Inverse Design for Text Generation with Accurate and Complex Causal Graph

Anonymous Author(s)
Affiliation
Address
email

Abstract

The development and evaluation of causal discovery methods requires large quantities of data with causal structure annotations. However, such real-world data with annotations is insufficient. Therefore, text generation with causal structure annotations serves as a critical foundational task for advancing causal discovery research. Nevertheless, existing data generation methods cannot ensure both causal structure accuracy and complexity. To address this, we apply inverse design from scientific computing to Chain-of-Thought (CoT) and propose a method named iTAG. Our method is capable of generating large quantities of text with accurate and complex causal graphs. Empirical evaluation demonstrates the substitutability of iTAG-generated data for real-world data through two experiments. First, annotation accuracy evaluation shows remarkable causal graph annotation accuracy across complexities (F1>96%, SHD<1, SID<0.5). Second, substitutability analysis reveals strong statistical correlation between generated and real-world text across various metrics computed on existing causal discovery algorithms (Pearson=0.96, Spearman=0.94, $R^2 = 0.93$).

1 Introduction

Causal discovery researches rely extensively on generated data for testing algorithm performance due to four critical limitations of existing text datasets with causal structure annotations: (1) extremely high manual annotation costs that impede dataset expansion; (2) difficulty for human to accurately identify complex causal relationships [38, 12], which consequently leads to; (3) insufficient data volume for robust model training and evaluation; and (4) overly simplistic causal structures that inadequately represent real-world complexity [16, 29]. Hence it is a fundamental task to generate data with accurate and complex causal structure annotations.

Causal discovery in text also necessitates the generation of data with complex causal structure annotation and high annotation accuracy. This is a persistent technical challenge that recent Large Language Models (LLMs) have attempted to address, yet they face difficulties in ensuring both annotation accuracy and causal structures complexities [12, 21]. Early generative approaches predefine causal graphs with specified effect sizes and employ parameterized generation methods (bag-of-words, LDA, and GPT-2) [37]. While these methods ensured causal structure annotation accuracy through controlled vocabulary and sentence structure, they produced oversimplified causal structures in text without flexible controlled details. In recent researches, despite modern LLMs' applications and attemptance in textual causal inference [7, 19], multiple studies have conclusively demonstrated their limitations in causal reasoning ability [6, 35, 25]. While modern LLMs can flexibly control complexity, this fundamental constraint prevents them from ensuring the annotation accuracy in text generation.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

To address the challenge of ensuring both annotation accuracy and causal complexity, we propose **iTAG**: **i**nverse design for **T**ext gener**A**tion with causal **G**raph. Unlike existing methods that directly convert causal graphs into corresponding text, iTAG innovatively applies inverse design from scientific computing to CoT prompting. Through a three-phase workflow, iTAG first controls the complexity of the generated causal graph by controlling variable quantity. Subsequently, in the latter two phases, it employs a reverse-design CoT approach to transform the causal graph into real-world concepts and textual representations, respectively. In our empirical evaluation, we evaluate the accuracy of iTAG-generated data across different complexities and its substitutability for real-world data. In summary, the main contribution of this paper includes:

- We propose iTAG, a novel methodology that applies inverse design from scientific computing to LLMs by CoT. iTAG is capable of generating large quantities of text with accurate and complex causal graph.
- Our first experiment in Section 4.1 evaluates the annotation accuracy of iTAG generated text across complexities. Results achieves remarkable causal graph annotation accuracy by manual expert verification (F1>96%, SHD<1, SID<0.5).
- Our second experiment in Section 4.2 further measures metrics of state-of-the-art (SOTA) causal discovery methods on generated and real-world data. The results exhibit strong statistical correlation, achieving Pearson=0.96, Spearman=0.94, and R^2=0.93, thereby validating our generated data as a viable substitute for real-world data.

The subsequent sections of this paper respectively recap the related work, introduce the proposed method, present the experimental results and analysis, and finally conclude the paper.

2 Related work

2.1 Text generation with causal graph

Text generation with causal graph is a task that transforms causal graphs into coherent natural language text while preserving all causal relationships. Formally, given a causal graph $G = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ represents concept nodes and $E \subseteq V \times V$ represents directed causal relationships such that $(v_i, v_j) \in E$ indicates v_i causally influences v_j, the objective is to generate text T comprising a set of sentences $S = \{s_1, s_2, ..., s_m\}$ that linguistically encode all relationships in E without introducing spurious connections not present in G. The task involves both generating the text T and establishing an annotation function A that maps T to a reconstructed causal graph $G' = A(T)$, where ideally G' is isomorphic to G.

Current research confronts the dual challenges of ensuring annotation accuracy and causal structure complexity [28]. A category of approaches predominantly focuses on causal structure complexity manipulation without guaranteeing the accuracy of the causal structure annotations. While these generative methods effectively control the causal complexity through predefined instruction templates and relationship definitions, they inherently compromise annotation accuracy by relying on LLMs' imperfect causal understanding capabilities. Therefore, they struggle to discern genuine causal relationships from mere correlational patterns or linguistic associations [22, 6]. Another line of work employs parameterized generation methods to ensure annotation accuracy while constraining the flexibility and complexity of causal frameworks. These approaches utilize structured parameterization schemes and predefined mappings from causal graphs to text representations. By employing rigorous mathematical formulations or template-based mechanisms, they achieve high fidelity in capturing explicit causal relationships. However, they inherently limit causal framework complexity. These methods restrict interventions to discrete specifications rather than enabling real-valued causal effects. Furthermore, they transform complex graph structures into linear narrative forms, which cannot fully capture intricate causal interrelationships [37, 28].

2.2 Inverse design and CoT

To our best knowledge, iTAG is the first method that applies inverse design from the domain of scientific computing into LLMs by CoT. The introduction of these two technologies is as follows:
Inverse design reverses traditional engineering approaches by enabling breakthrough applications in scientific computing domains such as fluid dynamics and aerodynamics [3, 26]. It parameterizes design spaces and optimizes performance objectives by iteratively simulating candidates and updating

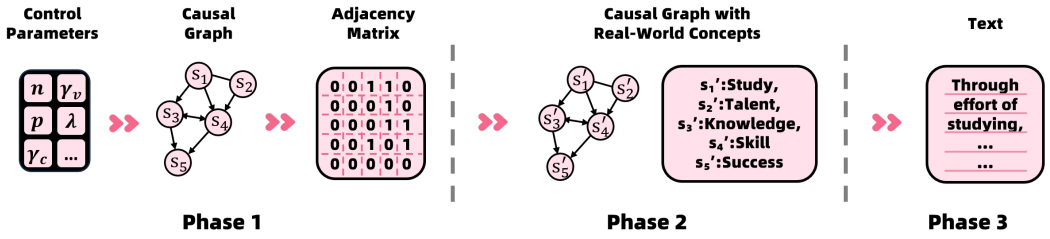

Figure 1: An example of the three-phase workflow of iTAG: INverse design for Variable-controlled tExt generation with counterfactual Reasoning Transformation.

parameters to minimize target-output gaps [27, 17]. Conventional methods employ specialized solvers with limited transferability and high computational costs [9, 20]. Whereas modern inverse design approaches enable efficient gradient-based optimization through differentiable surrogate models [31, 32]. **CoT** prompting is a prompt engineering technique that enables LLMs to perform complex reasoning by explicitly generating intermediate steps before reaching a final answer [36, 34]. This method is primarily designed for tasks requiring multi-step reasoning, including arithmetic and commonsense reasoning, where standard prompting often falls short [18, 40]. Current researches demonstrate that CoT prompting achieves substantial performance improvements, with accuracy gains across various reasoning benchmarks, particularly excelling in mathematical or reasoning problem-solving tasks [39, 41].

Existing methods that directly convert causal graphs into text overly rely on LLMs' intrinsic reasoning capabilities, while well-designed CoT can significantly enhance reasoning abilities [36]. Therefore, iTAG leverages inverse design principles to construct CoT, guiding LLMs to iteratively refine the final text by targeting the causal structure of causal graphs with varying complexity in a parameterized causal graph design space. This inverse design approach ensures causal structure accuracy without relying on the LLMs' inherent reasoning capabilities. Simultaneously, it enables flexible control over the target causal structure, thereby addressing the dual challenges presented in Section 2.1.

3 Method

In this section, we introduce iTAG and its components. We first outline the three-phase workflow of iTAG (Section 3.1), then detail its' phases in Sections 3.2, 3.3, and 3.4, respectively.

3.1 Overview of the three-phase workflow of iTAG

iTAG generates text with causal graph through a three-phase pipeline as shown in Figure 1. In **phase 1**, **control parameters** such as node count (n), expected graph density (p), colliders ratio (γ_v), mediator chains count (λ), and confounders ratio (γ_c) are transformed into a structured **causal graph** (nodes s_1 through s_5 in the example) and subsequently converted into an **adjacency matrix**. This matrix precisely encodes all causal relationships, with entries of 1 indicating direct causal influences (such as $s_1 \to s_3$ in the example) while 0s represent the absence of such relationships. In **phase 2**, abstract variables in the causal graph undergo substitution with **real-world concepts** (s_1':"Study", s_2':"Talent", s_3':"Knowledge", s_4':"Skill", s_5':"Success") while maintaining strict adherence to the causal structure defined in the adjacency matrix. In **phase 3**, these real-world concepts and their causal structure are transformed into coherent natural language **text** that implicitly embeds the defined causal relationships, such as the text generated in the example:

> Through effort of **studying**, individuals acquire **knowledge** while developing **skills** enhanced by their natural **talents**. **Knowledge** and **skills** reciprocally enhance one another, and those who simultaneously possess **knowledge** and refined **skills** typically achieve **success**.

The complete process of each phase is illustrated in Figure 2 and detailed as follows.

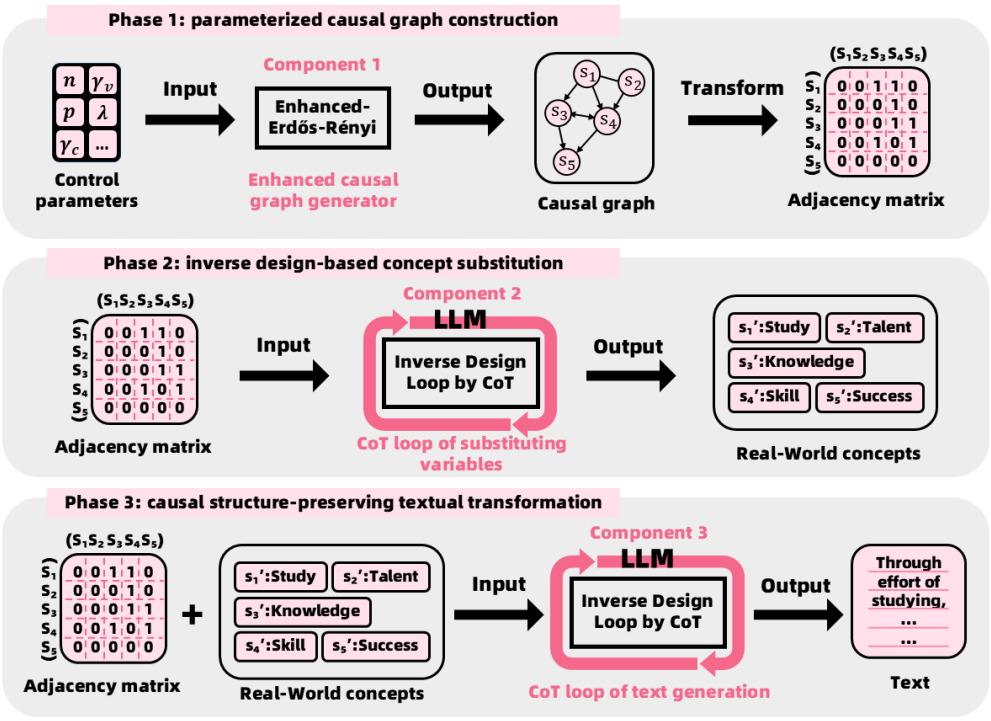

Figure 2: Detailed three phases of iTAG. Rectangle with rounded corners are different forms of data. Rectangle with square corners are components for the implemention of different phases.

3.2 Phase 1: parameterized causal graph construction

Phase 1 transforms control parameters into structured causal graphs and adjacency matrices. The **input** parameters include node count (n), expected graph density (p), maximum parents ($max\_parents$), maximum children ($max\_children$), confounders ratio (γ_c), colliders ratio (γ_v), and mediator chains count (λ), providing precise control over the structural complexity; the **output** is a directed acyclic graph (DAG) with its **transformed** corresponding adjacency matrix representation, where matrix elements $a_{ij} = 1$ indicate a direct causal relationship from node i to node j, while $a_{ij} = 0$ indicates the absence of a direct causal relationship between these nodes.

To execute this transformation, **Component 1** implements an enhanced Erdős-Rényi causal graph generator to construct DAGs [10, 11]. The implementation first calculates the expected number of edges ($expected\_edges = p \times \frac{n(n-1)}{2}$) and corresponding edge probability; then initializes an empty directed graph and adds edges using the enhanced Erdős-Rényi approach while enforcing structural constraints on $max\_parents$ and $max\_children$; subsequently adds specialized causal structures, specifically incorporating $\gamma_c \times n$ confounders (common causes of multiple variables), $\gamma_v \times n$ colliders (variables influenced by multiple independent causes), and λ mediator chains (sequences forming indirect causal pathways), through dedicated subroutines.

3.3 Phase 2: inverse design-based concept substitution

Phase 2 transforms abstract causal graph nodes into real-world concepts while preserving the causal structure defined by the adjacency matrix. The **input** is the adjacency matrix from Phase 1 and the **output** consists of domain-specific concepts assigned to the node that maintain all causal relationships indicated by 1s in the matrix while ensuring no spurious relationships are introduced where 0s appear.

Inverse design is to obtain **structures that exhibit desired performance targets** by using **optimization algorithms** to **iteratively** search through possible configurations and automatically generate and optimize structures through **forward analysis** and **backward analysis** [4]. **CoT** is a prompting

Algorithm 1 Inverse Design-Based Concept Substitution

Require: Adjacency Matrix A
Ensure: Real-World Concept Set C

$relationships \leftarrow AnalyzeCausalStructure(A)$ ▷ *Extract all 1s and 0s relationships*
$concepts \leftarrow InitialConceptAssignment()$ ▷ *Initial domain-specific assignment*
repeat
  $validation\_results \leftarrow CounterfactualVerification(concepts, relationships)$
  $fallacies \leftarrow FallacyAnalysis(validation\_results)$
  if $fallacies \neq \emptyset$ **then**
    $concepts \leftarrow RefineConceptAssignment(concepts, fallacies)$
until $fallacies = \emptyset$
return $concepts$

technique that guides LLMs to solve problems by explicitly articulating intermediate reasoning steps, similar to how humans "think step by step" to reach a conclusion [36].

Component 2 implements variable substitution through an inverse design methodology that employs a specialized prompt template shown in Appendix A that guides LLMs through a structured CoT reasoning loop as outlined in Algorithm 1. The $AnalyzeCausalStructure$ function identifies all existing (value 1) and non-existing (value 0) connections in the adjacency matrix, establishing the causal **structures that exhibit desired performance targets**; $InitialConceptAssignment$ assigns domain-specific concepts to abstract nodes while conforming to the predetermined causal structure; $CounterfactualVerification$ **forward analyzes** each proposed relationship against the target structure by implementing Pearl's Level 3 causal inference, examining whether effect B would still occur in the same manner if cause A had not occurred, with all other conditions held constant; and $FallacyAnalysis$ **backward analyzes** reasoning errors, triggering iterative refinement through **optimization algorithms** $RefineConceptAssignment$ that **iteratively** minimizes the gap between the concept assignments and the target causal structure. Therefore, this **CoT** implements the complete **inverse design** approach through iterative reasoning loops.

3.4 Phase 3: causal structure-preserving textual transformation

Phase 3 generates text through an inverse design methodology that transforms causal graphs with real-world concepts into natural language text while preserving the causal structure. The **input** consists of both the adjacency matrix from Phase 1 and the real-world concepts from Phase 2; the **output** is coherent natural language text that implicitly embeds all causal relationships defined in the adjacency matrix without introducing spurious relationships.

Component 3 implements text generation through an identical inverse design CoT loop used in Component 2, with a critical modification in step 2: replacing variable substitution with writing initiation to generate text that implicitly embeds established causal relationships. Simultaneously, concept control as detailed in Appendix A ensures that no irrelevant concepts or spurious relationships are introduced into the text, ultimately producing text with corresponding causal graph.

4 Empirical evaluation

4.1 Evaluating annotation accuracy across complexities

In this section, we first evaluate the annotation accuracy of text generated by SOTA method Davinci and our method iTAG across causal structure complexities (variable quantity 3-10) in Section 4.1.2 [28], then further explored the capability of iTAG to generate large quantities of data in Section 4.1.3. We compare with one baseline in Section 4.1.2 because the only other existing generation work's predefined components cannot meet our experiment's multi-theme generation requirements [37]. It is noteworthy that the range of variable quantity derives from two considerations: (1) It comprehensively represents realistic causal structure scenarios since current text involving human decision-making typically contain fewer than 10 variables. (2) Manual annotation costs increase dramatically with the number of variables for large-scale, multi-person sample validation. To ensure data diversity, we selected three text themes where AI participates in human decision-making (business, medi-

cal, and legal), with equal distribution in the generation. Implementation tools are provided in https://placeholder.com.

4.1.1 Experimental setup

Ground truth construction: Ground truth is established via a panel of 11 human annotation experts. Each annotator evaluated 1,000 text data for each of baseline Davinci and our method iTAG, where text data refers to generated text with causal graph as described in Section 2.1. The ground truth is determined through majority voting across annotators' assessments. We posit that the voting results constitute the accurate causal graph for the corresponding text, and that LLMs defined the true scope of variables in text encompassed within the causal graph. As verified through manual validation, there is no extraneous concepts beyond the causal graph in the text.

Evaluation metrics: Causal graph annotation accuracy metrics are F1, SHD and SID: Precision ($P = \frac{TP}{TP+FP}$), Recall ($R = \frac{TP}{TP+FN}$), and F1-score ($F1 = \frac{2PR}{P+R}$) assess edge-wise accuracy with higher values indicating better performance (↑), where TP, FP, and FN represent correctly identified, falsely identified, and missed causal edges, respectively. For structural comparison, we employ Structural Hamming Distance (SHD = $\sum_{i,j} I(G_{ij} \neq \hat{G}_{ij})$), which counts edge modifications needed to transform the predicted graph $\hat{G}$ into the ground truth G with lower values indicating better performance (↓), where $I(\cdot)$ equals 1 when the condition is true and 0 otherwise, and Structural Intervention Distance (SID = $\sum_{i \neq j} I(Pa_{\hat{G}}^{do(i)}(j) \neq Pa_{G}^{do(i)}(j))$), which measures causal inference accuracy by counting node pairs with different post-intervention parent sets with lower values indicating better performance (↓), where $\text{Pa}_{G}^{do(i)}(j)$ denotes the parent set of node j after intervening on node i in graph G.

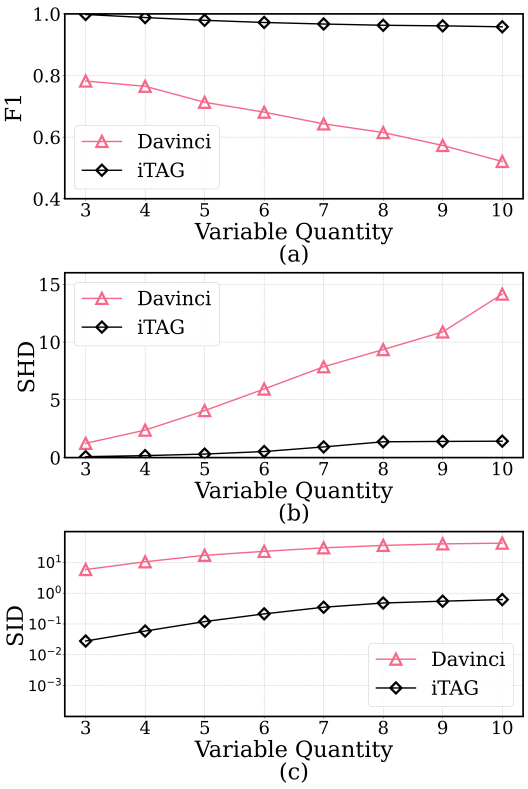

Figure 3: Causal graph annotation accuracy metrics F1, SHD, and SID on text generated by Davinci and iTAG across variable quantities.

4.1.2 Annotation accuracy study under varying variable quantities

Figure 3 presents comparative results for Davinci and iTAG across varying variable quantities. The x-axis represents an increasing number of variables, corresponding to progressively more complex causal structures. As causal complexity increases, the baseline Davinci model struggles to maintain performance, exhibiting deterioration across all evaluation metrics. In contrast, our proposed iTAG method consistently maintains near-perfect performance with minimal degradation, substantially outperforming Davinci. This is because iTAG's inverse design methodology ensures iteratively reasoning through CoT processes until most fallacies are resolved even facing complex causal structures. These results demonstrate that iTAG can effectively generate text with causal graph while simultaneously ensuring both complexity and accuracy. Notably, both iTAG and Davinci are built upon the same underlying LLM (GPT-4o), indicating that iTAG's superior performance is not merely attributable to the inherent reasoning capabilities of the base model.

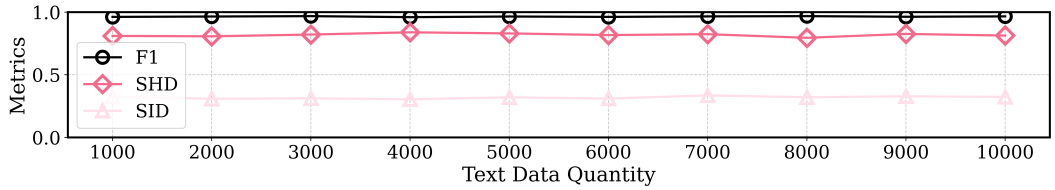

Figure 4: iTAG's annotation accuracy across text data quantities.

4.1.3 Annotation accuracy study under varying text data quantities

We investigated potential changes in iTAG's annotation accuracy when generating large quantities of text data, as shown in Figure 4. It is noteworthy that due to prohibitively high manual annotation costs, our metrics were computed using a randomly sampled 10% subset of the textual data with each text data quantity. Under the condition of equal proportions among variables in the generated data, the metrics remain nearly constant as the data quantity increases, achieving a high average F1 of **0.966** and low SHD and SID values of **0.813** and **0.323**, respectively. This demonstrates that iTAG's generated text quality remains stable even when generating large quantities of data.

4.2 Investigating the substitutability of generated data for real-world data

This experiment tests SOTA text causal discovery methods, including non-LLM algorithms (CLEANN, SA, PA, CGEN) and LLMs (Claude-3-7, Claude-3-5, GPT-4o, GPT-4o-mini) [30, 33, 23], on both iTAG-generated data and real-world data, with both datasets containing 1000 samples equally distributed across themes and variable quantities in Section 4.2.2. We then further explored the metrics stability on large quantities of generated data in Section 4.2.3, and analyze metrics' statistical correlation between generated and real-world data in Section 4.2.4. The LLM prompt template are detailed in Appendix A. Non-LLM methods are conducted in a controlled environment using an NVIDIA RTX 3090 GPU and Intel Xeon Platinum 8362 CPU.

4.2.1 Experimental setup

Ground truth construction: To evaluate the gap between text generated by iTAG and real-world data, we calculated every metric across variable quantities using both generated data as ground truth and real-world text data with manually constructed ground truth (maintaining the **same** construction methodology as in Section 4.1.1). For real-world text data across different domains, we selected datasets with identifiable causal structures with a range of simple to complex from medical, business, and legal fields: MIMIC IV ver.2.2 NOTE, FinCausal 2025, and JUSTICE [15, 24, 2].

Evaluation metrics: Metrics of causal discovery accuracy (F1, SHD, SID) maintain the same. Statistical correlation metrics are r, ρ and R^2: Pearson correlation coefficient ($r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$), measuring linear relationship strength, with higher values indicating better correlation (↑); Spearman's rank correlation ($\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$), where d_i is the difference between ranks of corresponding values, assessing monotonic relationships, with higher values indicating better correlation (↑); and coefficient of determination ($R^2 = 1 - \frac{\sum_i (y_i-\hat{y}_i)^2}{\sum_i (y_i-\bar{y})^2}$), representing variance explained proportion, with higher values indicating better fit (↑), where x_i and y_i represent performance metrics on generated and real-world data.

4.2.2 Causal discovery accuracy study under varying variable quantities

Figure 5 demonstrates the results computed on generated and real-world data using methods across variable quantities. We observe consistent and high convergence across all methods and metrics on both data. This consistency likely stems from iTAG's ability to simultaneously ensure causal structural complexity and annotation accuracy, providing evidence for the feasibility of using generated data as a viable substitute for real-world data in testing causal discovery algorithms. Furthermore, results across variable quantities reveal that the accuracy of existing methods decreases rapidly as the number

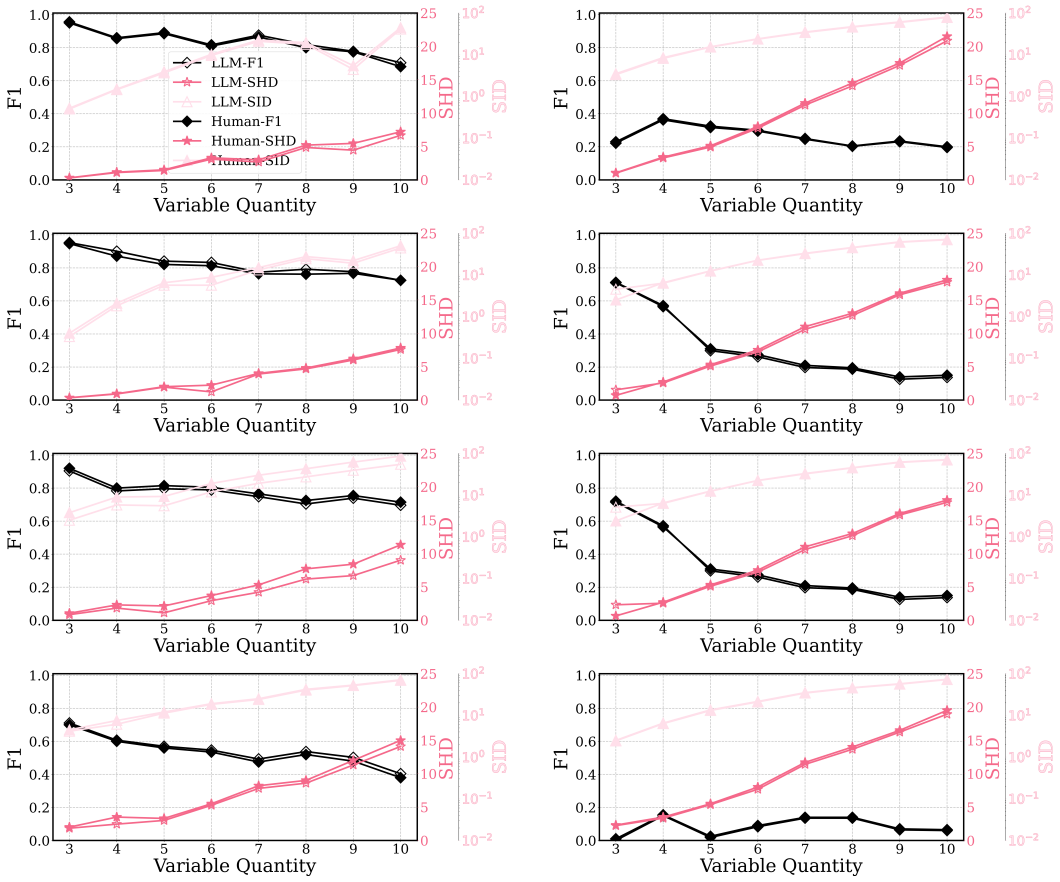

Figure 5: Causal graph annotation accuracy metrics F1, SHD, and SID calculated using ground truth provided by LLM and human of different causal discovery methods across variable quantities. The first column displays data for methods: Claude-3-7, Claude-3-5, GPT-4o, and GPT-4o-mini, respectively. The second column displays data for methods: CLEANN, PA, SA, and CGEN, respectively.

of variables increases, indicating a critical research direction for future studies in textual causal discovery to address the challenges of modeling complex causal structures.

4.2.3 Causal discovery accuracy study under varying text data quantities

We investigate the stability of metrics results with larger quantities, from 1000 to 10000, in potential practical applications, as shown in Figure 6. Since this does not involve manual annotation, we evaluate the complete text data quantities without sampling. Under the condition of equal proportions among variables in the generated data, different methods do not show changes across all metrics as quantity increases. This demonstrates that iTAG-generated data maintains high quality and stability in case of large-scale testing of causal discovery algorithms.

4.2.4 Statistical correlation study

The analysis provided in Table 1 is the statistical correlation analysis of metrics derived from generated and real-world data. This represents the most critical aspect of the experiment, quantitatively evaluating the substitutability of generated data compared for real-world data for causal discovery algorithm assessment: (1) The mean Pearson correlation coefficient of 0.962 indicates an extremely strong linear relationship between metrics derived from generated and real-world data. (2) The mean Spearman correlation coefficient of 0.927 demonstrates highly consistent ranking order of models' performance across both datasets. This is particularly significant for model selection decisions, as it indicates that models performing optimally on generated data are likely to remain optimal choices

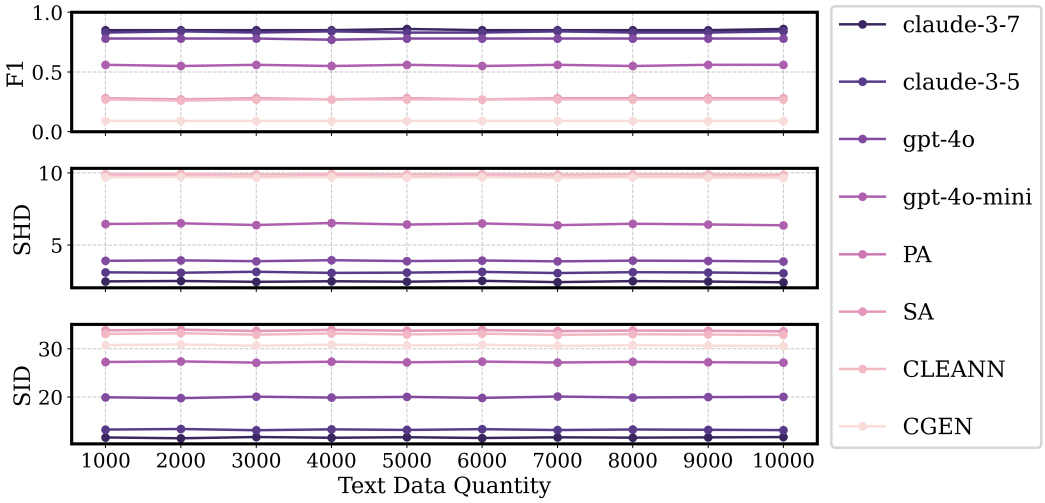

Figure 6: Causal discovery accuracy for methods across text data quantities.

Table 1: Statistical correlation analysis of metrics derived from generated and real-world data.

Pearson Corr.			**Spearman Corr.**		**Linear Regr.**	
Metric	Correlation	p-value	Metric	Correlation	Metric	R-squared
F1	0.960	0.0005	F1	0.970	F1	0.921
SHD	0.988	<0.0001	SHD	0.922	SHD	0.976
SID	0.938	0.0017	SID	0.922	SID	0.880
Average	**0.962**	/	Average	**0.938**	Average	**0.926**

in real-world applications. (3) The mean R^2 value of 0.926 approaching 1 demonstrates the high predictive capability of generated data for real-world performance.

These statistical analysis results collectively demonstrate that despite real-world data contains greater noise and natural variation, there exists an exceptionally strong statistical correlation between model evaluation results on generated data by iTAG and real-world performance, validating the substitutability of generated data for real-world data on causal discovery algorithm assessment.

5 Conclusion

We presented a method for batch text generation with complex accurate causal graphs. This contributes to filling the gap of the lack of text data with causal structure annotations, establishing foundational work for future causal discovery in textual context, which may lead to: (1) substantial reduction data and experimental costs for addressing research questions, and (2) extension of research inquiries into more complex and diverse scenarios.

Although iTAG currently demonstrates ideal performance, several fundamental **limitations** constrain its current applicability: First, in this groundbreaking generative work that potentially extends data generation to future applications, iTAG focuses on DAG as causal structure representations rather than complete structural causal models (SCMs). This is because encoding intricate functional relationships and effect magnitudes within natural language presents significant challenges that exceed the representational capabilities of existing LLMs and even humans. Second, based on our investigation of real-world data in Section 4.2.1, the graph density, another secondary parameter that determines causal structure complexity, ranges from 0.2 to 0.3 in real-world texts. We therefore adopt this range as our experimental configuration for data generation. However, future work ought to explore the potential for more flexible control over graph density.

References

[1] Abouei, A.M., Mokhtarian, E., Kiyavash, N., & Grossglauser, M. (2024). Causal Effect Identification in a Sub-Population with Latent Variables. *arXiv preprint arXiv:2405.14547*.

[2] Alali, M., Syed, S., Alsayed, M., Patel, S., & Bodala, H. (2021). Justice: A benchmark dataset for supreme court's judgment prediction. *arXiv preprint arXiv:2112.03414*.

[3] Allen, K., Lopez-Guevara, T., Stachenfeld, K.L., Sanchez Gonzalez, A., Battaglia, P., Hamrick, J.B., & Pfaff, T. (2022). Inverse design for fluid-structure interactions using graph network simulators. *Advances in Neural Information Processing Systems*, 35, 13759–13774.

[4] Bendsøe, M.P., & Kikuchi, N. (1988). Generating optimal topologies in structural design using a homogenization method. *Computer Methods in Applied Mechanics and Engineering*, 71, 197–224.

[5] Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2020). Neural ordinary differential equations. *SIAM Journal on Mathematics of Data Science*, 2(3), 628–666.

[6] Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., Liu, T., & Han, B. (2024). Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37, 96640–96670.

[7] Dhawan, N., Cotta, L., Ullrich, K., Krishnan, R.G., & Maddison, C.J. (2024). End-to-end causal effect estimation from unstructured natural language data. *arXiv preprint arXiv:2407.07018*.

[8] Dörfler, J., van der Zander, B., Bläser, M., & Liskiewicz, M. (2024). On the Complexity of Identification in Linear Structural Causal Models. *arXiv preprint arXiv:2407.12528*.

[9] Du, X., He, P., & Martins, J.R.R.A. (2021). Rapid airfoil design optimization via neural networks-based parameterization and surrogate modeling. *Aerospace Science and Technology*, 113, 106701.

[10] Erdős, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6, 290–297.

[11] Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publicationes Mathematicae, Institute of Mathematics, Hungarian Academy of Sciences*, 5, 17–61.

[12] Feder, A., Keith, K.A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M.E., & others (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10, 1138–1158.

[13] Gao, N., Xie, H., Wang, K., Jiang, S., & Guo, X. (2021). Deep learning for designing electric machines: A comprehensive review. *IEEE Transactions on Magnetics*, 57(9), 1–10.

[14] Huang, D., Allen, K. L., Oppe, E. E., Graham, R. L., & DeMello, A. J. (2021). Machine learning for microfluidics: automating and enhancing the design process. *Lab on a Chip*, 21(22), 4253–4273.

[15] Johnson, A., Pollard, T., Horng, S., Celi, L.A., & Mark, R. (2023). MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet.

[16] Khetan, V., Rizvi, M.I.H., Huber, J., Bartusiak, P., Sacaleanu, B., & Fano, A. (2021). MIMICause: Representation and automatic extraction of causal relation types from clinical notes. *arXiv preprint arXiv:2110.07090*.

[17] Kim, B., Azevedo, V.C., Thuerey, N., Kim, T., Gross, M., & Solenthaler, B. (2019). Deep fluids: A generative network for parameterized fluid simulations. *Computer Graphics Forum*, 38(2), 59–70.

[18] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.

[19] Le, H.D., Xia, X., & Chen, Z. (2024). Multi-agent causal discovery using large language models. *arXiv preprint arXiv:2407.15073*.

[20] Li, Y., Wu, J., Tedrake, R., Tenenbaum, J.B., & Torralba, A. (2018). Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*.

[21] Liang, X., Wang, H., Wang, Y., Song, S., Yang, J., Niu, S., Hu, J., Liu, D., Yao, S., Xiong, F., & others (2024). Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

[22] Liu, C., Chen, Y., Liu, T., Gong, M., Cheng, J., Han, B., & Zhang, K. (2024). Discovery of the Hidden World with Large Language Models. In *Advances in Neural Information Processing Systems*, 37, pages 102307–102365.

[23] Maisonnave, M., Delbianco, F., Tohme, F., Milios, E., & Maguitman, A.G. (2022). Causal graph extraction from news: a comparative study of time-series causality learning techniques. *PeerJ Computer Science*, 8, e1066.

[24] Moreno-Sandoval, A., Porta, J., Carbajo-Coronado, B., Torterolo, Y., & Samy, D. (2025). The Financial Document Causality Detection Shared Task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221.

[25] Nie, A., Zhang, Y., Amdekar, A.S., Piech, C., Hashimoto, T.B., & Gerstenberg, T. (2023). Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36, 78360–78393.

[26] Peurifoy, J., Shen, Y., Jing, L., Yang, Y., Cano-Renteria, F., DeLacy, B.G., Joannopoulos, J.D., Tegmark, M., & Soljačić, M. (2018). Nanophotonic particle simulation and inverse design using artificial neural networks. *Science Advances*, 4(6), eaar4206.

[27] Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., & Battaglia, P. (2020). Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*.

[28] Phatak, A., Mago, V.K., Agrawal, A., Inbasekaran, A., & Giabbanelli, P.J. (2024). Narrating causal graphs with large language models. *arXiv preprint arXiv:2403.07118*.

[29] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A.K., Webber, B.L., & others (2008). The penn discourse TreeBank 2.0. In *LREC*.

[30] Rohekar, R.Y., Gurwicz, Y., & Nisimov, S. (2023). Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36, 31450–31465.

[31] Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., & Battaglia, P. (2020). Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pages 8459–8468. PMLR.

[32] Thuerey, N., Weißenow, K., Prantl, L., & Hu, X. (2020). Deep learning methods for Reynolds-averaged Navier–Stokes simulations of airfoil flows. *AIAA Journal*, 58(1), 25–36.

[33] Tran, K.H., Ghazimatin, A., & Saha Roy, R. (2021). Counterfactual explanations for neural recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1627–1631.

[34] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171*.

[35] Wang, Z. (2024). CausalBench: A Comprehensive Benchmark for Evaluating Causal Reasoning Capabilities of Large Language Models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151.

[36] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., & others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.

[37] Wood-Doughty, Z., Shpitser, I., & Dredze, M. (2021). Generating synthetic text data to evaluate causal inference methods. *arXiv preprint arXiv:2102.05638*.

[38] Yao, B., Jindal, I., Popa, L., Katsis, Y., Ghosh, S., He, L., Lu, Y., Srivastava, S., Li, Y., Hendler, J., & others (2023). Beyond labels: Empowering human annotators with natural language explanations through a novel active-learning architecture. *arXiv preprint arXiv:2305.12710*.

[39] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems*, 36.

[40] Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). Automatic Chain of Thought Prompting in Large Language Models. *arXiv preprint arXiv:2210.03493*.

[41] Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., & others. (2023). Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *International Conference on Learning Representations*.

A Prompt template

Component 2 of iTAG

Adjacency Matrix:
[Matrix]

==========

Task: Please assign concepts from a meaningful real-world [domain/series of events] to the [N] nodes in the causal DAG represented by this adjacency matrix, while fully conforming to the causal relationships between nodes.

Requirements: In your thinking, please use the following separators to assist your reasoning, and only output the final result when you are satisfied with it:
—-Let me first analyze carefully—-
(First list all relationships between nodes represented by 1s in the matrix and all non-existent relationships represented by 0s in the matrix)
—-First attempt—-
(Then write out the concepts corresponding to the nodes)
—-Check for errors—-
(Please use the complete paradigm "'First, imagine that in the real world, [variable A] occurs (or takes some value) and [variable B] subsequently occurs (or takes some value). If [variable A] had not occurred (or had taken a different value), would [variable B] still occur in the same way (or maintain the same value) under the same background conditions? If in the counterfactual scenario where [variable A] did not occur, [variable B] significantly changes (either does not occur at all, or occurs in a substantially different way, time, intensity, or characteristics), and this change is systematic rather than accidental, while all other potential background conditions and common causes that might affect [variable B] remain constant, then we can reasonably infer a causal relationship between [variable A] and [variable B], meaning [variable A] is a cause of [variable B]. Conversely, if in the counterfactual scenario, even when [variable A] does not occur, [variable B] still occurs in essentially the same way, or changes in [variable B] can be fully explained by changes in other variables, and this situation stably repeats across various background conditions, this indicates there is no direct or substantial causal relationship between [variable A] and [variable B], and the observed correlation between them may be coincidental, a spurious association due to common causes, or an indirect effect mediated through other variables rather than a true causal connection."' to check whether the concepts conform to ALL relationships marked by 1s and do NOT conform to ALL relationships marked by 0s. If causal relationships are unreasonable, consider the reasons for errors and avoid them in the next attempt)
Begin second analysis
—-Second attempt—-
...

Your answer should be in JSON format:

```
{
  "Existing causal relationships (values of 1 in the matrix)": [
    "Node 0 → Node 1",
    ...
  ],
  "Non-existing causal relationships (values of 0 in the matrix)": [
    "Node 0 → Node 1",
    ...
  ],
  "Real concepts assigned to variables": [
    "Node 0: ___",
    ...
  ],
  "Relationship verification": {
    "Existing causal relationships": [
```

```
      "___ (natural language description conforming to the reasoning
      paradigm)",
      ...
    ],
    "Non-existing causal relationships": [
      "___ (natural language description conforming to the reasoning
      paradigm)",
      ...
    ]
  }
}
```

Component 3 of iTAG

Concepts:
[Concepts]

Adjacency matrix between concepts:
[Adjacency Matrix]

==========

Task: Please express all concepts clearly in a paragraph of natural language (implicitly conveying relationships between concepts rather than explicitly stating them), without introducing any additional concepts.

Requirements: In your thinking, please use the following separators to assist your reasoning, and only output the final result when you are satisfied with it:
—-Let me first analyze carefully—-
(Which relationships between concepts should be indirectly described and which should not appear)
—-First attempt—-
(Try writing your paragraph)
—-Check for implicitness—-
(Even though all concepts appear in this paragraph, the causal relationships between them are not clearly stated, ensuring readers must make their own judgments)
—-Check for errors—-
(Check if the description avoids expressing relationships that don't exist, i.e., 0s in the matrix. If the description is not rigorous or not implicit, consider the reasons and begin a second analysis)
Begin second analysis
—-Second attempt—-
...

Your answer should be in JSON format:

```
{
  "Natural language description": "..."
}
```

LLM causal discovery prompt

Text:
[Text]

Important concepts appearing in the text:
[Important concepts]

==========

Task: For the text and the important concepts appearing in it, please infer the **direct causal relationships** between each concept based on the text and common sense reasoning (causal relationships are not the same as correlations. For example, high temperature has causal relationships with both the number of drownings and ice cream sales, but the number of drownings and ice cream

sales only have correlation without direct causal relationship).

Requirements: The format for annotating causal relationships for each text should be:
0101 (means that the first concept has direct causal relationships with the second and fourth concepts, and the first concept is the cause of the second and fourth concepts) 0010 (means that the second concept has a direct causal relationship with the third concept, and the second concept is the cause of the third concept) 0000 (means that the third concept is not the cause of any other concept) 0100 (similarly, ...)

Your response must be in JSON format containing the following:

```
{
  "adjacency matrix": [
    [0,1,0,...],
    [0,0,1,...],
    ...
  ]
}
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly delineate the scope of the paper's work on "text generation with causal graph" and the fundamental impact of substantial high-quality generated data on causal discovery in text.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have elaborated on the limitations in Section 5 regarding the relatively limited scope of our experiments in relation to the broader contributions to the field.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper attempts to propose simple and effective heuristic methods, however, it struggles to mathematically model the key principles related to natural language processing.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all specific prompt templates in Appendix A, comprehensive code for data generation and processing in the Supplementary Material, and will make the code publicly available through the link in Section 4 in case of publication.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will make the code link public in Section 4 after the double-blind review phase.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our methodology does not involve fine-tuning but rather directly utilizes specially designed prompts to call APIs. The implementation of baselines involving training and testing details strictly follows the guidance provided in the relevant papers for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: To control and present the reliability and variability of experimental data, we have presented necessary standard deviations, critical results under different data quantities, and statistical correlation studies.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: For LLM-based methods, we directly call APIs, while for non-LLM methods, we describe the experimental environment in Section 4.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We are law-abiding citizens.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

The primary potential societal impacts come from the LLMs themselves, presenting challenges that require our ongoing and future collaborative efforts across the field to address.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper does not directly release new data or pre-trained models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We do not directly provide code or datasets derived from other works; however, where these are used in the paper and explicitly stated and cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Detailed introductions are presented in the readme of the Supplementary Material, and the links in Section 4 will be made public together after the double-blind phase.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our experiment does not incorporate human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our experiment does not incorporate human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our research employs LLM to address grammatical errors, enhance writing quality, and resolve specific LaTeX formatting issues.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.