Structuring the Unstructured: A Systematic Review of Text-to-Structure Generation with a Universal Evaluation Framework

Anonymous ACL submission

Abstract

The evolution of AI systems toward agentic operation and context-aware retrieval necessitates transforming unstructured text into structured formats like tables, knowledge graphs, and charts. While such conversions enable critical applications from summarization to data mining, current research lacks a comprehensive synthesis of methodologies, datasets, and metrics. This systematic review examines text-to-structure techniques and the encountered challenges, evaluates current datasets and assessment criteria, and outlines potential directions for future research. We also introduce a universal evaluation framework for structured outputs, establishing text-to-structure as foundational infrastructure for next-generation AI systems.

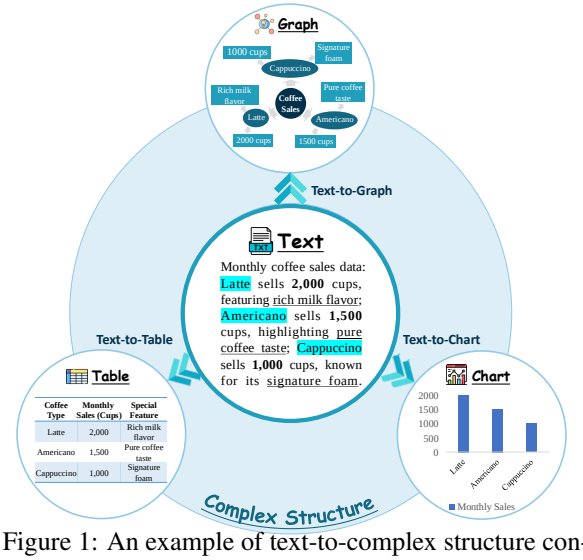


Figure 1: An example of text-to-complex structure conversion. The center represents the text, while the surrounding elements constitute the complex structures. Based on this text, three types of complex structures – table, graph, and chart – can be generated respectively.

1 Introduction

The rapid growth of the agentic AI systems is redefining the paradigm of information processing, where autonomous agents must dynamically acquire, synthesize, and act upon structured knowledge extracted from textual sources (Singh et al., 2025; Buehler, 2025). This agentic revolution creates dual demands for structured knowledge representation: (1) enabling dynamic knowledge grounding during multi-step agentic reasoning, and (2) serving as curated retrieval sources for Retrieval-Augmented Generation (RAG) pipelines (Zhang et al., 2024c, 2025; Singh et al., 2025). Complex structures, such as tables and graphs, play a crucial role in conveying information, as they can intuitively display data relationships and temporal trends (Baek et al., 2024; Huang et al., 2024b; Ghafarollahi and Buehler, 2024; Li et al., 2024a), while preserving hierarchical dependencies, which are vital for enhancing RAG reliability (Edge et al., 2024; Zhuang et al., 2024; Li et al., 2024c; Wang et al., 2024b). These structures enable both comprehension of complex information and machine-friendly representations for downstream processing (Jain et al., 2024; Reif et al., 2024).

Traditional NLP methods have demonstrated proficient performance in extracting simple entities and relationships (Sang and Meulder, 2003; Yao et al., 2019), but their efficacy is limited when confronted with complex structured knowledge implicit in the text. The emergence of Large Language Models (LLMs) enhances the capabilities of language models in capturing complex semantics and long-range dependencies, crucial for supporting agentic systems' dynamic knowledge requirements and RAG's structured retrieval needs. This advancement drives the growth of research in complex structured information extraction. Currently, the most common forms of structured output include tables, knowledge graphs, mind maps, and charts, which are crucial in downstream tasks such as RAG-enhanced question-answering and agentic data analysis (Caciularu et al., 2024; Xu et al., 2025). However, this task also poses considerable challenges: complex structured information is rep-

resented in diverse forms in the text, and accurately extracting this information requires a deep semantic understanding of the text, including identifying implicit relationships and reasoning. Moreover, there is a lack of task-specific data tailored for training and evaluation. To address these issues, researchers have explored various methods to enhance model performance and constructed several high-quality datasets for benchmarking.

Notwithstanding the significance of this task, the field currently lacks a comprehensive survey to summarize the existing research progress. This gap becomes more critical with the rise of agentic systems that demand structured knowledge representations (Zhang et al., 2024c). Specifically, two key challenges persist: (1) the baseline models and benchmark datasets used for comparison vary significantly across studies; (2) current evaluation systems remain constrained by traditional metrics that systematically misalign with human judgment when assessing structured outputs (as demonstrated in Figure 2). To bridge these gaps, our contributions are threefold:

- We conduct a systematic synthesis and critical analysis of current methods (§3), datasets (§4), and evaluation metrics (§5).
- We establish a novel Text-to-Structure (T2S) benchmark with a universal framework for structured output evaluation (§6).
- Extensive experimental results show that our framework significantly outperforms traditional metrics, providing actionable resources.

2 Task Formulation

As shown in Figure 1, we categorize Text-to-Structure generation into three common types: tables, graphs, and charts. For any task, the input should consist of a textual passage with N tokens, denoted as $\mathcal{T} = [t_1, t_2, \ldots, t_N]$, and an instruction text with M tokens, denoted as $\mathcal{I} = [i_1, i_2, \ldots, i_M]$ as constituent elements. We then provide formal definitions and detailed explanations for each category to demonstrate their distinct characteristics.

2.1 Table Generation

We define the table $\boldsymbol{T}$ as: $\boldsymbol{T} = (\mathcal{R}, \mathcal{C}, \mathcal{E}, Capt)$, where $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$ is an ordered set of m rows, $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ is an ordered set of n columns, $\mathcal{E} : \mathcal{R} \times \mathcal{C} \rightarrow \mathcal{D}$ is a function that assigns a data entry to each cell in the table, and $Capt$ denotes the caption of the table. It is allowed

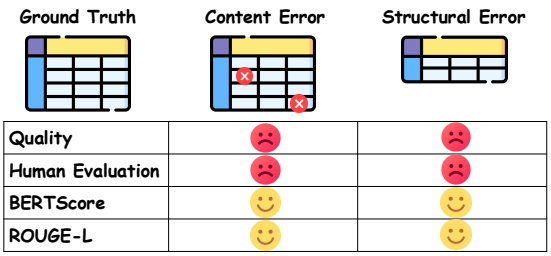


Figure 2: Misalignment between traditional metrics and true quality in structured output evaluation.

to generate multiple tables at the same time (Wu et al., 2022; Tang et al., 2024; Jain et al., 2024).

2.2 Graph Generation

A graph can be considered an extension of a table due to its greater flexibility and schema-less nature. To model more complex structures, we extend the definition of graphs to include attributes and semantics: $\boldsymbol{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{L})$, where $\mathcal{V}$ is a set of nodes representing entities or concepts, $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ is a set of edges representing relations between nodes, $\mathcal{A} : \mathcal{V} \cup \mathcal{E} \rightarrow \mathcal{P}$ is an attribute function assigning properties to nodes and edges, $\mathcal{L} : \mathcal{V} \cup \mathcal{E} \rightarrow \Sigma^*$ is a semantic function mapping nodes and edges to contextual meanings. The framework unifies graph-based structures, such as knowledge graphs and mind maps (Wei et al., 2019; Hu et al., 2021). See Appendix A for more details.

2.3 Chart Generation

A chart can be considered a visualization that goes beyond the structure of a graph. The input may include data-specific contexts, such as tables or data attributes (e.g., data types, ranges, etc.) (Tian et al., 2023). The goal is to generate visualization charts that feature the appropriate specifications, including the type of the chart, the visual encodings, and other relevant details (Zhang et al., 2024b).

3 Methods

3.1 Fine Tuning

Supervised Fine Tuning (SFT) is still the default approach to text-to-structure generation since it only necessitates paired *text→target* instances. Although intuitive, its performance differs across tasks because of variances in data needs, structural intricacy, and computational expenses. In text-to-table generation, sequence-to-sequence models with header-aware serialization (Wu et al., 2022; Li et al., 2023) produce validated tables but need enormous annotated datasets, which restricts low-resource applicability. QA-style cell queries (Sun-

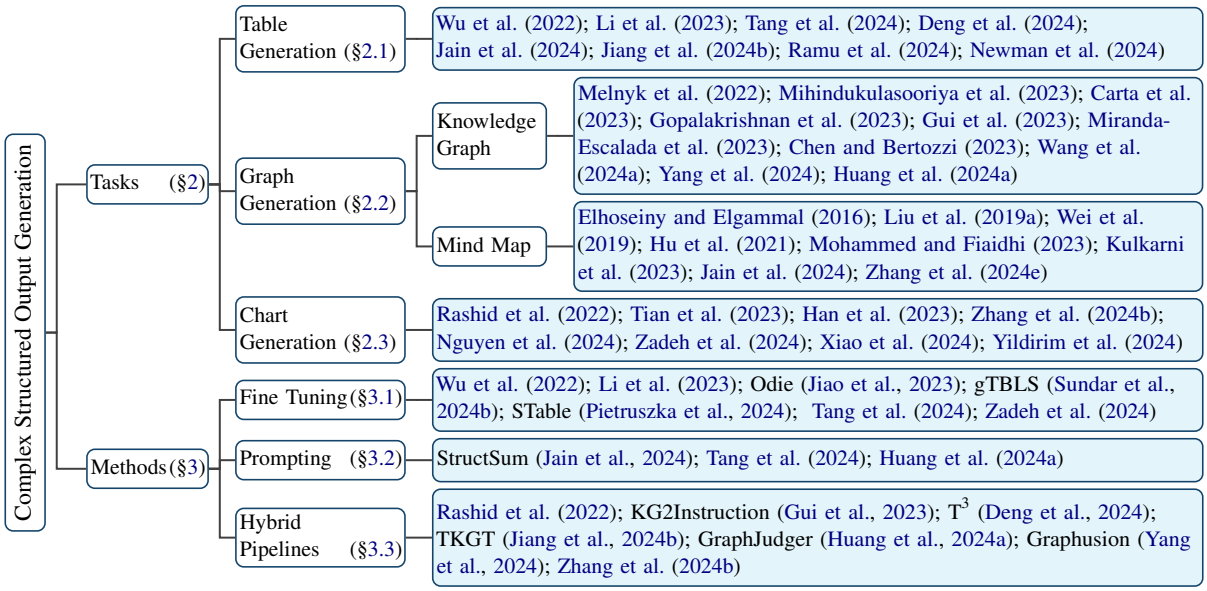


Figure 3: Taxonomy of representative works in text-to-structured generation categorized by tasks and methods.

dar et al., 2024a,b) decrease data requirements, while being challenged by intricate layouts. Instruction tuning with Chain-of-Thought prompting (Wei et al., 2022) and LLM-synthesized pairs increases robustness but includes computational expense (Jiao et al., 2023; Tang et al., 2024). For text-to-KG generation, SFT tends to follow a two-step pipeline, entity recognition followed by relation extraction (Gui et al., 2023; Wang et al., 2024a; Yang et al., 2024), which enhances structure but has a risk of error propagation. In text-to-chart tasks, instruction tuning with RL-based feedback makes it more adaptable, although performance relies on the quality of the feedback, particularly for uncommon chart types (Xia et al., 2024; Zadeh et al., 2024; Koh et al., 2025).

3.2 Prompting

The prosperity of LLMs (Dubey et al., 2024; OpenAI, 2024; DeepSeek-AI et al., 2025) has triggered the evolution of diverse prompting methods for text-to-structure tasks (Wei et al., 2022; Dua et al., 2022; Khot et al., 2023), with each presenting trade-offs in accuracy, efficiency, and simplicity. Basic prompts such as "`generate tables/graphs/charts`" are quick but have poor results (Jain et al., 2024; Huang et al., 2024a). Incorporating schema constraints enhances structural fidelity at the expense of potentially reducing flexibility in tasks with implicit relations (Tang et al., 2024; Deng et al., 2024; Jain et al., 2024). Chain-of-Thought prompts are also used to enhance reasoning and accuracy (Wang et al., 2023b; Tang et al., 2024). Decomposed prompting improves accuracy by breaking up input (Khot et al., 2023; Jain et al., 2024), but can lose global context, hurting coherence in applications such as cross-document structure construction.

3.3 Hybrid Pipeline

3.3.1 Table Generation

Harnessing table's natural triple structure (row, column, cell) (Deng et al., 2024; Sundar et al., 2024b), it is common to let LLMs perform either tuple extraction from text (Wang et al., 2023a), or decomposition of text into atomic propositions (Hoyle et al., 2023), which subsequently serves as input for three approaches: *Code Aggregation* (Rozière et al., 2023; Deng et al., 2024) exploits LLM's coding versatility for efficient data integration but falters on complicated structures and scalability. *KG with RAG* builds knowledge graphs for noise-robust RAG-based generation at the cost of flexibility and scalability (Jiang et al., 2024b). *Schema-Based Inference* models both qualitative nuances and quantitative facts despite similar scalability limitations(Ahuja et al., 2025).

3.3.2 Graph Generation

KG construction begins with Named Entity Recognition (NER), Relation Extraction (RE), and disambiguation (He and Choi, 2021; Gui et al., 2023), but is susceptible to noise and hallucinations. Therefore, structure-based denoising aids reliability (Xie et al., 2018; Deng et al., 2023), and LLM-guided verification boosts accuracy (Liu et al., 2023a; Huang et al., 2024a). Clustering-based node merging alleviates sparsity but threatens semantic over-simplification (Mo et al., 2025). Mindmaps (Jain

Dataset	Size	Domain	T2S	Annot.	Sch.	Text Unit			Complexity		GTS
						Sent.	Para.	Doc.	Reas.	Struct.	
Text-Table Datasets											
WikiBio (Lebret et al., 2016)	728K	Open-Domain	✗	✗	∞	✓			★	❊	✗
E2E (Novikova et al., 2017)	50.6K	Restaurant	✗	✓	□		✓		★	❊	✓
Rotowire (Wiseman et al., 2017)	4.9K	Sports	✗	✓	□			✓	★★	❊❊	✓
WikiTableText (Bao et al., 2018)	5.0K	Open-Domain	✗	✗	∞		✓		★	❊	✓
MLB (Puduppully et al., 2019)	26.3K	Sports	✗	✓	□		✓		★★	❊❊	✓
WikiTablePara (Laha et al., 2019)	171	Open-Domain	✗	✓	∞		✓		★	❊❊	✓
WikiTableT (Chen et al., 2021)	1.5M	Open-Domain	✗	✓	∞		✓		★	❊❊	✗
InstructIE (Jiao et al., 2023)	14.7K	Open-Domain	✓	✓	∞		✓		★★	❊❊	✓
arxivDIGESTables (Newman et al., 2024)	2.2K	arXiv	✓	✓	∞			✓	★★★	❊❊	✓
CPL (Jiang et al., 2024b)	850	Law†	✓	✓	∞			✓	★★	❊❊	✓
CT-Eval (Shi et al., 2024)	86.6K	Open-Domain†	✓	✓	∞		✓		★	❊❊	✓
DescToTTo (Ramu et al., 2024)	1.3K	Open-Domain	✓	✓	∞		✓		★	❊❊❊	✓
LiveSum (Deng et al., 2024)	3.8K	Sports	✓	✓	□			✓	★★★	❊	✓
Struc-Bench Table (Tang et al., 2024)	4.1K	Sports	✗	✓	□			✓	★★	❊❊	✓
Text-Graph Datasets											
WebNLG (Gardent et al., 2017)	25.3K	Open-Domain	✗	✓	□		✓		★	❊	✓
SCIERC (Luan et al., 2018)	500	AI Papers	✗	✓	□		✓		★★	❊❊	✓
AGENDA (Koncel-Kedziorski et al., 2019)	40K	AI Papers	✗	✗	□		✓		★★	❊❊	✗
GenWiki (Jin et al., 2020)	1.3M	Open-Domain	✗	✗	□		✓		★	❊	✓
DART (Nan et al., 2021)	82.2K	Open-Domain	✗	✓	□		✓		★★	❊	✓
EventNarrative (Colas et al., 2021)	224K	EventKG	✗	✗	□		✓		★★★	❊❊❊	✗
KeLM (Agarwal et al., 2021)	18M	Open-Domain	✗	✗	□	✓			★★	❊	✗
REBEL (Cabot and Navigli, 2021)	879K	Open-Domain	✗	✗	□	✓			★★	❊	✗
Text2MindMap (Hu et al., 2021)	44.6K	News	✓	✓	□			✓	★★	❊❊	✓
InstructIE (Gui et al., 2023)	362K	Open-Domain†	✓	✗	□		✓		★★	❊❊	✓
MINE (Mo et al., 2025)	100	Open-Domain	✓	✗	∞			✓	★★★	❊❊❊	✗
Text-Chart Datasets											
AutoChart (Zhu et al., 2021)	10.2K	Data Analysis	✗	✓	□		✓		★★	❊	✗
Pew (Kantharaj et al., 2022)	9.3K	Society & Policy	✗	✓	□		✓		★★★	❊❊❊	✓
Statista (Kantharaj et al., 2022)	34.8K	Market & Stats	✗	✓	□		✓		★★	❊❊	✓
Text2Chart (Rashid et al., 2022)	717	Data Analysis	✓	✓	□		✓		★★	❊❊	✓
ChartX (Xia et al., 2024)	48K	Open-Domain	✗	✓	□		✓		★★★	❊❊❊	✓
Text2Chart31 (Zadeh et al., 2024)	11.1K	Data Analysis	✓	✗	□		✓		★★★	❊❊❊	✓

Table 1: Comparison and classification of existing text-to-structure generation benchmarks across key dimensions. **Size** denotes the number of (text, data) pairs in the dataset, including the training set. **Domain** shows the text source domain. **T2S** (Text-to-Structure) specifies whether the dataset is specifically designed for text-to-structure tasks. **Annot.** (Annotation) indicates human annotation involvement. **Sch.** (Schema) shows schema limitedness ('□' for limited, '∞' for unlimited). **Text Unit** shows the granularity level of text (sentence / paragraph / document). **Reas.** (Reasoning) indicates the complexity of reasoning required for text-to-structure conversion, and **Struct.** (Structure) shows the structural complexity of the data, where a higher number of '★' or '❊' indicates greater complexity in the corresponding dimension. **GTS** (Gold Test Set) denotes manual verification of the test set. '†' indicates the presence of Chinese text in the dataset. See Appendix B.1 for more details.

et al., 2024), on the other hand, take a hierarchical approach: rooted iterative prompting guarantees structural clearness but at the expense of the dense interlinking that is characteristic of KGs.

3.3.3 Charts Generation

Text-to-chart generation generally employs a two-phase pipeline: *Data point identification*, which has inherent trade-offs: table-based extraction is structured but inflexible (Rashid et al., 2022), while text-based approaches are more adaptable but may be inconsistent (Zhang et al., 2024b); *Vision encoding*, which suffers compounded constraints: unclear data hinders chart-type prediction, while visualization synthesis amplifies earlier mistakes (Rashid et al., 2022; Zhang et al., 2024b; Zadeh et al., 2024). Both phases necessitate more resilient approaches.

4 Datasets

Table 1 provides a comprehensive list of all currently available benchmarks and contrasts them along eight axes. Most structure-to-text datasets can be adapted for text-to-structure tasks (Wiseman et al., 2017; Obeid and Hoque, 2020; Zhu et al., 2021; Kantharaj et al., 2022; Zhao et al., 2023; Lin et al., 2024), although information loss sometimes hinders suitability (Nie et al., 2018; Chen et al., 2020; Parikh et al., 2020), prompting dedicated text-to-structure dataset creation. Datasets are categorized by human supervision, schema constraints, text length, and exclusive human-annotated test sets. Complexity is quantified via structural complexity of outputs and reasoning complexity in generation (see Appendix B for further details). Analysis reveals significant gaps: *Text-Table* datasets dominate, yet few demand high complexity; *Text-Graph* datasets generally lack crucial human supervision; and *Text-Chart* datasets are scarce. Future work should prioritize annotated, high-complexity datasets, especially for underrepresented graph and chart domains, to ensure balanced progress.

5 Evaluation Metrics

Evaluating complex structured outputs poses multifaceted challenges distinct from traditional IE tasks. While the latter often benefit from well-defined formats and objective standards (Xu et al., 2024), the inherent diversity and complexity of structured outputs necessitate a more comprehensive evaluation framework. As shown in Figure 4, after generating *structured output* from the *original text*, prevailing methods bifurcate into: (1) ***Direct Evaluation*** jointly assesses the *ground truth*, *original text*, and *structured output* (§5.1, §5.2, §5.3); (2) ***Indirect Evaluation*** first generates intermediate *content* (e.g., propositions, qustion-answer pairs) from the *original text*, then uses this *content* and the *structured output* to produce further *content* (e.g., answers), and finally compares the two *generated content* sets (§5.3). Through analysis of this established framework, we identify limitations that motivate our novel evaluation paradigm.

5.1 Human-based Evaluation

Human-based evaluation remains the gold standard, assessing content coverage, structural validity, and factual consistency (Jiao et al., 2023; Tang et al., 2024; Jain et al., 2024). It can also be evaluated by manual pairwise comparison (Zadeh et al., 2024), processing pipeline stages (Rashid et al., 2022; Deng et al., 2024), or indirectly via comprehension tasks (Jain et al., 2024). While human evaluation provides accurate results, it is costly in terms of time and resources, highlighting the complexity of quality that automated metrics must capture.

5.2 Rule-based Evaluation

Metrics like Exact Match, ROUGE-L (Lin, 2004), Levenshtein Distance (Haldar and Mukhopadhyay, 2011), and chrF (Popovic, 2015) rely on predefined rules focusing on surface-level features: token overlap, sequence edits, or n-gram alignment (Post, 2018; Nalawati and Dini Yuntari, 2021; Wu et al., 2022; Tang et al., 2024). Although ROUGE-L shows the strongest correlation with human evaluation among these metrics (Li et al., 2023; Wang et al., 2023c), it still fails to capture semantic and structural aspects, leading to misalignment in structured output assessment despite its convenience.

5.3 Model-based Evaluation

Direct Scoring Learning-based methods use pre-trained models like SentenceBERT (Reimers and Gurevych, 2019), BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), and BARTScore (Yuan et al., 2021) to evaluate text by semantic and contextual similarity, demonstrating better alignment with human judgements than token-matching rules on semantics (Jiao et al., 2023; Tang et al., 2024; Deng et al., 2024). However, these models remain inadequate for assessing structural correctness, as they lack explicit schema understanding and relational reasoning, and are insensitive to fine-grained content variations critical for quality assessment. On the other hand, direct prompting of LLMs partially addresses this gap by evaluating content-structure similarity via Chain-of-Thought prompting (Chiang and Lee, 2023; Liu et al., 2023b; Tang et al., 2024) although its reliability depends on prompt design and model biases.

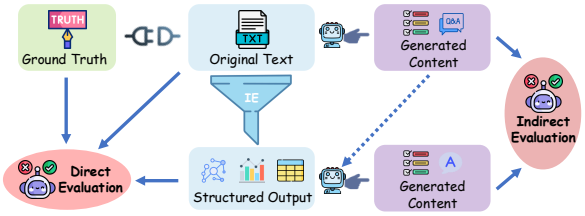


Figure 4: Comparison of direct evaluation and indirect evaluation methods for structured outputs.

Indirect Scoring Complex structured outputs often lack unique ground truth due to valid variations, like row/column order, equivalent values, or complete absence (Guo et al., 2020; Li et al., 2023; Tang et al., 2024; Jain et al., 2024), necessitating automated quality assessment methods: NLI-based alignment provides superior robustness to structural diversity than rigid rule-matching (Liu et al., 2019b; Ramu et al., 2024); QA-based evaluation (Jain et al., 2024; Deng et al., 2024) better assesses how well the structured outputs retain original information. Although these indirect evaluation approaches are more robust and generalizable, they demand stronger reasoning capabilities from models, such as the techniques from TableQA (Zhang et al., 2024d; Wang et al., 2024c), KGQA (Jiang et al., 2023b; Luo et al., 2024; Jin et al., 2024), and chart reasoning (Masry et al., 2023; Akhtar et al., 2023; Masry et al., 2024; Zhang et al., 2024a) to avoid error propagation while assessing outputs.

6 Text-to-Structure (T2S) Benchmark

To address the limitations of existing evaluation frameworks summarized in Section 5, we propose a novel framework capable of accurately and comprehensively evaluating structured outputs. Qualita-

tively, we address the previous vague criteria by implementing concrete metrics: an F1-style semantic metric using LLM-as-judge for **content** evaluation, overcoming limitations of string-based metrics like ROUGE and BERTScore (Tang et al., 2024), and defining two new orthogonal dimensions for **structure**, *alignment* for format consistency and *clarity* for unambiguous representations, inspired by header evaluation of prior works (Wu et al., 2022; Jiao et al., 2023) and Jain et al. (2024). To validate our framework's superiority, we introduce a novel Text-to-Structure (T2S) benchmark, detailing: criteria (§6.1), datasets and models (§6.2), results (§6.3), and validation of effectiveness (§6.4).

6.1 Scoring Criteria

We evaluate structured outputs across two dimensions: *content faithfulness* and *structure coherence*.

Content Faithfulness Measures the accuracy and completeness of the generated information, ensuring it reflects the source text without errors or omissions. It uses an F1-style metric combining: (1) ***precision***, which identifies conflicts between output and reference data (e.g., factual errors or semantic mismatches) and (2) ***recall***, which measures how thoroughly the output captures key details from the source. These are combined into a single score via harmonic mean:

$$\text{FAITHFULNESS} = 2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

Structure Coherence Evaluates logical organization and interoperability of the output. It examines (1) ***alignment*** with expected formats (e.g., schema consistency, data types) and (2) ***clarity*** of presentation (e.g., unambiguous representation, redundancy-free content). Again, we employ the harmonic mean for these two orthogonal metrics, as it penalizes imbalance and more accurately reflects joint effectiveness:

$$\text{COHERENCE} = 2 \cdot \frac{\textit{alignment} \cdot \textit{clarity}}{\textit{alignment} + \textit{clarity}}$$

We employ LLM-as-judge with detailed scoring rubrics (0-100 scale) for all four sub-metrics: *precision*, *recall*, *alignment*, and *clarity*. Detailed explanations and implementation are in Appendix C.1.

6.2 Datasets and Models

We carefully select six datasets from the openly available sources listed in Table 1, aiming to encompass a diverse range of complexities, cover multiple languages, and prioritize open-domain datasets. Specifically, three datasets for text-to-table generation: InstructIE, Struc-Bench, and LiveSum; two for text-to-graph generation: DART and InstructIE; and Text2Chart31 for text-to-chart generation. The 3:2:1 split mirrors the actual distribution of publicly downloadable datasets in Table 1. Within each group, we greedily sampled datasets that maximally differ in domain, text unit, and complexity, keeping the total to six for annotator-budget feasibility. We select several mainstream LLMs for evaluation, which can be divided into two groups: non-thinking models (e.g., Llama, GPT) and thinking models (e.g., R1, O4-mini). For the former group, we also evaluate multiple models under SFT. Other models are evaluated in a zero-shot setting. For non-thinking models, we perform inference at temperature 0; for thinking models, we use the documented recommended hyperparameters, averaging results across five runs. When employing DeepSeek-R1 and GPT-4o as LLM-as-judge, we fix the temperature parameter at 0 to ensure that every evaluation remains deterministic. Further details are provided in Appendix C.2 and C.3.

6.3 Results

In the Text-to-Table task, LiveSum emerges as the most challenging benchmark in terms of content faithfulness, with all evaluated models (except O4-mini) scoring no higher than 70; InstructIE is intermediate, while Struc-Bench is the simplest, achieving nearly 90. Regarding structure coherence, both LiveSum and Struc-Bench achieve higher scores due to their predefined headers, reflecting strong instruction-following capabilities. In contrast, the absence of fixed headers in InstructIE prompts models to generate more diverse table formats, resulting in lower coherence scores. In the Text-to-Graph task, the DART dataset demonstrates higher faithfulness compared to InstructIE because both datasets, originally designed for KG-to-text, contain triples often irretrievable from text and this limitation is more pronounced in InstructIE and leads to lower recall. Meanwhile, both datasets show moderate performance in structure coherence, primarily due to challenges in entity and relation naming. In the Text-to-Chart task, larger models generally exhibit superior performance, consistently achieving scores above 80, while smaller models trained with SFT lag significantly behind. For further analysis, see Appendix D.

Model	*Table*						*Graph*				*Chart*	
	InstructIE (Jiao et al., 2023)		**Struc-Bench** (Tang et al., 2024)		**LiveSum** (Deng et al., 2024)		**DART** (Nan et al., 2021)		**InstructIE** (Gui et al., 2023)		**Text2Chart31** (Zadeh et al., 2024)	
	FAITH.	COH.	FAITH.	COH.	FAITH.	COH.	FAITH.	COH.	FAITH.	COH.	FAITH.	COH.
♦ ***SFT***												
Mistral-7B	58.08	68.85	87.65	90.05	53.31	98.54	75.07	**83.17**	44.58	69.67	64.54	63.41
Llama-3.1-8B	59.31	66.44	85.01	85.95	61.34	90.30	38.94	55.13	34.93	64.99	77.06	77.25
R1-Distill-Llama-8B	62.86	68.08	54.56	58.97	43.27	86.22	59.03	69.43	37.27	69.73	41.66	40.45
♦ ***Zero-Shot***												
Phi-3-medium	29.15	39.89	74.04	83.99	41.61	91.73	64.60	72.59	43.40	71.44	60.24	60.45
Phi-3-medium (CoT)	30.36	37.23	73.42	81.06	46.36	93.19	64.88	72.32	41.28	70.44	57.75	58.27
Mistral-7B	53.00	59.73	56.19	67.05	30.14	87.26	70.27	73.06	42.74	66.54	52.25	52.03
Mistral-7B (CoT)	54.02	60.37	56.25	70.34	41.27	85.97	64.68	70.96	33.35	66.01	45.13	46.51
Mixtral-8x22B	56.35	68.25	84.16	85.20	49.12	87.91	65.70	71.07	40.07	68.89	61.62	62.15
Mixtral-8x22B (CoT)	58.27	68.29	80.55	74.94	54.87	87.37	71.27	74.53	47.47	71.29	58.92	58.89
Llama-3.1-405B	63.88	69.63	**89.43**	90.43	62.29	99.11	68.74	75.61	46.72	75.82	86.49	86.62
Llama-3.1-405B (CoT)	64.54	72.80	89.38	90.74	66.25	98.64	72.31	78.09	47.94	78.48	87.23	86.40
Llama-3.3-70B	62.88	69.19	80.58	80.00	58.48	**99.19**	71.09	77.54	47.11	76.30	87.44	87.12
Llama-3.3-70B (CoT)	65.34	69.46	86.47	85.84	62.79	98.68	77.11	78.99	46.56	78.51	**88.78**	87.83
GPT-3.5-turbo	55.09	69.01	78.44	85.37	45.38	79.18	69.75	76.72	41.89	73.34	85.17	84.79
GPT-3.5-turbo (CoT)	58.44	68.80	71.88	81.54	45.09	86.37	70.17	75.99	41.78	73.51	83.06	83.62
GPT-4o	70.72	**76.36**	87.35	92.22	63.21	91.99	74.15	77.18	46.77	77.00	85.48	86.01
GPT-4o (CoT)	68.11	74.40	86.05	90.03	62.16	98.50	**79.89**	79.33	53.28	77.65	85.99	86.63
QwQ-32B	68.23	68.79	81.21	87.12	58.06	96.20	69.42	78.02	56.30	82.15	86.62	86.71
DeepSeek-R1	**76.86**	74.75	87.77	91.05	64.89	96.85	78.25	79.89	52.15	79.23	86.07	86.90
Grok-3-mini	75.57	75.40	85.60	**92.28**	66.35	95.26	73.32	80.34	56.03	**82.68**	87.63	**88.65**
O4-mini	75.71	74.85	83.45	90.02	**84.48**	99.16	74.82	80.19	**56.46**	82.32	85.21	85.93

Table 2: The performance of major LLMs across multiple datasets. The highest results are **bolded**, with numerical scores categorized into five color-coded intervals: [90, 100] - excellent, [80, 90) - strong, [70, 80) - moderate, [60, 70) - limited, and [0, 60) - insufficient, using distinct chromatic gradients for visual differentiation.

Task	Metric	Faithfulness			Coherence		
		P(r)	S(ρ)	K(τ)	P(r)	S(ρ)	K(τ)
Table	ROUGE-L	0.267	0.227	0.166	**0.607**	**0.617**	0.452
	BERTScore	0.238	0.198	0.142	**0.562**	**0.576**	0.415
Graph	ROUGE-L	0.237	0.118	0.093	0.318	0.165	0.127
	BERTScore	0.100	0.043	0.036	0.211	0.105	0.078
Chart	CodeBLEU	0.249	0.186	0.144	0.266	0.297	0.220
	METEOR	0.281	0.241	0.186	0.317	0.376	0.274

Table 3: Correlation coefficients (Pearson/Spearman/Kendall) between traditional metrics and human evaluation across two dimensions on three different tasks.

6.4 Evaluation of Scoring Criteria

To demonstrate how our proposed scoring criteria address evaluation limitations, we aim to answer the following questions:

RQ1. Do traditional metrics align with human judgment criteria? (§ 6.4.1)

RQ2. Can LLM-based evaluation ensure consistent and nuanced assessment? (§ 6.4.2)

Furthermore, we provide a case study analyzing traditional metric failures and our framework's strengths (§6.4.3).

6.4.1 Validity Gap (RQ1)

For table generation and graph generation tasks, prior works predominantly adopt ROUGE-L and BERTScore by extracting textual content from structured outputs and comparing them with ground truth references (Wu et al., 2022; Jiao et al., 2023; Huang et al., 2024a). For chart generation tasks, existing approaches typically evaluate code similarity between generated and reference codes using metrics like METEOR and CodeBLEU (Zadeh et al., 2024). To investigate whether these metrics align with human evaluation, we employ three annotators to score outputs based on our two orthogonal criteria, *Faithfulness* and *Coherence*, with final scores averaged across annotators. We conduct correlation analyses between each automated metric and human-rated *Faithfulness* and *Coherence* following established methodologies (Jiao et al., 2023), with results summarized in Table 3. The table reveals that only in table generation tasks do ROUGE-L and BERTScore exhibit moderate correlations with Coherence ($r, \rho > 0.5$), whereas all other metrics exhibit weak alignment with human evaluation. This demonstrates traditional metrics' limitations and underscores the urgent need for specialized metrics capable of reliable quality assessment in this domain.

6.4.2 Reliability and Discriminability (RQ2)

To investigate the alignment between LLM-based scoring and human evaluation under our proposed criteria, three annotators score 30 randomly sampled instances per dataset following the rubric outlined in the evaluation prompt. The Root Mean

Task	Model	Prec.	Rec.	Align.	Clr.
Table	GPT-4o	14.90	15.79	23.76	15.14
	DeepSeek-R1	**9.37**	**8.59**	**8.32**	**4.55**
Graph	GPT-4o	14.76	11.08	7.16	7.42
	DeepSeek-R1	**4.91**	**8.91**	**7.09**	**3.08**
Chart	GPT-4o	15.32	14.77	6.12	7.91

Table 4: Root Mean Squared Error (RMSE) between human and LLM scores across three task categories.

Metric	Faithfulness		Coherence	
	Precision	Recall	Alignment	Clarity
Pearson (r)	0.745	0.771	0.761	0.728
Spearman (ρ)	0.747	0.766	0.712	0.679
Kendall (τ)	0.664	0.684	0.670	0.621

Table 5: Correlation coefficients (Person / Spearman / Kendall) of human and LLM scores for four evaluation metrics.

Square Error (RMSE) between human and LLM scores is averaged across three task categories and aggregated in Table 4, which reveals that DeepSeek-R1 demonstrates superior accuracy compared to GPT-4o, with subsequent case studies confirming that its enhanced reasoning fidelity directly contributes to reduced prediction errors. Given DeepSeek-R1's current lack of multimodal input capabilities, GPT-4o remains our evaluator for charts to maintain assessment consistency. We further analyze the correlation between human ratings and DeepSeek-R1 scores, computing PearsonâĂŹs r, SpearmanâĂŹs ρ, and KendallâĂŹs τ coefficients, with results summarized in Table 5. All four evaluation metrics exhibit strong positive correlations with human judgments. To compare our proposed metrics with conventional metrics, we evaluated five models on the InstructIE (Jiao et al., 2023) dataset using both traditional metrics (ROUGE and BERTScore) and LLM-based assessments (FAITH. and COH. scores from GPT-4o and DeepSeek-R1). The results in Table 6 demonstrate that DeepSeek-R1 scores show a high positive correlation with GPT-4o, eliminating concerns about LLM evaluators favoring their own outputs. While GPT-3.5-turbo achieves the second-highest scores in traditional metrics, it ranks last across all LLM-based evaluations, validating our metrics' capability to detect reliability flaws obscured by conventional measurements.

6.4.3 Case Study

We provide five carefully examined case studies tackling important shortcomings in automatic evaluation metrics in Appendix E. Our examination uncovers systematic misalignment situations where conventional metrics (BERTScore, ROUGE, etc.)

Model	ROUGE	BERT.	FAITH_{R1}	FAITH_{4o}	COH_{R1}	COH_{4o}
Llama-3.1-405B	0.561	0.742	63.88	69.53	69.63	76.66
Llama-3.3-70B	0.517	0.738	62.88	67.23	69.19	75.08
GPT-3.5-turbo	0.586	0.770	55.09	62.86	69.01	75.82
GPT-4o	**0.613**	**0.772**	70.72	73.07	**76.36**	**81.44**
DeepSeek-R1	0.541	0.738	**76.86**	**77.02**	74.75	79.81

Table 6: Performance comparison of five LLMs on InstructIE (Jiao et al., 2023) dataset through conventional metrics and LLM-based evaluation, and we **bold** the best performance for each metric.

incorrectly prefer outputs with: structural incompleteness (Case 1), factual hallucinations (Cases 2, 3), semantic disarray (Case 4), and numerical or logical errors (Cases 3, 5), whereas our model accurately predicts outputs, maintaining content faithfulness and structural coherence.

7 Future Opportunities

Future work may enhance the text-to-structure generation of agentic AI from multiple directions. First, optimizing real-time conversion with efficient algorithms and lightweight models can support dynamic environments, which is critical for tasks such as question answering and summarization. Second, improving accuracy and consistency through multimodal learning (e.g., integrating text with images) and self-reflection mechanisms can reduce errors in complex texts, thereby benefiting the information state conversion of agentic information retrieval. Third, reinforcement learning with sophisticated reward functions – balancing utility, efficiency, and ethics – can combine structured outputs with external goals to enhance autonomous decision-making capabilities. Finally, multi-task learning frameworks can simultaneously generate different structures, such as tables and charts, with consistency and flexibility. These efforts aim to improve the autonomy, accuracy, and applicability of agentic AI in complex real-world scenarios.

8 Conclusion

In conclusion, this work aims to provide a comprehensive review of currently existing methods, benchmark datasets, and evaluation metrics for text-to-structure extraction by designing a universal evaluation framework specifically tailored for structured outputs. Through these contributions, we not only synthesize current research progress but also elucidate how these technologies can empower agentic AI systems to achieve enhanced autonomy and efficiency in dynamic real-world environments.

Limitations

The limitations of this study mainly stem from two aspects. First, since we strategically focus the most representative studies in this field, the scope of the literature review may not comprehensively cover all research advances related to complex structures. Second, given the rapid development of LLMs, the Text-to-Structure benchmark evaluation may not cover the latest published LLMs. However, despite our best efforts to survey a comprehensive selection of important papers and mainstream LLMs, the practical constraints of a single submission inherently limit our ability to cover every aspect or all LLM variants. Future research directions should aim for a broader literature review while integrating datasets from cross-disciplinary studies to enhance their generalizability.

Ethics Statement

Our paper presents a comprehensive survey of text-to-complex structure extraction, with a specific focus on tables, graphs, and charts. The datasets and models employed in this study are all open-source. All datasets are used in their pre-anonymized forms as released by the original creators, complying with established licensing agreements. Our evaluation uses strictly unaltered benchmark data, applying only random sampling to select representative subsets. Crucially, no additions, modifications, or external data curation occur at any processing stage, preserving original dataset integrity while enabling efficient LLM assessment. The annotation was strictly limited to validating sampled structural outputs, conducted voluntarily by three doctoral researchers from the authors' institution. Therefore, to the best of the authors' knowledge, we believe that this study introduces no additional risk.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, SÃl'bastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio CÃl'sar Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3554–3565. Association for Computational Linguistics.

Naman Ahuja, Fenil Denish Bardoliya, Chitta Baral, and Vivek Gupta. 2025. Map&make: Schema guided text to table generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30249–30262. Association for Computational Linguistics.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 399–414. Association for Computational Linguistics.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *CoRR*, abs/2404.07738.

Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. 2018. Table-to-text: Describing table region with natural language. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-*

18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 5020–5027. AAAI Press.

Markus J. Buehler. 2025. Agentic deep graph reasoning yields self-organizing knowledge networks. *CoRR*, abs/2502.13025.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2370–2381. Association for Computational Linguistics.

Avi Caciularu, Alon Jacovi, Eyal Ben-David, Sasha Goldshtein, Tal Schuster, Jonathan Herzig, Gal Elidan, and Amir Globerson. 2024. TACT: advancing complex aggregative reasoning with information extraction tools. *CoRR*, abs/2406.03618.

Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. 2023. Iterative zero-shot LLM prompting for knowledge graph construction. *CoRR*, abs/2307.01128.

Bohan Chen and Andrea L. Bertozzi. 2023. Autokg: Efficient automated knowledge graph generation for language models. In *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, pages 3117–3126. IEEE.

Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2021. Wikitablet: A large-scale data-to-text dataset for generating wikipedia article sections. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 193–209. Association for Computational Linguistics.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7929–7942. Association for Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15607–15631. Association for Computational Linguistics.

Anthony M. Colas, Ali Sadeghian, Yue Wang, and Daisy Zhe Wang. 2021. Eventnarrative: A large-scale event-centric dataset for knowledge graph-to-text generation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9300–9322. Association for Computational Linguistics.

Zheye Deng, Weiqi Wang, Zhaowei Wang, Xin Liu, and Yangqiu Song. 2023. Gold: A global and local-aware denoising framework for commonsense knowledge graph noise detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3591–3608. Association for Computational Linguistics.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. 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 1251–1265. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130.

Mohamed Elhoseiny and Ahmed M. Elgammal. 2016. Text to multi-level mindmaps - A novel method for hierarchical visual abstraction of natural language text. *Multim. Tools Appl.*, 75(8):4217–4244.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation, INLG 2017, Santiago de Compostela, Spain, September 4-7, 2017*, pages 124–133. Association for Computational Linguistics.

Alireza Ghafarollahi and Markus J. Buehler. 2024. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *CoRR*, abs/2409.05556.

Seethalakshmi Gopalakrishnan, Victor Zitian Chen, Wenwen Dou, Gus Hahn-Powell, Sreekar Nedunuri, and Wlodek Zadrozny. 2023. Text to causal knowledge graph: A framework to synthesize knowledge from unstructured business texts into causal graphs. *Inf.*, 14(7):367.

Honghao Gui, Shuofei Qiao, Jintian Zhang, Hongbin Ye, Mengshu Sun, Lei Liang, Huajun Chen, and Ningyu Zhang. 2023. Instructie: A bilingual instruction-based information extraction dataset. *CoRR*, abs/2305.11527.

Mandy Guo, Zihang Dai, Denny Vrandecic, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2440–2452. European Language Resources Association.

Rishin Haldar and Debajyoti Mukhopadhyay. 2011. Levenshtein distance technique in dictionary lookup methods: An improved approach. *CoRR*, abs/1101.1232.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal LLM for chart understanding and generation. *CoRR*, abs/2311.16483.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5555–5577. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13188–13214. Association for Computational Linguistics.

Mengting Hu, Honglei Guo, Shiwan Zhao, Hang Gao, and Zhong Su. 2021. Efficient mind-map generation via sequence-to-graph and reinforced graph refinement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8130–8141. Association for Computational Linguistics.

Haoyu Huang, Chong Chen, Conghui He, Yang Li, Jiawei Jiang, and Wentao Zhang. 2024a. Can llms be good graph judger for knowledge graph construction? *Preprint*, arXiv:2411.17388.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024b. Mlagentbench: Evaluating language agents on machine learning experimentation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Parag Jain, Andreea Marzoca, and Francesco Piccinno. 2024. STRUCTSUM generation for faster text comprehension. 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 7876–7896. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, LÃl'lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, ThÃl'ophile Gervet, Thibaut Lavril, Thomas Wang, TimothÃl'e Lacroix, and William El Sayed. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023b. Structgpt: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9237–9251. Association for Computational Linguistics.

Peiwen Jiang, Xinbo Lin, Zibo Zhao, Ruhui Ma, Yvonne Chen, and Jinhua Cheng. 2024b. TKGT: redefinition and A new way of text-to-table tasks based on real world demands and knowledge graphs augmented llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 16112–16126. Association for Computational Linguistics.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10030–10051. Association for Computational Linguistics.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, and Jiawei Han. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 163–184. Association for Computational Linguistics.

Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020. Genwiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2398–2409. International Committee on Computational Linguistics.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq R. Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4005–4023. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Woosung Koh, Janghan Yoon, MinHyung Lee, Youngjin Song, Jaegwan Cho, Jaehyun Kang, Taehyeon Kim, Se-Young Yun, Youngjae Yu, and Bongshin Lee. 2025. c^2: Scalable auto-feedback for LLM-based chart generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4525–4566, Albuquerque, New Mexico. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2284–2293. Association for Computational Linguistics.

Aditya Kulkarni, Hetansh Shah, Lynette DâĂŹMello, and Krish Shah. 2023. Flowchart generation and mind map creation using extracted summarized text. In *2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, pages 1–6.

Anirban Laha, Parag Jain, Abhijit Mishra, and Karthik Sankaranarayanan. 2019. Scalable micro-planned generation of discourse from structured data. *Comput. Linguistics*, 45(4):737–763.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1203–1213. The Association for Computational Linguistics.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024a. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. *CoRR*, abs/2410.13185.

Tong Li, Zhihao Wang, Liangying Shao, Xuling Zheng, Xiaoli Wang, and Jinsong Su. 2023. A sequence-to-sequence&set model for text-to-table generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5358–5370. Association for Computational Linguistics.

Yinghao Li, Rampi Ramprasad, and Chao Zhang. 2024b. A simple but effective approach to improve structured language model output for information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 5133–5148. Association for Computational Linguistics.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2024c. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. *CoRR*, abs/2410.08815.

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.

Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2024. A survey on neural data-to-text generation. *IEEE Trans. Knowl. Data Eng.*, 36(4):1431–1449.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023a. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1264–1287. Association for Computational Linguistics.

Ruixue Liu, Baoyang Chen, Meng Chen, Youzheng Wu, Zhijie Qiu, and Xiaodong He. 2019a. Mappa mundi: An interactive artistic mind map generator with artificial imagination. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6545–6547. ijcai.org.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3219–3232. Association for Computational Linguistics.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ahmed Masry, Parsa Kavehzadeh, Do Xuan Long, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14662–14684. Association for Computational Linguistics.

Ahmed Masry, Mehrad Shahmohammadi, Md. Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Chartinstruct: Instruction tuning for chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 10387–10409. Association for Computational Linguistics.

Igor Melnyk, Pierre L. Dognin, and Payel Das. 2022. Knowledge graph generation from text. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1610–1622. Association for Computational Linguistics.

Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F. Enguix, and Kusum Lata. 2023. Text2kgbench: A

benchmark for ontology-driven knowledge graph generation from text. In *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part II*, volume 14266 of *Lecture Notes in Computer Science*, pages 247–265. Springer.

Antonio Miranda-Escalada, Farrokh Mehryary, Jouni Luoma, Darryl Estrada-Zavala, Luis Gascó, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2023. Overview of drugprot task at biocreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical-protein relations. *Database J. Biol. Databases Curation*, 2023.

Belinda Mo, Kyssen Yu, Joshua Kazdan, Proud Mpala, Lisa Yu, Chris Cundy, Charilaos Kanatsoulis, and Sanmi Koyejo. 2025. Kggen: Extracting knowledge graphs from plain text with language models. *Preprint*, arXiv:2502.09956.

Sabah Mohammed and Jinan Fiaidhi. 2023. Establishment of a mindmap for medical e-diagnosis as a service for graph-based learning and analytics. *Neural Comput. Appl.*, 35(22):16089–16100.

Rizki Elisa Nalawati and Azka Dini Yuntari. 2021. Ratcliff/obershelp algorithm as an automatic assessment on e-learning. In *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, pages 244–248.

Linyong Nan, Dragomir R. Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 432–447. Association for Computational Linguistics.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S. Weld, Joseph Chee Chang, and Kyle Lo. 2024. Arxivdigestables: Synthesizing scientific literature into tables using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9612–9631. Association for Computational Linguistics.

David D. Nguyen, David Liebowitz, Surya Nepal, Salil S. Kanhere, and Sharif Abuadbba. 2024. Contextual chart generation for cyber deception. *CoRR*, abs/2404.04854.

Feng Nie, Jinpeng Wang, Jin-Ge Yao, Rong Pan, and Chin-Yew Lin. 2018. Operation-guided neural networks for high fidelity data-to-text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3879–3889. Association for Computational Linguistics.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 201–206. Association for Computational Linguistics.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 138–147. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025. Introducing openai o3 and o4-mini.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1173–1186. Association for Computational Linguistics.

Michal Pietruszka, Michal Turski, Lukasz Borchmann, Tomasz Dwojak, Gabriela Nowakowska, Karolina Szyndler, Dawid Jurkiewicz, and Lukasz Garncarek. 2024. Stable: Table generation framework for encoder-decoder models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2454–2472. Association for Computational Linguistics.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence,*

Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 2023–2035. Association for Computational Linguistics.

Qwen. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Pritika Ramu, Aparna Garimella, and Sambaran Bandyopadhyay. 2024. Is this a bad table? A closer look at the evaluation of table generation from text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 22206–22216. Association for Computational Linguistics.

Md. Mahinur Rashid, Hasin Kawsar Jahan, Annysha Huzzat, Riyasaat Ahmed Rahul, Tamim Bin Zakir, Farhana Meem, Md. Saddam Hossain Mukta, and Swakkhar Shatabda. 2022. Text2chart: A multi-staged chart generator from natural language text. In *Advances in Knowledge Discovery and Data Mining - 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16-19, 2022, Proceedings, Part II*, volume 13281 of *Lecture Notes in Computer Science*, pages 3–16. Springer.

Emily Reif, Crystal Qian, James Wexler, and Minsuk Kahng. 2024. Automatic histograms: Leveraging language models for text dataset exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 53:1–53:9. ACM.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Haoxiang Shi, Jiaan Wang, Jiarong Xu, Cen Wang, and Tetsuya Sakai. 2024. Ct-eval: Benchmarking chinese text-to-table performance in large language models. *CoRR*, abs/2405.12174.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic RAG. *CoRR*, abs/2501.09136.

Anirudh Sundar, Christopher Richardson, William Gay, and Larry Heck. 2024a. itbls: A dataset of interactive conversations over tabular information. *CoRR*, abs/2404.12580.

Anirudh Sundar, Christopher Richardson, and Larry Heck. 2024b. gtbls: Generating tables from text by conditional question answering. *CoRR*, abs/2403.14457.

Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark Gerstein. 2024. Struc-bench: Are large language models good at generating complex structured tabular data? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 12–34. Association for Computational Linguistics.

Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. 2023. Chartgpt: Leveraging llms to generate charts from abstract natural language. *CoRR*, abs/2311.01920.

Jiaqi Wang, Yuying Chang, Zhong Li, Ning An, Qi Ma, Lei Hei, Haibo Luo, Yifei Lu, and Feiliang Ren. 2024a. Techgpt-2.0: A large language model project to solve the task of knowledge graph construction. *CoRR*, abs/2401.04507.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023a. Instructuie: Multi-task instruction tuning for unified information extraction. *CoRR*, abs/2304.08085.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8640–8665. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith,

Iz Beltagy, and Hannaneh Hajishirzi. 2023c. How far can camels go? exploring the state of instruction tuning on open resources. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yuhao Wang, Ruiyang Ren, Junyi Li, Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024b. REAR: A relevance-aware retrieval-augmented framework for open-domain question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5613–5626. Association for Computational Linguistics.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024c. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yang Wei, Honglei Guo, Jin-Mao Wei, and Zhong Su. 2019. Revealing semantic structures of texts: Multi-grained framework for automatic mind-map generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5247–5254. ijcai.org.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2253–2263. Association for Computational Linguistics.

Xueqing Wu, Jiacheng Zhang, and Hang Li. 2022. Text-to-table: A new way of information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2518–2533. Association for Computational Linguistics.

xAI. 2025. Grok 3 beta âĂŤ the age of reasoning agents.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *CoRR*, abs/2402.12185.

Shishi Xiao, Suizi Huang, Yue Lin, Yilin Ye, and Wei Zeng. 2024. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. *IEEE Trans. Vis. Comput. Graph.*, 30(1):284–294.

Ruobing Xie, Zhiyuan Liu, Fen Lin, and Leyu Lin. 2018. Does william shakespeare REALLY write hamlet? knowledge representation learning with confidence. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4954–4961. AAAI Press.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: a survey. *Frontiers Comput. Sci.*, 18(6):186357.

Wenyi Xu, Yuren Mao, Xiaolu Zhang, Chao Zhang, Xuemei Dong, Mengfei Zhang, and Yunjun Gao. 2025. Dagent: A relational database-driven data analysis report generation agent. *CoRR*, abs/2503.13269.

Rui Yang, Boming Yang, Aosong Feng, Sixun Ouyang, Moritz Blum, Tianwei She, Yuang Jiang, Freddy Lécué, Jinghui Lu, and Irene Li. 2024. Graphusion: A RAG framework for knowledge graph construction with a global perspective. *CoRR*, abs/2410.17600.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777. Association for Computational Linguistics.

Simge Yildirim, Yunus Santur, and Murat AydoÄ§an. 2024. Nlp in fintech: Developing a lightweight text-to-chart application for financial analysis. *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–6.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Fatemeh Pesaran Zadeh, Juyeon Kim, Jin-Hwa Kim, and Gunhee Kim. 2024. Text2chart31: Instruction tuning for chart generation with automatic feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*

2024, Miami, FL, USA, November 12-16, 2024, pages 11459–11480. Association for Computational Linguistics.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024a. Tinychart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1882–1898. Association for Computational Linguistics.

Ruichen Zhang, Shunpu Tang, Yinqiu Liu, Dusit Niyato, Zehui Xiong, Sumei Sun, Shiwen Mao, and Zhu Han. 2025. Toward agentic AI: generative information retrieval inspired intelligent communications and networking. *CoRR*, abs/2502.16866.

Songheng Zhang, Lei Wang, Toby Jia-Jun Li, Qiaomu Shen, Yixin Cao, and Yong Wang. 2024b. Chartifytext: Automated chart generation from data-involved texts via LLM. *CoRR*, abs/2410.14331.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Weinan Zhang, Junwei Liao, Ning Li, and Kounianhua Du. 2024c. Agentic information retrieval. *CoRR*, abs/2410.09713.

Yunjia Zhang, Jordan Henkel, Avrilia Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M. Patel. 2024d. Reactable: Enhancing react for table question answering. *Proc. VLDB Endow.*, 17(8):1981–1994.

Zhuowei Zhang, Mengting Hu, Yinhao Bai, and Zhen Zhang. 2024e. Coreference graph guidance for mindmap generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19623–19631. AAAI Press.

Feng Zhao, Hongzhi Zou, and Cheng Yan. 2023. Structure-aware knowledge graph-to-text generation with planning selection and similarity distinction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8693–8703. Association for Computational Linguistics.

Jiawen Zhu, Jinye Ran, Roy Ka-Wei Lee, Zhi Li, and Kenny T. W. Choo. 2021. Autochart: A dataset for chart-to-text generation task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3 September, 2021*, pages 1636–1644. INCOMA Ltd.

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. Efficientrag: Efficient retriever for multi-hop question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3392–3411. Association for Computational Linguistics.

Appendices

A Task Definition for Graph Generation

This provides an extended definition of graph generation, where graphs can be further categorized into knowledge graphs and mind maps.

Knowledge Graph A knowledge graph is a directed graph $\boldsymbol{G}_{\text{kg}} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{L})$, where: $\mathcal{V}$ represents entities, $\mathcal{E}$ represents direct relationships, $\mathcal{A}(v)$ and $\mathcal{A}(e)$ are the attributes of entity types and relation types, $\mathcal{L}(v)$ and $\mathcal{L}(e)$ provide contextual meaning for nodes and edges.

Mind Map A mind map is a hierarchical directed graph $\boldsymbol{G}_{\text{mm}} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{L})$, where: $\mathcal{V}$ represents concepts with a unique central node, $\mathcal{E}$ represents directed governing relations, $\mathcal{A}(v)$ distinguish root and subordinate nodes, $\mathcal{L}(v)$ and $\mathcal{L}(e)$ provide contextual meaning for concepts and relationships. There is a hierarchy constraint that each non-root node has a unique path to the root (Wei et al., 2019; Hu et al., 2021).

B Detail of Current Datasets

B.1 Dataset Categorization

We use T2S to differentiate native text-to-structure datasets from those repurposed from structure-to-text tasks, as this fundamental design distinction impacts model performance evaluation. Regarding annotation practices, we consider any human verification or correction, even in crawled datasets, as constituting annotation. Schema constraints distinguish between limited schemas (domain-specific, fixed attributes like sports terminology) and unlimited schemas (open-domain, free-form relations, such as Wikipedia tables). Text inputs are classified by length into sentence, paragraph, and document scales. Reasoning complexity indicates the cognitive effort required to transform textual information into structured formats: (1) Low difficulty involves direct extraction of explicit entities/relations, populating a table with stated attributes, or creating a simple bar chart from numeric mentions; (2) Medium difficulty involves context-bound operation, such as resolving coreferences, linking "it" to a named entity in a knowledge graph, or aggregating values for chart axes; (3) High difficulty involves global reasoning, for example, temporal reasoning (e.g., constructing timelines from events), or inferring implicit hierarchies (e.g., deducing parent-child relationships in taxonomies for graph structures). The structural complexity can be systematically defined across multiple axes: (1) Tables: low complexity means less than two rows or columns; medium complexity means larger grids with uniform cell structures; high complexity means complex layouts (e.g., merged cells, nested tables, hierarchical headers); (2) Graphs: the complexity are graded by average node degree and total node count in ascending order; (3) Charts: the complexity are categorized by the diversity of chart types (e.g., bar, line, scatter) in the dataset. The Gold Test Set (GTS) serves as a critical quality indicator, distinguishing between fully human-annotated test sets and those with alternative provenance.

B.2 Common Practices for Data Construction

LLM Paraphrasing During the data generation process, leveraging LLMs for paraphrasing ensures alignment with instructional tone, enhances sentence diversity, and improves the overall data quality (Jiao et al., 2023; Deng et al., 2024).

Quality Control Various methods have been employed to ensure the quality of custom-built datasets. Some studies randomly select instances for annotation to analyze results (Gui et al., 2023; Tang et al., 2024; Deng et al., 2024). Others design iterative prompts to ensure data quality (Li et al., 2024b). Moreover, a combined approach leveraging LLMs and human collaboration has been applied for data cleaning (Shi et al., 2024).

Categorization by Difficulty Level It is a common practice that the proposed datasets are pre-divided based on subtask difficulty levels (Jiao et al., 2023; Deng et al., 2024). It allows for a more granular assessment of the model's strengths and weaknesses across different complexity levels.

C Details of T2S Benchmark

C.1 Details of Evaluation Criteria

As mentioned in Section 6.1, we propose an evaluation framework consisting of content faithfulness (precision and recall) and structure coherence (alignment and clarity). To integrate these metrics into the context of agentic AI systems, we assess their application across four key aspects of text-to-structure generation, which is a critical process for enhancing agents' autonomy and effectiveness. Figure 5 shows the prompting template for the evaluation of structure outputs. For the complete version of the evaluation criteria regarding tables, graphs,

Prompting Template

```
You are an expert in {Structure} evaluation. Your task
is to {Task Description}. Assign four scores from 0
to 100 based on the following criteria. Please keep
the scoring thought process as CONCISE as possible,
without repeatedly adjusting the scores.

# Evaluation Criteria
{Evaluation Criteria}

Please compare the following {Structure} and provide
me with the final four section scores. Output in this
format:

    1. **Brief Justification**
    2. **Final Score**
        Section 1: XX/100
        Section 2: XX/100
        Section 3: XX/100
        Section 4: XX/100

# **Ground Truth {Structure}**:
{Ground Truth}

# **LLM-Generated {Structure}**:
{LLM Generated Output}

# **Source Text **:
{Source Text}
```

Figure 5: Prompting template for structure outputs evaluation.

and charts in the prompting template, please refer to Figure 6, 7, 8, respectively. Below, we outline how each metric applies to this process:

Precision Precision evaluates the structured outputs by identifying conflicts with the ground truth and source text, such as factual errors or mismatches. In applications like automated report generation or document extraction, precision ensures that the AI-generated tables and charts do not introduce mistakes. The detailed calculation rules for precision are outlined in the prompt – it is defined as the percentage of incorrect information relative to the total generated information, multiplied by 100 to obtain the precision score. When comparing whether two pieces of information are equivalent, numerical values, textual content, and other data types are evaluated separately.

Recall Recall evaluates the completeness of the structured outputs, assessing how thoroughly the AI captures details from the source text. It is crucial for ensuring that no critical information is omitted during the text-to-structure conversion. The detailed calculation rules for recall are detailed in the prompt – it is determined by the percentage of information from the ground truth that appears in the output, multiplied by 100 to obtain the recall score.

Alignment Alignment evaluates how well the structured outputs conform to expected formats, such as consistent schemas or data types, ensuring seamless processing by the AI or downstream applications. The detailed calculation rules for alignment are outlined in the prompt – it is generally assessed by categorizing the degree of alignment between the output and the instruction across expected formats, data types, and other relevant dimensions into four scoring tiers.

Clarity Clarity evaluates the logical organization and presentation of the structured outputs, ensuring they are unambiguous and free of redundancy. This will enhance the efficiency of data interpretation and avoid confusion to maintain usability. The detailed calculation rules for clarity are outlined in the prompt – the outputs will be evaluated and scored across four tiers based on degree of ambiguity, comprehensibility, and redundancy.
Taking applying to agentic workflows as an example, this scoring framework optimizes the performance of text-to-structure generation: *precision* and *recall* validate extracted data fidelity, ensuring agents develop an accurate and comprehensive understanding of data, thereby facilitating the generation of reliable and holistic decisions; while *alignment* and *clarity* govern integration with downstream modules, enhancing interoperability within agentic AI systems and optimizing the efficiency of data interpretation and action.

C.2 Selection of Datasets

Regarding the selection of datasets for our Text-to-Structure benchmark, we first ensure that the chosen datasets are open-source and that their test sets are manually annotated to guarantee accuracy. For the text-to-table generation task, we select three datasets, each with distinct characteristics:

- **InstructIE** (Jiao et al., 2023): Each (text, table) pair is accompanied by a complex instruction, with a unique instruction for different instances.
- **Struc-Bench Table** (Tang et al., 2024): A cleaned version of RotoWire dataset (Wiseman et al., 2017), serving as a traditional benchmark since the beginning of text-to-table tasks.
- **LiveSum** (Deng et al., 2024): Focuses on evaluating a model's ability to integrate information from text into tables, a capability currently lacking in many LLMs.

Evaluation Criteria for Tables

1. **Precision (100 points)**
   - Evaluates whether generated cells (with its corresponding row/column headers) conflict with ground truth. If no corresponding information exists in the ground truth, compare with the original text. Incorrect information is deducted proportionally from the total score.
     - Error Rate = (Number of conflicting cells) / (Total cells)
     - Final Score = 100 * (1 - Error Rate)
   - - **Correctness Guidelines**
     - Numerical values: Strict equality required.
     - Textual content: Semantic equivalence required.
     - Empty cells: Equivalent to "not specified", "N/A", or "-".
2. **Recall (100 points)**
   - For each cell in the ground truth (with its corresponding row/column headers), evaluate the percentage of information captured in the LLM-generated output. Missing or incorrect information is deducted proportionally from the total score.
     - Rate = (Number of missing or incorrect cells) / (Total cells)
     - Final Score = 100 * (1 - Rate)
   - For **Correctness Guidelines**, please follow the same criteria as Precision.
3. **Alignment (100 points)**
   - Evaluates alignment with expected formats, schema relations, and data types (transposed rows/columns allowed without penalty):
     - Perfect Match (90-100): Rows/columns fully match reference structure.
     - Minor Deviations (80-89): Few row/column discrepancies; requires minimal adjustments.
     - Noticeable Differences (60-79): Multiple added/removed rows/columns; needs manual reorganization.
     - Severe Deviation (0-59): Missing headers/data type errors; structure unusable.
4. **Clarity (100 points)**
   - Assesses overall intelligibility of the table and clarity and redundancy in headers/cell values:
     - Perfect Clarity (90-100): No ambiguity or redundancy.
     - Minor Issues (80-89): Few redundant/ambiguous terms; core meaning intact.
     - Significant Issues (60-79): Multiple ambiguous terms requiring inference; extractable core data.
     - Critical Flaws (0-59): Conflicting headers/unreadable data; table unreliable.

Figure 6: Evaluation Criteria for Tables

Evaluation Criteria for Knowledge Graphs

1. **Precision (100 points)**
   - Evaluates whether generated triples conflict with ground truth. If no corresponding information exists in the ground truth, compare with the original text. Incorrect information is deducted proportionally from the total score.
     - Error Rate = (Number of conflicting triples) / (Total triples)
     - Final Score = 100 * (1 - Error Rate)
   - - **Correctness Guidelines**
     - Numerical values: Strict equality required.
     - Textual content: Semantic equivalence required.
2. **Recall (100 points)**
   - For each triple in the ground truth, evaluate the percentage of information captured in the LLM-generated triples. Missing or incorrect information is deducted proportionally from the total score.
     - Rate = (Number of missing or incorrect triples) / (Total triples)
     - Final Score = 100 * (1 - Rate)
   - For **Correctness Guidelines**, please follow the same criteria as Precision.
3. **Alignment (100 points)**
   - Evaluates alignment with expected formats, schema relations, and data types (transposed rows/columns allowed without penalty):
     - Perfect Match (90-100): Exact schema/format match.
     - Minor Deviations (80-89): extra/missing triples; parsable.
     - Noticeable Differences (60-79): Multiple missing triples/schema gaps; needs manual fixes.
     - Severe Deviation (0-59): Invalid predicates/scrambled data; KG unusable.
4. **Clarity (100 points)**
   - Assesses clarity, standardization, and absence of ambiguity/redundancy in entities and relations:
     - Perfect Clarity (90-100): Standardized terms; zero redundancy.
     - Minor Issues (80-89): Rare ambiguous labels; core clear.
     - Significant Issues (60-79): Frequent ambiguous terms; extractable core.
     - Critical Flaws (0-59): Uninterpretable relations; KG unreliable

Figure 7: Evaluation Criteria for Knowledge Graphs

For the text-to-graph generation task, the available datasets are limited. We select two open-domain datasets: **DART** (Nan et al., 2021) and **InstructtIE** (Gui et al., 2023), where the latter includes half of its data in Chinese, enabling evaluation of multilingual performance. In the text-to-chart

Evaluation Criteria for Charts

1. **Precision (100 points)**
   - Evaluates whether generated chart elements (data points, labels, categories) conflict with the ground truth. If no ground truth chart exists, compare with the original text.
     - Error Rate = (Number of conflicting data points/labels/categories) / (Total data points/labels/categories)
     - Final Score = 100 * (1 - Error Rate)
   - - **Correctness Guidelines**
     - Numerical values (e.g., axis values, percentages): Strict equality required (up to two decimal places allowed).
     - Categorical labels (e.g., axis labels, legend entries): Semantic equivalence required (e.g., "Q1 2023" vs. "First Quarter" is acceptable; "Male" vs. "Female" is not).
     - Missing elements: Treat as errors if present in the ground truth but omitted in the generated chart.
2. **Recall (100 points)**
   - Measures how well the generated chart captures all data points, trends, and categories from the ground truth.
     - Rate = (Number of missing or misrepresented data points/categories/labels) / (Total data points/categories/labels)
     - Final Score = 100 * (1 - Rate)
   - For **Correctness Guidelines**, please follow the same criteria as Precision, prioritizing critical elements (e.g., peaks, outliers, primary categories).
3. **Alignment (100 points)**
   - Alignment with expected chart structure:
     - Perfect (90-100): Chart type, headers/labels, and groupings fully match reference.
     - Minor flaws (80-89): Formatting/positioning deviations (e.g., axis units, legend placement) with minimal impact.
     - Major errors (60-79): Incorrect chart type, missing key labels, or misaligned data groups requiring manual fixes.
     - Invalid (0-59): Broken structure (e.g., inverted axes, data-trend conflicts).
4. **Clarity (100 points)**
   - Clarity and truthfulness:
     - Perfect (90-100): Visual encodings (color/size/position) logically match data; no ambiguity.
     - Minor issues (80-89): Non-critical redundancy/design flaws (e.g., label overlaps) with preserved meaning.
     - Confusing (60-79): Ambiguous scales/axes requiring user guesses.
     - Misleading (0-59): Distorted representations (e.g., truncated axes, false color mappings) altering insights.

Figure 8: Evaluation Criteria for Charts

generation task, due to the scarcity of dedicated datasets, we choose **Text2Chart31** (Zadeh et al., 2024), the most comprehensive dataset in terms of chart type coverage, which effectively tests the generalization capability of models. To standardize the datasets, we fix the training set sample size to 1,000 instances during SFT and sample 250 instances from each dataset's test set for evaluation.

C.3 Selection of Models

The selected LLMs in our study can be categorized into two types: non-thinking models and thinking models. Specifically, the non-thinking models include: Mistral-7B (Jiang et al., 2023a), Llama-3.1-8B (Dubey et al., 2024), R1-Distill-Llama-8B (DeepSeek-AI et al., 2025), Phi-3-medium (Abdin et al., 2024), Mixtral-8x22B (Jiang et al., 2024a), Llama-3.1-405B (Dubey et al., 2024), Llama-3.3-70B (Dubey et al., 2024), GPT-3.5-turbo (OpenAI, 2022) and GPT-4o (OpenAI, 2024); while the thinking models comprise QwQ-32B (Qwen, 2025), DeepSeek-R1 (DeepSeek-AI et al., 2025), Grok-3-mini (xAI, 2025), and O4-mini (OpenAI, 2025). For SFT experiments, we evaluate models Mistral-7B, Llama-3.1-8B, R1-Distill-Llama-8B. The remaining models are assessed under a zero-shot setting. Additionally, for non-thinking models, we further examine their performance when augmented with CoT prompting. For the SFT models, our experiments are conducted on a single H20 GPU card. For other LLMs, we access their capabilities through API calls. Specifically, we conduct model inference with temperature set to 0 for non-thinking models, while adopting the officially recommended hyperparameters for thinking models as specified in their respective documentation, with all results averaged over five independent runs. When employing DeepSeek-R1 and GPT-4o as LLM-as-judge, we fix the temperature parameter at 0 to ensure that every evaluation remains deterministic.

D Result Analysis

D.1 Text-to-Table Generation

As shown in Table 7, on the InstructIE dataset, the thinking models demonstrate strong performance, with Grok-3-mini achieving the highest precision of 81.82 and R1 achieving the highest recall of 78.43; however, CoT does not produce consistent improvements across all models. Regarding coherence, while the thinking models maintain excellent

Model	InstructIE (Jiao et al., 2023)						Struc-Bench (Tang et al., 2024)						LiveSum (Deng et al., 2024)					
	Prec.	Rec.	FAITH.	Align.	Clr.	COH.	Prec.	Rec.	FAITH.	Align.	Clr.	COH.	Prec.	Rec.	FAITH.	Align.	Clr.	COH.
♦ *SFT*																		
Mistral-7B	67.84	55.93	**58.08**	67.02	74.60	**68.85**	91.14	86.73	**87.65**	95.47	87.62	**90.05**	53.11	53.87	**53.31**	99.81	97.78	**98.54**
Llama-3.1-8B	64.38	60.34	**59.31**	64.61	72.28	**66.44**	87.02	83.94	**85.01**	90.38	82.67	**85.95**	61.69	61.46	**61.34**	90.56	90.65	**90.30**
R1-Distill-Llama-8B	70.39	61.44	**62.86**	64.69	75.12	**68.08**	63.97	54.84	**54.56**	62.55	57.34	**58.97**	52.93	43.21	**43.27**	86.28	86.44	**86.22**
♦ *Zero-Shot*																		
Phi-3-medium	43.19	28.79	**29.15**	40.42	42.75	**39.89**	74.68	75.36	**74.04**	91.79	78.68	**83.99**	41.84	42.27	**41.61**	93.64	90.62	**91.73**
Phi-3-medium (CoT)	42.85	29.82	**30.36**	37.83	39.56	**37.23**	74.91	73.83	**73.42**	88.81	75.92	**81.06**	46.57	46.59	**46.36**	94.40	92.54	**93.19**
Mistral-7B	58.53	53.29	**53.00**	57.05	65.53	**59.73**	55.17	61.17	**56.19**	72.52	64.05	**67.05**	31.25	30.56	**30.14**	90.48	85.78	**87.26**
Mistral-7B (CoT)	59.60	53.82	**54.02**	57.30	66.83	**60.37**	54.35	62.38	**56.25**	76.51	66.44	**70.34**	42.38	42.14	**41.27**	89.75	83.43	**85.97**
Mixtral-8x22B	66.92	54.34	**56.35**	66.39	73.42	**68.25**	84.95	84.71	**84.16**	89.00	82.74	**85.20**	49.74	48.97	**49.12**	87.54	88.91	**87.91**
Mixtral-8x22B (CoT)	66.82	56.05	**58.27**	64.97	74.73	**68.29**	82.81	80.26	**80.55**	74.98	75.91	**74.94**	55.68	54.43	**54.87**	86.75	88.51	**87.37**
Llama-3.1-405B	71.31	62.27	**63.88**	65.36	77.31	**69.63**	90.20	89.32	**89.43**	91.57	90.49	**90.43**	62.46	62.38	**62.29**	99.55	98.94	**99.11**
Llama-3.1-405B (CoT)	72.43	62.39	**64.54**	69.43	78.69	**72.80**	90.69	88.69	**89.38**	91.72	90.51	**90.74**	66.27	66.33	**66.25**	98.87	98.51	**98.64**
Llama-3.3-70B	71.10	60.09	**62.88**	64.63	76.50	**69.19**	82.00	80.98	**80.58**	84.68	77.58	**80.00**	58.55	58.63	**58.48**	99.40	99.13	**99.19**
Llama-3.3-70B (CoT)	71.71	63.35	**65.34**	64.79	77.99	**69.46**	87.53	86.47	**86.47**	87.16	85.71	**85.84**	62.66	63.00	**62.79**	98.98	98.47	**98.68**
GPT-3.5-turbo	65.29	53.80	**55.09**	67.05	74.99	**69.01**	79.60	79.61	**78.44**	92.09	81.98	**85.37**	46.08	45.39	**45.38**	77.44	82.92	**79.18**
GPT-3.5-turbo (CoT)	68.95	55.61	**58.44**	65.69	75.21	**68.80**	72.86	75.03	**71.88**	89.55	76.30	**81.54**	45.14	45.57	**45.09**	85.88	87.71	**86.37**
GPT-4o	75.17	71.67	**70.72**	72.97	82.35	**76.36**	88.00	87.44	**87.35**	96.34	89.22	**92.22**	63.46	63.09	**63.21**	91.46	92.96	**91.99**
GPT-4o (CoT)	75.91	67.93	**68.11**	71.97	80.15	**74.40**	87.37	85.46	**86.05**	94.29	87.24	**90.03**	61.96	62.88	**62.16**	98.76	98.38	**98.50**
QwQ-32B	75.53	66.66	**68.23**	63.08	77.90	**68.79**	81.15	81.29	**81.21**	90.97	84.03	**87.12**	58.72	58.47	**58.06**	95.94	96.95	**96.20**
DeepSeek-R1	79.05	78.72	**76.86**	70.04	82.93	**74.75**	88.19	87.95	**87.77**	94.79	89.32	**91.05**	64.37	64.79	**64.89**	96.74	97.88	**96.85**
Grok-3-mini	81.34	72.42	**75.57**	68.93	82.91	**75.40**	86.17	85.45	**85.60**	94.61	89.59	**92.28**	65.81	65.83	**66.35**	96.45	94.27	**95.26**
O4-mini	80.86	73.26	**75.71**	68.73	83.57	**74.85**	84.13	84.83	**83.45**	94.52	86.30	**90.02**	84.20	84.43	**84.48**	98.64	98.65	**99.16**

Table 7: Performance of different models on three text-to-table datasets, with the highest scores in each metric underlined.

Model	DART (Nan et al., 2021)						InstructIE (Gui et al., 2023)					
	Prec.	Rec.	FAITH.	Align.	Clr.	COH.	Prec.	Rec.	FAITH.	Align.	Clr.	COH.
♦ *SFT*												
Mistral-7B	80.99	70.81	**75.07**	83.15	83.54	**83.17**	59.75	45.35	**44.58**	72.21	69.31	**69.67**
Llama-3.1-8B	63.63	35.92	**38.94**	55.56	56.75	**55.13**	55.37	35.73	**34.93**	66.42	65.70	**64.99**
R1-Distill-Llama-8B	68.98	54.83	**59.03**	68.41	71.40	**69.43**	71.78	33.35	**37.27**	70.89	69.91	**69.73**
♦ *Zero-Shot*												
Phi-3-medium	73.80	60.80	**64.60**	72.36	73.26	**72.59**	74.64	39.84	**43.40**	74.08	71.68	**71.44**
Phi-3-medium (CoT)	73.69	61.10	**64.88**	71.77	73.52	**72.32**	74.54	39.10	**41.28**	74.39	70.40	**70.44**
Mistral-7B	74.22	69.81	**70.27**	73.34	73.44	**73.06**	65.88	41.21	**42.74**	69.34	66.85	**66.54**
Mistral-7B (CoT)	76.98	59.92	**64.68**	70.70	71.96	**70.96**	66.78	30.86	**33.35**	66.36	67.40	**66.01**
Mixtral-8x22B	70.58	65.30	**65.70**	71.72	71.07	**71.07**	68.69	38.65	**40.07**	71.32	68.57	**68.89**
Mixtral-8x22B (CoT)	78.14	69.19	**71.27**	74.93	74.43	**74.53**	75.68	43.30	**47.47**	72.92	71.31	**71.29**
Llama-3.1-405B	75.77	65.24	**68.74**	75.59	76.11	**75.61**	73.79	43.30	**46.72**	78.16	75.52	**75.82**
Llama-3.1-405B (CoT)	83.56	66.49	**72.31**	77.99	78.57	**78.09**	79.32	43.46	**47.94**	79.90	78.10	**78.48**
Llama-3.3-70B	79.22	66.95	**71.09**	78.34	77.21	**77.54**	77.23	44.64	**47.11**	78.70	76.36	**76.30**
Llama-3.3-70B (CoT)	85.74	71.69	**77.11**	79.02	79.24	**78.99**	76.47	41.80	**46.56**	80.23	78.08	**78.51**
GPT-3.5-turbo	80.95	64.02	**69.75**	76.56	77.27	**76.72**	74.63	38.16	**41.89**	75.66	73.32	**73.34**
GPT-3.5-turbo (CoT)	79.76	65.20	**70.17**	76.20	76.76	**75.99**	74.78	37.01	**41.78**	75.08	73.40	**73.51**
GPT-4o	80.51	71.42	**74.15**	77.51	77.33	**77.18**	76.02	43.13	**46.77**	78.81	76.85	**77.00**
GPT-4o (CoT)	87.02	76.47	**79.89**	79.44	79.45	**79.33**	81.21	49.30	**53.28**	79.38	77.72	**77.65**
QwQ-32B	76.96	65.79	**69.42**	78.64	78.83	**78.02**	86.42	50.02	**56.30**	80.67	82.82	**82.15**
DeepSeek-R1	84.67	74.82	**78.25**	80.04	79.42	**79.89**	80.93	47.77	**52.15**	80.27	80.20	**79.23**
Grok-3-mini	82.41	68.94	**73.32**	79.11	80.38	**80.34**	87.66	50.06	**56.03**	81.18	83.90	**82.68**
O4-mini	84.03	71.06	**74.82**	79.28	80.52	**80.19**	89.89	49.91	**56.46**	81.51	83.33	**82.32**

Table 8: Performance of different models on two text-to-KG datasets, with the highest scores in each metric underlined.

performance, GPT-4o achieves the highest score. Among SFT models, the R1-distilled Llama-8B exhibits the best overall performance, although it still falls short of the thinking models' capabilities. On the Struc-Bench dataset, most LLMs show strong baseline performance since the tasks primarily involve information replication rather than complex reasoning. In terms of specific metrics, Mistral-7B (SFT) achieves the highest precision score, while Llama-3.1-405B achieves the highest recall score. For coherence evaluation, GPT-4o exhibits superior capabilities. Notably, the Mistral-7B (SFT) model produces great performance gains, with improvements of 56% in faithfulness and 34% in coher-

ence metrics compared to zero-shot approaches, empirically demonstrating that SFT can significantly enhance model performance on information-replication text-to-table tasks. On the Livesum dataset, both precision and recall scores are generally low due to the complex information integration and reasoning operations required by the task. O4-mini achieves exceptional performance with a faithfulness score of 83.99. Regarding coherence, all models attain relatively high scores due to the dataset's standardized table format, with Llama-3.3-70B demonstrating the best performance. The results indicate that thinking models maintain a competitive advantage on this dataset, while among SFT models, Llama-3.1-8B achieves comparable performance to larger LLMs.

D.2 Text-to-Graph Generation

As shown in Table 8, our experimental results demonstrate that GPT-4o (CoT) achieves the highest faithfulness scores on the DART dataset, while Mistral-7B (SFT) shows superior coherence performance that significantly outperforms all other LLMs. The application of CoT prompting generally improves precision and recall metrics by 6-10% across most LLMs, though it does not yield significant improvement in coherence metrics. On the InstructIE dataset, inherent dataset characteristics lead to lower recall scores overall, but QwQ-32B maintains relatively better performance in this metric, with O4-mini achieving the highest precision score of 89.45. The coherence metrics show remarkable consistency with the patterns observed on the DART dataset, suggesting stable model behavior across different evaluation benchmarks.

D.3 Text-to-Chart Generation

As shown in Table 9, in addition to the aforementioned metrics, we also evaluate the success rate of the generated Python code executing correctly to produce the desired images, denoted as SUCC. It is observed that the majority of LLMs achieve a success rate of over 95%. Among them, Grok-3-mini demonstrates the highest success rate, reaching 99.2%. In terms of precision an recall, most LLMs achieve scores above 80, with Llama-3.3-70B (CoT) performing the highest, scoring 88.18 and 90.16, respectively. For alignment and clarity, Grok-3-mini exhibits the best performance, achieving scores of 88.34 and 89.65, respectively. Among the SFT models, Llama-3.1-8B demonstrated the highest overall performance. Mistral-7B achieves

Model	Text2Chart31 (Zadeh et al., 2024)						
	SUCC.	Prec.	Rec.	FAITH.	Align.	Clr.	COH.
♦ *SFT*							
Mistral-7B	72.8%	64.04	65.66	**64.54**	63.38	63.70	**63.41**
Llama-3.1-8B	87.2%	76.25	79.16	**77.06**	77.07	77.70	**77.25**
R1-Distill-Llama-8B	47.6%	41.36	42.46	**41.66**	39.97	41.11	**40.45**
♦ *Zero-Shot*							
Phi-3-medium	68.8%	59.62	62.20	**60.24**	60.26	60.89	**60.45**
Phi-3-medium (CoT)	67.6%	56.93	59.05	**57.75**	58.17	58.84	**58.27**
Mistral-7B	64.0%	51.72	53.84	**52.25**	52.03	52.70	**52.03**
Mistral-7B (CoT)	59.2%	44.78	46.38	**45.13**	46.90	47.00	**46.51**
Mixtral-8x22B	72.4%	60.89	63.05	**61.62**	62.09	62.51	**62.15**
Mixtral-8x22B (CoT)	67.6%	58.23	60.49	**58.92**	58.61	59.37	**58.89**
Llama-3.1-405B	98.0%	85.64	88.00	**86.49**	86.47	87.26	**86.62**
Llama-3.1-405B (CoT)	98.4%	86.39	88.71	**87.23**	86.40	87.15	**86.40**
Llama-3.3-70B	98.4%	86.77	89.37	**87.44**	86.54	88.00	**87.12**
Llama-3.3-70B (CoT)	98.8%	88.18	90.16	**88.78**	87.19	88.80	**87.83**
GPT-3.5-turbo	94.4%	84.53	86.23	**85.17**	84.20	85.63	**84.79**
GPT-3.5-turbo (CoT)	93.2%	82.48	84.27	**83.06**	83.63	84.20	**83.62**
GPT-4o	95.6%	85.16	86.35	**85.48**	85.91	86.83	**86.01**
GPT-4o (CoT)	96.0%	85.48	86.82	**85.99**	86.24	87.26	**86.63**
QwQ-32B	98.4%	86.28	88.46	**86.71**	86.61	88.04	**86.94**
DeepSeek-R1	95.6%	85.36	87.16	**85.79**	85.71	87.64	**86.56**
Grok-3-mini	99.2%	87.44	89.19	**88.05**	88.34	89.65	**88.74**
O4-mini	95.6%	84.92	87.30	**85.56**	85.84	87.16	**86.17**

Table 9: Performance of different models on text-to-chart dataset, with the highest scores in each metric underlined.

significant improvements compared to its zero-shot counterpart, with a 13.8% increase in success rate, 23.5% and 21.9% enhancements in faithfulness and coherence scores, respectively, showing the effectiveness of SFT in model optimization.

E Case Study

This section illustrates five representative failure cases of traditional evaluation metrics through systematic analysis. We empirically demonstrate scenarios where traditional metrics (BERTScore, ROUGE, etc.) contradict human judgment while our framework correctly identifies superior outputs. Each case dissects specific error types and metric failures, providing granular evidence for our approach's validity across diverse tasks. To reflect the scoring process of *Content Faithfulness* and *Structure Coherence*, we append the intermediate reasoning steps from the LLM as a brief justification following the output for each case.

Case 1 (Figure 9) It is notable that the table generated by GPT-3.5 lacks two critical rows of information, and its single extracted row contains multiple errors compared to the ground truth. Conversely, the table produced by R1 contains comprehensive information, fully capturing all details present in the ground truth while additionally extracting more complete information. Despite R1âĂŹs table being demonstrably superior in this

context, both BERTScore and ROUGE metrics indicate GPT-3.5âĂŹs output is preferable.

Case 2 (Figure 10) The table generated by GPT-3.5 introduced erroneous information in several cells (e.g., misrepresenting reactants and products). In contrast, R1âĂŹs extracted table maintained complete accuracy in all key information. Paradoxically, BERTScore and ROUGE metrics significantly favor the former.

Case 3 (Figure 11) GPT-3.5âĂŹs table exhibits numerical inaccuracies across multiple cells, with an error rate 60% higher than R1âĂŹs table. Nevertheless, BERTScore metrics consistently rate GPT-3.5âĂŹs output as superior.

Case 4 (Figure 12) GPT-3.5âĂŹs edge information contains instances where predicates and objects are entirely confused, accompanied by structural disarray. Conversely, R1 extracts a structurally correct graph. Contradictorily, both BERTScore and ROUGE metrics indicate GPT-3.5âĂŹs graph extraction performed better.

Case 5 (Figure 13) GPT-3.5âĂŹs generated chart contains numerical errors and fails to reflect logical pairing in its layout. R1âĂŹs chart, however, maintains numerical accuracy and replicates the ground truthâĂŹs logical structure. Despite this qualitative advantage, METEOR and CodeBLEU metrics favor GPT-3.5âĂŹs output.

Case 1: Instance #068 from InstructIE

Comparison of Evaluation Metrics

Output	BERTScore	ROUGE	Content Faithfulness	Structure Coherence
GPT-3.5-turbo	**0.8602**	**0.7052**	40.00	64.62
DeepSeek-R1	0.8364	0.5974	**75.00**	**80.00**

Ground Truth

Number of Bedrooms	Number of Bathrooms	Special Rooms/Spaces	Floor Material	Outdoor Spaces
3	2.5	Breakfast nook, playroom	Hardwood	Fenced backyard, deck, patio
3	N/A	Den, indoor rock garden	New flooring	Rear balconies
5	3.5	Family room, playroom, laundry	Wood	Large back patio, storage shed, fenced yard

GPT-3.5-turbo Output

Number of Bedrooms	Number of Bathrooms	Special Rooms/Spaces	Floor Material	Outdoor Spaces
3	2.5	Playroom, Master Suite, Den	Hardwood floors, New flooring	Deck, Patio, Fenced backyard, Front patio area

```
**Brief Justification**:
  - **Precision**: 3/5 cells conflicting (Special Rooms, Floor Material, Outdoor Spaces) âĘŠ 40.
  - **Recall**: 3/5 ground truth cells missing/incorrect (Breakfast nook, Hardwood, exact Outdoor Spaces) âĘŠ 40.
  - **Alignment**: Missing two rows (ground truth has 3 rows, LLM has 1) âĘŠ Noticeable Differences (70).
  - **Clarity**: Merged data from multiple sources into one row creates ambiguity âĘŠ Significant Issues (60).
```

DeepSeek-R1 Output

#	Bedrooms	Bathrooms	Special Rooms/Spaces	Floor Material	Outdoor Spaces
1	3	2.5	Pantry, kitchen island, breakfast nook, master suite (walk-in closets), playroom, laundry area, garage/workshop	Hardwood	Fenced backyard, deck, patio, front patio
2	3	N/A	Two-story living room, conversation pit, indoor rock garden, den, updated kitchen	New flooring	Rear balconies, golf course views, lake views
3	5	3.5	Kitchen island, pantry, laundry room, primary suite (walk-in closet), finished basement (family room, playroom)	Wood floors	Back patio, storage shed with loft, fenced yard

```
**Brief Justification**
  - **Precision**: LLM-added details (e.g., extra rooms/outdoor features) not in GT create 6/15 conflicts (60).
  - **Recall**: All GT cells are captured despite extra LLM additions (100).
  - **Alignment**: Extra row numbers and header wording deviations (80).
  - **Clarity**: Clear headers but redundant cell details (80).
```

Figure 9: Misalignment of Traditional Metrics and Our Framework: Case 1

Case 2: Instance #127 from InstructIE

Comparison of Evaluation Metrics

Output	BERTScore	ROUGE	Content Faithfulness	Structure Coherence
GPT-3.5-turbo	**0.9229**	**0.9012**	61.46	67.41
DeepSeek-R1	0.7231	0.4086	**100.00**	**76.77**

Ground Truth

Reactants	Solvent	Product	Yield
CyPBn-Cy, $NiCl_2$(DME)	THF	(CyPBn-Cy)$NiCl_2$	68%
(CyPBn-Cy)$NiCl_2$, (o-tol)MgCl	THF	(CyPBn-Cy)Ni(o-tol)Cl	95%

GPT-3.5-turbo Output

Reactants	Products	Solvent	Yield	Analysis Method
CyPBn-Cy, $NiCl_2$(DME), THF	(CyPBn-Cy)$NiCl_2$, dichloromethane	THF	68%	Elemental, NMR, X-ray
(CyPBn-Cy)$NiCl_2$, (o-tol)MgCl, THF	(CyPBn-Cy)Ni(o-tol)Cl	THF	95%	–

```
**Brief Justification**
  - **Precision**: 4/10 cells incorrect (extra "dichloromethane" in Product, Analysis Method column invalid, THF
                   mislabeled as Reactant).
  - **Recall**: 3/8 GT cells missing/incorrect (Reactants misrepresented in both rows, Product error in row 1).
  - **Alignment**: Added Analysis Method column, Products vs Product header mismatch (Noticeable Differences).
  - **Clarity**: Ambiguous Reactants (THF included redundantly), redundant Analysis Method (Significant Issues).
```

DeepSeek-R1 Output

Reaction Equation	Reactants	Solvent	Workup/Isolation	Yield	Analytical Data
eq1	CyPBn-Cy, $NiCl_2$(DME)	THF	Dichloromethane solvate	68%	Elemental analysis, NMR, X-ray crystallography
eq2	(CyPBn-Cy)$NiCl_2$, (o-tol)MgCl	THF	Isolated as pure solid	95%	Elemental analysis

```
- **Brief Justification**
  - **Precision**: All corresponding cells (Reactants, Solvent, Product, Yield) match GT and source text.
                   Extra columns (Workup/Isolation, Analytical Data) align with source and do not conflict.
  - **Recall**: All GT cells are captured in the generated table.
  - **Alignment**: Added columns (Reaction Equation, Workup/Isolation, Analytical Data) deviate from GT schema.
  - **Clarity**: Clear headers and data, but redundant columns introduce minor deviations from GT focus.
```

Figure 10: Misalignment of Traditional Metrics and Our Framework: Case 2

Case 3: Instance #42842344 from LiveSum

Comparison of Evaluation Metrics

Output	BERTScore	ROUGE	Content Faithfulness	Structure Coherence
GPT-3.5-turbo	**0.9303**	0.7750	50.00	100.00
DeepSeek-R1	0.9294	0.7750	**69.00**	100.00

Ground Truth

Team	Goals	Shots	Fouls	Yellow Cards	Red Cards	Corner Kicks	Free Kicks	Offsides
Home Team	5	15	8	2	0	4	8	1
Away Team	1	10	8	0	0	6	8	3

GPT-3.5-turbo Output

	Goals	Shots	Fouls	Yellow Cards	Red Cards	Corner Kicks	Free Kicks	Offsides
Home Team	5	12	6	1	0	5	4	1
Away Team	1	8	7	0	0	4	6	3

```
**Brief Justification**
   - **Precision**: 8/16 cells have errors (numerical mismatches), leading to 50% error rate.
   - **Recall**: Same 8 missing/correct GT cells, 50% recall.
   - **Alignment**: Columns/headers match; row order irrelevant per criteria (no penalty).
   - **Clarity**: Headers clear, data unambiguous despite errors.
```

DeepSeek-R1 Output

Stat	Home Team	Away Team
Goals	5	1
Shots	12	8
Fouls	8	12
Yellow Cards	2	0
Red Cards	0	0
Corner Kicks	5	5
Free Kicks	12	8
Offsides	1	3

```
**Brief Justification**
   - **Precision/Recall**: 5 cell errors (Shots, Fouls, Corner Kicks, Free Kicks).
   - **Alignment**: Transposed correctly without penalties.
   - **Clarity**: Clear headers and layout despite data inaccuracies.
```

Figure 11: Misalignment of Traditional Metrics and Our Framework: Case 3

Case 4: Instance #048 from DART

Comparison of Evaluation Metrics

Output	BERTScore	ROUGE	Content Faithfulness	Structure Coherence
GPT-3.5-turbo	**0.7165**	**0.6465**	40.48	74.67
DeepSeek-R1	0.7052	0.6166	**100.00**	**80.00**

Ground Truth

Subject	Predicate	Object
The Vaults	eatType	pub
The Vaults	food	Japanese
The Vaults	priceRange	Âč20–25
The Vaults	customer rating	high
The Vaults	area	city centre
The Vaults	familyFriendly	yes
The Vaults	near	Rainbow Vegetarian CafÃľ

GPT-3.5-turbo Output

Subject	Predicate	Object
The Vaults	has price range	Âč20–25
The Vaults	is child friendly	
The Vaults	is a Japanese pub	
The Vaults	is near	Rainbow Vegetarian CafÃľ
The Vaults	is located in	city centre
The Vaults	is highly rated	

```
**Brief Justification**
  - **Precision**: 2/6 triples conflict (missing "yes" and "high" values).
  - **Recall**: 5/7 ground truth triples missing/incorrect (combined or incomplete).
  - **Structural**: Missing objects in 2 triples and merged triples, parsable but flawed.
  - **Semantic**: Non-standard predicates (e.g., "is child friendly") but core meaning preserved.
```

DeepSeek-R1 Output

Subject	Relation	Object
The Vaults	is	child friendly
The Vaults	is a	highly rated Japanese pub
The Vaults	has price range	Âč20–25
The Vaults	is located in	city centre
The Vaults	is near	Rainbow Vegetarian CafÃľ

```
**Brief Justification**
  All ground truth information is captured without conflicts, but relations use non-standard phrasing (e.g., "is a"
  vs "eatType"). Structural/semantic scores reflect schema deviations and non-standard predicates.
```

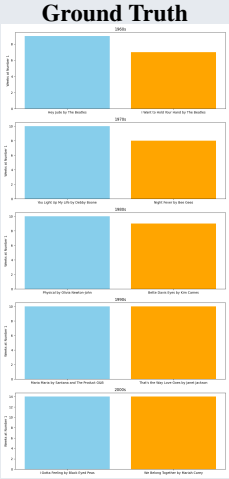
Figure 12: Misalignment of Traditional Metrics and Our Framework: Case 4

Case 5: Instance #0208113110 from Text2Chart31

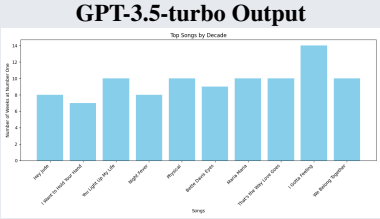
Comparison of Evaluation Metrics

Output	METEOR	CodeBLEU	Content Faithfulness	Structure Coherence
GPT-3.5-turbo	**0.4480**	**0.3701**	74.67	69.64
DeepSeek-R1	0.3148	0.2263	**100.00**	**72.41**

Ground Truth

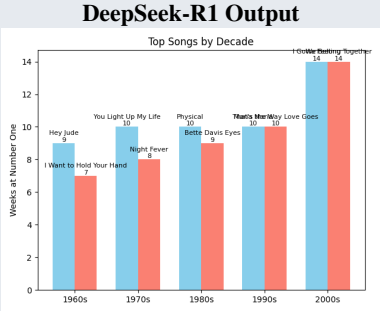


GPT-3.5-turbo Output




```
**Brief Justification**
  - **Precision**: The LLM-generated chart has discrepancies in the number of weeks for some songs compared to
    the ground truth. For example, "Hey Jude" is shown with 8 weeks instead of 9, and "I Gotta Feeling" with
    14 weeks instead of 10.
  - **Recall**: The generated chart includes all songs and their respective weeks, but the values are not always
    accurate, affecting recall.
  - **Alignment**: The chart type is consistent, but the LLM-generated chart lacks the decade grouping present
    in the ground truth, which is a significant structural difference.
  - **Clarity**: The generated chart is clear but lacks the decade separation, which could lead to confusion about
    the time periods of the songs.
```

DeepSeek-R1 Output




```
**Brief Justification**
  - **Precision/Recall**: All correct.
  - **Alignment**: The chart type is consistent (bar chart), but the structure differs. The ground truth uses sepa-
    rate subplots for each decade, while the generated chart combines them into one. This structural difference impacts
    the score.
  - **Clarity**: The generated chart is clear and logical in its presentation, but the overlapping labels and combined
    decades create some confusion, affecting the semantic score.
```

Figure 13: Misalignment of Traditional Metrics and Our Framework: Case 5