$ for all j {reuse counter}
$S \leftarrow 0, W \leftarrow 0$
for $i = 1$ to $|T_A|$ **do**
  $d_i \leftarrow \text{Depth}(p_i)$
  $w_i \leftarrow w(d_i)$
  $s_{\max} \leftarrow 0, j^* \leftarrow -1$
  for $j = 1$ to $|T_B|$ **do**
    $s \leftarrow \text{Score}(p_i, q_j, i, j, C_j)$ {Eq. (3)}
    if $s > s_{\max}$ **then**
      $s_{\max} \leftarrow s$
      $j^* \leftarrow j$
    end if
  end for
  $S \leftarrow S + w_i \cdot s_{\max}$
  $W \leftarrow W + w_i$
  if $j^* \neq -1$ **then**
    $C_{j^*} \leftarrow C_{j^*} + 1$
  end if
end for
return S/W

C Evaluation Details

C.1 RSTsim Pseudocode

The pseudocode of RSTsim is shown in Algorithm 1.

C.2 Automatic RST Parser Capability

The automatic RST parser we used demonstrates sufficient capability for our task. For explicit constraint data, to satisfy discourse-level constraints, the generated text is often structurally coherent with prominently featured discourse development-related vocabulary. We erase the spans of parser results and the spans of the gold tree, regarding them as fully compliant; the Standard-Parseval score (full) reaches a sufficiently high value of 93.1, with the only limitation being ambiguity in some

vague relationships. For explicit constraint data, as we mentioned before, the Standard-Parseva score (full) reaches 67.0, showing sufficient capability. After segmenting the GUM corpus, the parser's capability is by no means a limiting factor for our evaluation.

C.3 Failure Mode Case Study

Using the same $D = 4$ and $B = 5$ data, we can easily elicit the four failure modes of the models.

"prompt": "Compare the differences between the thinking processes of large language models and humans. List three differences. For each difference, explain the reason that leads to it. Within each reason, provide a concrete example to illustrate the explanation. After listing the differences, enumerate three limitations that large language models face in current applications. After listing the limitations, separately explain the cause of the first limitation. After explaining the cause, describe under what conditions it might be improved. After describing the conditions, propose a research direction for large language models using an analogy. "

Constraint Neglect Constraint neglect refers to the complete omission of one or more required discourse units or relations specified by the target discourse-level constraints. Here is an example showing that the model neglected "Reason", corresponding to "Cause" relation constraint.The locations where errors occurred have been bolded.

"response": "Human knowledge updating relies on individual learning and the accumulation of personal experience, whereas the knowledge updating of large language models depends on updates to their training data. For example, humans can immediately learn about the latest scientific discoveries, while large language models may need to wait for the release of new training datasets to acquire such information. In communication, humans are able to deeply understand others' emotions and communicate effectively through non-verbal cues such as tone and facial expressions. In contrast, large language models process emotions primarily through textual data and lack the ability to perceive non-verbal information. For instance, when a person expresses sadness in a conversation, a large language model may only infer the emotion by analyzing keywords in the text, rather than truly understanding the specific context and depth of the sadness. **Humans possess a high level of creative thinking, enabling them to generate innovative ideas and solve problems creatively. By contrast, the creative capabilities of large language models are constrained by their training data and algorithmic design.** Three major limitations currently faced by large language models include: Data bias: The training data of large language models may contain inherent biases. Limitations in logical reasoning: Large language models still exhibit constraints when performing complex logical reasoning. Privacy and security concerns: When handling sensitive information, large language models may face risks related to privacy leakage and data security. Among these three limitations, data bias is primarily caused by imbalanced training data and insufficient representativeness. To mitigate data bias, it is necessary to increase the diversity and representativeness of datasets, ensuring that different groups and contexts are adequately covered. By analogy, the reinforcement learning process of large language models can be compared to a "student" who continuously learns and grows."

Constraint Misordering Constraint misordering occurs when all required discourse components are generated, but their relative order violates the specified discourse structure. The bolded part has reversed the nested order of Elaboration and Cause.

"response": "Large language models rely on statistical associations rather than causal reasoning. **For example, when answering "roads are slippery on rainy days," they merely reproduce common co-occurring patterns without understanding changes in friction, whereas humans build causal chains through experience, such as forming the physical understanding that "water reduces friction" after stepping on a wet surface. This is because their training is based on pattern matching over massive amounts of textual data.**Large language models lack persistent autonomous goals, as their operation is driven by user inputs. For instance, when asked "how to lose weight," they list various methods but do not actively track outcomes or adjust recommendations. Humans, by contrast, possess intrinsic motivation and long-term planning abilities, such as proactively recording dietary intake and modifying plans for the sake of health. The knowledge of large language models is static and lagging, since their training data are fixed up to a certain cutoff point. For example, they are unable to know policy changes introduced after 2023. Humans, however, can perceive environmental changes in real time and update their cognition accordingly; for instance, upon personally encountering newly enacted regulations, they immediately adjust their behavior. The three limitations encountered in current applications are: inability to guarantee factual accuracy, difficulty in handling complex multi-step reasoning, and lack of embodied real-world experience. The cause of the first limitation is that model generation is based on probabilistic prediction rather than fact-checking mechanisms. For example, a model may describe fictional events as if they were historical facts because it has never been exposed to signals that verify truth. This limitation may be mitigated under conditions where external knowledge bases with real-time retrieval and fact-consistency verification modules are introduced. This is analogous to equipping a blind painter with tactile maps and a proofreading assistant, enabling the painter to draw not solely from fragmented memories but by consulting references while drawing. From this perspective, a research direction of "retrieval-augmented generative language models under truthfulness constraints" can be proposed."

Infinite Loops Infinite loops describe a pathological generation behavior in which the model repeatedly produces the same or highly similar discourse units, often attempting to satisfy a constraint but failing to progress to subsequent ones. We have bolded the infinite loop section.

"response": "There are the following three differences between large language models and human thinking:Knowledge update speed: Human knowledge updating depends on individual learning and the accumulation of experience, whereas the knowledge updating of large language models depends on updates to training data. For

example, humans can instantly learn about the latest scientific discoveries, while large language models need to wait for the release of new training datasets in order to obtain this information.Emotional understanding and expression: Humans are able to deeply understand and express complex emotions, whereas large language models mainly rely on predefined emotional vocabulary and patterns for emotional analysis and expression. For example, humans can convey complex emotions through tone, facial expressions, and body language, while large language models can only perform simple emotion recognition through emotional words in text. Creativity and intuition: Humans possess a high degree of creativity and intuition, enabling innovative thinking and problem solving. In contrast, the creativity of large language models is constrained by their training data and mainly relies on pattern matching to generate responses. For example, humans can propose innovative solutions based on intuition, whereas large language models can only provide answers based on existing data. The three limitations that large language models encounter in current applications include: Data bias: Large language models may inherit biases from their training data, leading to unfair outputs. For example, if the proportion of female engineers in the training data is low, large language models may exhibit gender bias when answering questions about female engineers. Lack of contextual understanding: When processing long texts or complex dialogues, large language models may have difficulty understanding context, resulting in inaccurate answers. For example, in a long conversation, large language models may fail to accurately understand the conversational background, causing the response to deviate from the topic. Originality of generated content: The content generated by large language models may lack originality, because their responses are mainly based on existing data. For example, poems generated by large language models may lack unique creativity, as the generation process relies on existing poetic patterns. Cause of the first limitation: Data bias mainly stems from the imbalance of training data. For example, if the proportion of female engineers in the training data is low, large language models may exhibit gender bias when answering questions about female engineers. Conditions for improvement: Improving data bias requires collecting more diverse training data to ensure that all groups in the dataset have sufficient representation. **For example, increasing data samples of female engineers to reduce gender bias and ensure that all groups in the dataset have sufficient representation. For example, increasing data samples of female engineers to reduce gender bias and ensure that all groups in the dataset have sufficient representation. For example, increasing data samples of female engineers to reduce gender bias and ensure that all groups in the dataset have sufficient representation. For example, increasing data samples of female engineers to reduce gender bias**"

Constraint Diffusion Constraint diffusion refers to the phenomenon where discourse constraints are weakly or partially reflected across multiple segments instead of being realized as distinct, well-bounded discourse units. As a result, the intended discourse relations become blurred. The bolded part demonstrates that the later-positioned "Analogy" constraint has been obviously generalized to more positions and stylized.

"response": "First, large language models rely on statistical patterns rather than causal reasoning, because their training is based on co-occurrence relationships in massive text corpora. For example, when encountering “rain,” they often associate it with “bringing an umbrella,” without understanding that rain causes people to get wet and therefore requires an umbrella-**much like knowing that soldiers need to carry bulletproof vests**. Humans, by contrast, infer behavior by understanding causal mechanisms.Second, large language models lack persistent self-awareness and intentionality, as their architecture does not include an intrinsic goal system. For example, when responding to “I want to drink coffee,” they merely imitate linguistic conventions, whereas human expressions arise from genuine physiological needs-**much like a parrot mimicking human speech**. Third, the “knowledge” of large language models is static and lacks experiential grounding. Because their training data cannot be updated after a cutoff point, they know nothing about events occurring after, for example, 2024. Humans, however, can continuously revise their cognition through real-time perception, **like a vehicle that keeps moving forward rather than remaining stationary**. Current applications of large language models are limited in three aspects: inability to reliably handle facts that require real-time verification, difficulty in performing long-term coherent reasoning and planning, and lack of embodied understanding of the physical world. The cause of the first limitation lies in the static nature of training data and the lack of dynamic connections to external information sources. For example, a model cannot know whether today is a holiday unless such information is contained in its data or accessed through plugins, **much like a person who never goes online**. If a model can securely and efficiently connect to verifiable real-time databases and dynamically cite sources during generation, this limitation can be mitigated. This is analogous to equipping a scholar who only reads history books with a telephone that can instantly call authoritative institutions. Research on large language models can thus be likened to constructing a “verifiable generation” architecture, in which outputs are proactively checked against trusted external sources, just as scientists cite experimental data before publishing conclusions."

C.4 PPL Calculation

We calculated Global PPL and Local PPL. For global PPL, we compute it directly; for local PPL, we first locate the non-compliant segments (if full compliance is achieved, we instead compute the global PPL). We then calculate the local PPL for these segments and perform a correlation analysis between the obtained PPL scores and the compliance scores of the corresponding text. The conclusion drawn is that LLMs' poor adherence to discourse-level constraints is not intended to pursue fluency, and the goals of these two are irrelevant.

C.5 Model Score under Different P(i) Settings

Figure 10 shows the irrelevance between the relative position of the deepest node in the constraint tree and the model's performance in adhering to constraints.

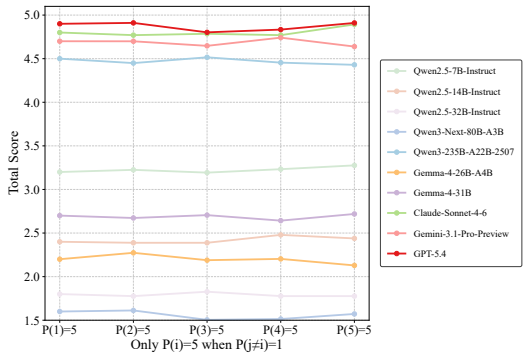

Figure 10: Model Total Scores with Different P(i) Settings. We set only one i each time such that $P(i) = 5$, while all other $P(j)$ are set to 1.

C.6 Training

Training Setup To ensure parameter efficiency and training stability, we adopt LoRA-based fine-tuning with rank 16 and scaling factor 8. The actor is optimized with a learning rate of 1×10^{-6}, and KL regularization is applied during training with a coefficient of 0.001 to prevent policy drift. At each GRPO step, the model generates multiple candidate responses per prompt, with a maximum rollout batch size of 128 sequences. For each prompt, we sample $n = 4$ candidate responses to compute discourse-level rewards. Both prompts and responses are truncated to a maximum length of 500 tokens, and overlong prompts are filtered to avoid incomplete discourse structures.Training is performed for three epochs. Rollouts are executed using vLLM. We split the explicit constraint dataset into training, validation, and test sets in an 8:1:1 ratio for each stage of currilulum learning.

Capability-Aware Curriculum Design The successful application of curriculum learning is attributed to the fact that we have obtained the LLM's compliance capability boundaries on explicit datasets through previous evaluation experiments. For instructions within the boundaries (correct samples 9/10 - 10/10) , near the boundaries (correct samples 6/10 - 8/10) , and beyond the boundaries (correct samples 0/10 - 5/10) , we synthesize them at a $1 : 3 : 1$ ratio, while ensuring: (1) For instructions near the capability boundaries, each round of GRPO rollout usually contains distinct positive and negative samples to maximize the effectiveness of GRPO; (2) Instructions within and beyond the boundaries enable the model to attempt to further expand its capabilities without compromising its basic capabilities. After one round of curriculum learning, we re-evaluate the capability boundaries and conduct another round. Significant improvements are observed after two rounds.