Thesis Proposal: Intentional Inference for Insight Generation

Kristýna Onderková
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
onderkova@ufal.mff.cuni.cz

Abstract

Large language models (LLMs) show strong capabilities in natural language generation (NLG) and have been applied to translate complex structured data into human-readable insights. While these models excel at surface-level fluency, they remain unreliable as they produce factually inaccurate outputs and struggle with consistent logical inference beyond surface-level patterns. Moreover, they often lack a clear sense of relevance and produce shallow or uninformative insights.

This proposal argues that a key source of these limitations is task underspecification, which requires models to make implicit assumptions about missing context. We investigate how such underspecification leads to unintentional assumptions and how these affect faithfulness and evaluation. We examine how models can identify missing premises and surface multiple plausible interpretations to make evaluation more rigorous. We also explore how to improve reasoning to enable deeper inferences, focusing on code generation and qualitative reasoning. Finally, we will evaluate how the underlying assumptions and depth of inference influence the perceived interestingness of the insights.

By shifting focus from surface-level generation to assumption-aware deeper inferences, this work aims to improve reliability, interpretability, and user controllability in NLG.

1 Introduction

The advances of Transformer architecture (Vaswani et al., 2017) and large-scale training of LLMs (Minaee et al., 2024) have significantly expanded the capabilities of NLG. LLMs have high potential to translate complex structured data into human-readable reports. Their strong instruction following and in-context learning capabilities (Wei et al., 2022a) have led to widespread use in answering questions about data, generating summaries, reports, code and insights.

Despite these advances, data-to-text generation remains a challenging problem (Wiseman et al., 2017; Osuji et al., 2024), as it requires (i) faithfulness to the input data, (ii) inferences over that data, and (iii) domain-aware intuition for what is relevant, important, or interesting. For example, a surface-level summary might state: "*Region A recorded 12 incidents this year.*", a deeper insight would be "*Despite accounting for nearly a third of all incidents, region A decreased by 20% from last year, while most other regions increased.*

LLMs suffer from hallucinations (producing outputs unsupported or contradicting the input data) (Obaid ul Islam et al., 2023) and factual inaccuracies (Thomson et al., 2023; Dusek et al., 2019). They struggle with inference operations (Creswell et al., 2023; Bubeck et al., 2023; Wu et al., 2024) despite their fluency and surface-level coherence. And they lack reliable relevance judgment (Soboroff, 2025) often producing shallow or unhelpful outputs. These limitations highlight that surface generation and genuine data understanding remain distinct capabilities.

1.1 Challenges

Underspecification Xu (2025) argues that hallucinations are inevitable under the open world assumption and should be understood as a manifestation of a generalization problem. Moreover, tasks are often not well defined (Zhou and Shbita, 2026), leaving outputs open to multiple interpretations. Underspecification is a fundamental issue generally in machine learning (D'Amour et al., 2022).

From a logical perspective, inference is invalid without a complete set of premises (Copi et al., 2016). When premises are missing, humans and models supply them implicitly. These implicit assumptions (presuppositions) may vary across users, models, and contexts, causing errors in inference and confounded faithfulness evaluations. A striking example of incorrect assumption is the emer-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1236–1252
July 2-7, 2026 ©2026 Association for Computational Linguistics

gence of the phrase *vegetative electron microscopy*, which entered the scientific literature likely through language models trained on corrupted texts.[1] The optical character recognition implicitly assumed a single-column page layout, merging unrelated text fragments into a nonexistent term.

We argue that increasing model capacity does not resolve this issue; more capable models may simply make more subtle but still unintentional assumptions. Therefore, we want to treat the task as inherently incomplete by design. Rather than attempting to eliminate assumptions, the key challenge is to surface, represent, and reason about them explicitly, similarly to what is done with enthymeme detection and reconstruction (finding unsaid premises of arguments; Stahl et al., 2023).

Deeper Inferences Another limitation of current systems is their tendency to produce shallow or uninteresting outputs that lack diversity. The models frequently overlook deeper insights (Berglund et al., 2024). This raises the question of whether current systems are truly performing inference, or merely reproducing patterns based on latent knowledge and prior assumptions learned during training (Shojaee et al., 2025). Interestingness of an insight also depends on the human assumptions and prior knowledge, novel or surprising insights for one user may be obvious or irrelevant to another.

Humans routinely operate in incomplete and underspecified settings using approximate and rule-of-thumb reasoning, which relies on heuristics, abstractions, and mental models rather than precise numerical computation. For this, it is important to assess uncertainty (Xiong et al., 2024) and relevance, at which the models were shown not to perform well (Xiong et al., 2024; Soboroff, 2025). One of the ways to formalize this is the research area of *qualitative reasoning* (Kuipers, 1994).

Finally, insight generation often requires adopting an open-world perspective, where data is interpreted relative to real-world expectations, norms, or prior knowledge. In such cases, factuality and consistency with real-world knowledge becomes important alongside strict faithfulness to the input data alone (Xu, 2025).

Evaluation Evaluating correctness of generated insights is challenging. Current approaches rely largely on benchmark-driven evaluation with poorly defined objectives and subjective interpretations (Raji et al., 2021). Generally, benchmarks contain label errors and ambiguities, which leads to introduction of re-evaluated "platinum benchmarks" (Vendrow et al., 2025). Most benchmarks are satisfied with surface-level information and require minimal inference, e.g. WebNLG (Gardent et al., 2017b; Cripwell et al., 2023) and LogicNLG (Chen et al., 2020a). As a result, models are rarely evaluated on deeper inference or comprehensive content selection (Puduppully et al., 2019). Beltoft and Galke (2025) also criticize the fact that current AI systems and benchmarks overemphasize quantitative, numerical reasoning at the expense of qualitative understanding. This motivates the need for better methods that account for assumptions and evaluate inferences.

1.2 Goals

This thesis aims to improve insight generation from structured data and address these limitations by treating assumptions as an important part of data-to-text generation and automated insight discovery.

- **RQ1**: *To what extent do implicit assumptions compromise the factual validity of generated insights?*
- **RQ2**: *How can we use symbolic and qualitative reasoning to generate factual insights with deeper inferences?*
- **RQ3**: *In what way do assumptions and depth of inferences determine the perceived interestingness of an insight?*

Figure 1 illustrates the conceptual structure of the proposed research agenda with examples and thesis sections. This work focuses on data-to-text generation, primarily table-to-text generation. We may also explore dialogue or text summarization in order to test how well our approach generalizes. Our experiments will mainly use open models to ensure reproducibility; closed models may be used mainly for preliminary evaluation or comparison.

1.3 Outline

Section 2 reviews the theoretical foundations and methodology. In section 3, we explore the first and third research question, reviewing relevant literature and detailing published and ongoing work, and the aims for future work. Section 4 covers the second and third research question similarly. Finally, section 5 summarizes our findings and objectives.

[1] https://retractionwatch.com/2025/02/10/vegetative-electron-microscopy-fingerprint-paper-mill/

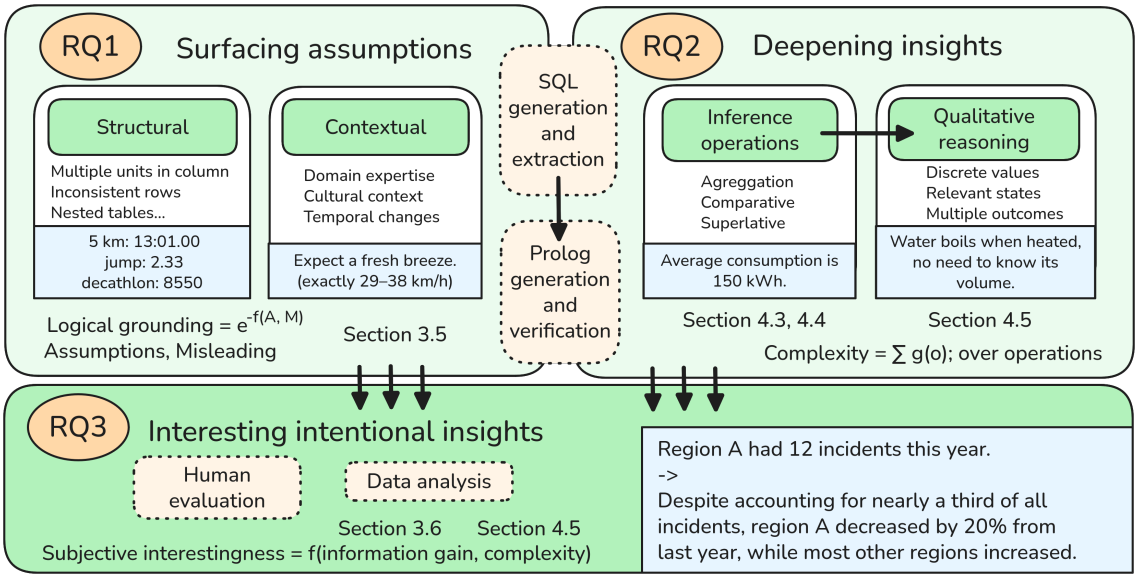

Figure 1: Surfacing assumptions (*RQ1*) provides explicit premises to logically ground the generations. *RQ2* constructs deeper inferences, and both jointly support the evaluation of interestingness of insights (*RQ3*).

2 Background

Modern LLMs are based on the transformer architecture (Vaswani et al., 2017), contemporary models are fine-tuned to follow instructions and aligned using reinforcement learning from human feedback (Ouyang et al., 2022). Recently, reasoning models have been trained using reinforcement learning to encourage the generation of intermediate chain-of-thought reasoning before producing the final answer (Guo et al., 2025). This training typically involves math and coding tasks, where the accuracy can be automatically verified and used as a reward signal (Guo et al., 2025).

Prompting strategies for these models include zero-shot and few-shot prompting (Brown et al., 2020), chain-of-thought (CoT) reasoning (Wei et al., 2022b), and agentic methods (Wang et al., 2024a). Prompt optimization techniques improve consistency and performance (Khattab et al., 2023; Pryzant et al., 2023; Wang et al., 2024b). For tasks requiring structured output, constrained decoding methods are particularly useful, as they enforce adherence to a predefined schema (Geng et al., 2023).

2.1 Table-to-Text Generation

Table-to-text generation (Liu et al., 2018; Parikh et al., 2020a) is a subtask of data-to-text generation in which table summaries or textual insights are produced from a structured table. It is regarded as one of the most challenging NLG tasks (Wiseman et al., 2017; Osuji et al., 2024) and has been studied for decades (Barzilay and Lapata, 2005). Due to LLMs limitations in faithfulness, high-stakes domains still rely on rule and template based pipelines (Reiter and Dale, 1997; Dale, 2023). While such approaches offer strong control and reliability, they are inflexible due to reliance on specific templates, and difficult to scale as they are labor-intensive.

Neural end-to-end approaches (Gatt and Krahmer, 2018) excel in producing fluent text but struggle with content planning, i.e., selecting information to be communicated. End-to-end neural models (Wiseman et al., 2017) include table-aware architectures (Xing and Wan, 2021), fine-tuned neural language models (Nan et al., 2022; Zhao et al., 2023a; Kantharaj et al., 2022), hybrid models (Andrejczuk et al., 2022) and LLMs (Kasner and Dusek, 2024a). Most of these approaches lack explicit intermediate meaning representations, making them difficult to interpret and prone to hallucinations (Obaid ul Islam et al., 2023).

2.2 Structured Reasoning

Some hybrid methods incorporate symbolic content planning or explicit multi-step logical reasoning with neural generation, often through handcrafted logical forms (Chen et al., 2020a; Liu et al., 2022a; Alonso and Agirre, 2024; Saha et al., 2023). While these representations improve interpretability and can be checked for validity, they are typically benchmark-specific and require task-specific fine-tuning and extensive manual effort.

Recent works explore the use of LLMs (Zhao et al., 2023b; Pérez et al., 2025). Zhao et al. (2023b) outperform prior methods with few-shot and chain-of-thought prompting, without task-specific training. Others fine-tune models for tabular tasks, including Li et al. (2023c) and Bian et al. (2024), using a two-step approach to first highlight relevant table regions before generating insights.

In tabular question answering (QA), code generation improves the accuracy of the answers by executing generated SQL queries or Python code on the data (Cheng et al., 2023; Jiang et al., 2023), and generating synthetic data for table-aware fine-tuning (Li et al., 2023d). Pérez et al. (2025) extend this to table-to-text generation, employing SQL to retrieve data and enhance output faithfulness.

However, insight generation requires inferring relevance rather than completing a question. These methods also improve accuracy but do not explicitly address interpretability or the generation of genuinely interesting insights, which have primarily been studied in the context of visual analytics (Law et al., 2020; Brath and Hagerman, 2021).

2.3 Evaluation

Evaluation of generated insights involves multiple dimensions, including faithfulness to the input data, factuality regarding real-world knowledge (Maynez et al., 2020), and diversity (Zhu et al., 2018). Automatic comparisons to human-written references are generally unsuitable, as there are many valid insights in the data; even hybrid metrics like PARENT (Dhingra et al., 2019) do not fully account for this. Reference-free metrics often rely on pretrained models such as TAPAS (Herzig et al., 2020a), TAPEX (Liu et al., 2022b), or natural language inference scores (Chen et al., 2020a). However, these metrics exhibit biases and miss deeper insights. As a result, NLG evaluation often relies on human evaluation (Sai et al., 2023; van der Lee et al., 2021) or LLM-as-a-judge methods (Zheng et al., 2023; Gu et al., 2024).

Human evaluation remains the most reliable approach when following best practices (Howcroft et al., 2020), though replicability remains challenging (Belz et al., 2021). The ReproHum project (Belz et al., 2023), which we participated in (Onderková et al., 2025), aims to improve it by establishing guidelines and recommendations. Human and LLM-based evaluations also enable richer analysis, such as ranking outputs and error span annotation (Thomson and Reiter, 2020).

Evaluation is commonly based on benchmarks such as WebNLG (Gardent et al., 2017a) and DART (Nan et al., 2021) for triples-to-text, LogicNLG (Chen et al., 2020a) and LoTNLG (Zhao et al., 2023b) datasets of Wikipedia tables based on TabFact (Chen et al., 2020b) and WikiTableQuestions (Pasupat and Liang, 2015), ToTTo (Parikh et al., 2020b), or databases in BIRD (Li et al., 2023a). LLMs have been shown to memorize benchmarks during pre-training or fine-tuning (Oren et al., 2024; Xu et al., 2024; Sainz et al., 2023; Balloccu et al., 2024), inflating performance. Benchmarks also exhibit uneven performance across domains (Hu et al., 2025; Diao et al., 2025; Zhu et al., 2025).

3 Presuppositions

This section focuses on how underspecified data and tasks lead to unintentional assumptions during data-to-text generation, affecting faithfulness, factuality (*RQ1*), and inference quality including interestingness of insights (*RQ3*). We study this problem in both neural text generation systems and neuro-symbolic approaches combining code generation with execution to improve accuracy.

We treat inference as an interaction between data and assumptions, rather than a purely data-driven process. In particular, we want to investigate whether models can identify and reconstruct missing premises (Stahl et al., 2023). Rather than committing to a single interpretation, we want to explore methods for surfacing multiple plausible assumptions to mitigate premature closure (Inayat et al., 2024), where an LLM-generated answer discourages further human exploration.

Just as a photographer's intention determines the focus and exposure of a photograph, making assumptions explicit, we aim to make systems more controllable and give users greater agency. Explicit premises also enable rigorous and interpretable evaluation, as stated assumptions can be examined and verified using formal or semi-formal verification methods (Quan et al., 2025, 2024).

For *RQ1*, we propose "logical grounding" (*LG*) of an insight as a function of its implicit assumptions A and misleading claims M:

$$LG = e^{-f(A,M)}, \quad LG \in [0, 1]. \tag{1}$$

Here, *A* quantifies the implicit assumptions and *M* is the number of misleading claims. We define:

$$A = \sum_{i \in \mathcal{I}_A} w(i), \quad A \in [0, \infty). \tag{2}$$

where I_A is the set of missing atomic assumptions that must be added for the inference to be logically valid, given there are no previous contradictions. *w(i)* measures the ambiguity of assumption *i* as the number of its plausible interpretations. The misleading claims *M* suggest that a property holds uniquely, while it is not discriminative (e.g. *"Alice has two cookies"* if Bob also has two). We aim to estimate both the functional forms and behavior of LLMs in under-specified conditions, using (i) iterative assumption generation with symbolic verification and (ii) systematic data corruption.

There are many aspects of interestingness, including actionability, conciseness, coverage, diversity, novelty, peculiarity, reliability, surprisingness, and utility (Geng and Hamilton, 2006). For *RQ3*, we focus on interestingness as a function of information gain (Shannon, 1948) under the reader's assumptions, which is closest to novelty and surprisingness. We will begin by analyzing human-written data-to-text artifacts and then conduct a human evaluation.

3.1 Linguistic Foundations

Presuppositions are central in language understanding, as humans routinely rely on shared assumptions and leave parts of meaning implicit. Linguistic theory characterizes them as unstated background conditions that are necessary for an utterance to be meaningful (Parrish et al., 2021). They are often introduced by presupposition triggers, including clause-embedding predicates (e.g, *know*, *think*), implicatives (e.g., *manage to*, *fail to*), change-of-state verbs, clefts, comparatives, aspectual verbs, temporal adverbs and numerical determiners (Parrish et al., 2021). For example, the determiner in "*the sun*" triggers the assumption that it is only one (our) Sun, not just any star.

Presuppositions often project through negation (e.g., *"Chet didn't finish law school"* presupposes he attended it), though their strength varies by context (Parrish et al., 2021). While some triggers impose hard constraints (e.g., "*both apples*" requires exactly two), "soft" triggers, unlike entailment, allow presuppositions to be canceled or revised (e.g., *"Chet didn't finish law school. He attended med school."* where the second sentence cancels the law school presupposition) (Parrish et al., 2021).

In addition to semantically grounded presuppositions tied to specific expressions, pragmatic presuppositions arise from discourse context and shared expectations between participants (Yu et al., 2023).

3.2 Presuppositions in Question Answering

Work in QA examined questions that are unanswerable, ambiguous or contain false presuppositions to improve generations (Kim et al., 2021; Han et al., 2023). For example, "*Who won the Ethiopia-Italy war?*" is ambiguous since there were two wars, "*Who won the Nobel peace prize for DNA structure?*" has a false presupposition (Han et al., 2023).

Kim et al. (2021) show that a many unanswerable questions can be explained by unverifiable presuppositions, and that generating and verifying presuppositions improves system responses. Han et al. (2023) build on this idea using LLM prompting, demonstrating improved performance without parameter updates by explicitly extracting presuppositions and using them as as working memory. More recent frameworks decompose claims into presupposition-free subquestions and verify them iteratively using LLMs (Dipta and Ferraro, 2025). LLMs may also adopt false presuppositions (even obviously so), particularly in politically sensitive contexts, highlighting how they can subtly introduce misleading information (Sieker et al., 2025).

3.3 Resources

Several datasets support the study of presuppositions in LLMs. The NOPE corpus (Parrish et al., 2021) evaluates a model's ability to predict human inferences across multiple presupposition triggers. They show that older natural language inference transformer-based models succeed in simple cases but fail on more complex ones (Parrish et al., 2021). The CREPE dataset (Yu et al., 2023) focuses on detecting and correcting false presuppositions in Reddit questions. The models asses if a sentence contains a false presupposition and, if so, output it alongside a corrected sentence. Francis et al. (2024) introduced a task to explicitly generate presuppositions for an input sentence. They release two datasets: PGen containing premises and corresponding lists of presuppositions to be generated, and PECaN containing premise-hypothesis pairs for classifying presuppositions.

3.4 Published Work

To address limitations in table-to-text generation evaluation, we developed FreshTab (Onderková et al., 2025), a dynamic benchmark for insight generation, following prior work on dynamic dataset construction (Kasner and Dusek, 2024b; White et al., 2024). FreshTab collects tables from

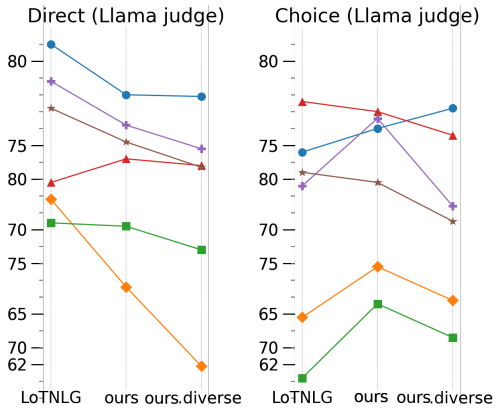

Figure 2: Comparison of faithfulness on old (LoTNLG) and new datasets (FreshTab), viz Section 3.4.

Wikipedia using only articles created after a specified knowledge cutoff date to combat LLM memorization. It supports non-English Wikipedias and domain-sensitive evaluation through the Wikidata ontology (Zhu et al., 2025). This work received the Best Short Paper Award at INLG 2025, and its findings motivate research questions in this proposal.

We compare the insights generated on the older LoTNLG dataset (Zhao et al., 2023b) with those on FreshTab, with the same domain distribution and with a diverse domain distribution. Experiments show LLMs behaving differently on the new data, especially with diverse domains (Figure 2), revealing model-domain weaknesses. We also generated insights in German, Russian, and French, observing trends similar to English, although direct comparison is not possible due to differing tables.

We conducted a human study to correlate automatic metrics with human judgments and assess both faithfulness and interestingness. We found very weak correlations for trained entailment-based metrics (Pearson correlation 0.12 for TAPAS (Herzig et al., 2020b) and 0.11 for TAPEX (Liu et al., 2022b)). LLM-based evaluation achieved substantially higher correlation (0.53).

3.5 Ongoing Work

Making presuppositions explicit can support several goals: understanding and processing messy input data, personalizing generation to align with readers' needs, and enabling more rigorous evaluation of faithfulness by clarifying the conditions under which an output is valid.

Structural assumptions During my work on table-to-text generation from Wikipedia tables (Onderková et al., 2025) (Section 3.4), I observed that presuppositions about table structure are frequently violated. A manual analysis revealed several recurring issues: (i) inconsistent columns that mix multiple units (e.g., *kg*, *g*, *lbs*) or value types (e.g., time, meters, points); (ii) inconsistent rows whose semantics differ substantially from the rest of the table, such as aggregate rows of referee between players; (iii) transposed tables, where attributes appear in the first column and values populate subsequent columns; (iv) tables containing nested sub-tables; and (v) other structural irregularities, including primarily visual tables lacking a consistent schema.

I conducted preliminary experiments prompting LLMs to detect such problematic tables, with partial success. I evaluated these cases using the *SQuirreL* pipeline (Section 4.4), where text-to-SQL execution proved less reliable for such difficult tables. To address *RQ1*, I stress-test such structural presuppositions by systematically corrupting clean tables from the BIRD dataset (Li et al., 2023b), altering the semantic structure and injecting inconsistencies. This setup evaluates whether models can detect violations and generate explicit structural assumptions before producing multiple plausible outputs.

Contextual assumptions Beyond structure-related assumptions, I identified several content-related presupposition issues: (i) domain expertise, for example terms such as *fresh breeze* correspond to specific wind speed ranges for meteorologists but not for lay users; (ii) cultural context, such as whether lower values on a rating scale indicate better or worse performance; (iii) temporal assumptions, including changes in rules or conventions over time; and (iv) other context-dependent expectations. These presuppositions may differ across annotators, posing challenges for evaluation when they are not made explicit. Moreover, formal or semi-formal logical verification requires all relevant premises to be stated explicitly.

I am now exploring methods to systematically identify missing assumptions in human-written NLGs gold-references and quantify underspecification. To this end, I use a neuro-symbolic pipeline based on Prolog: (i) input data are encoded as facts in a Prolog knowledge base; (ii) entities and logical inferences are extracted from textual insights using an LLM and translated into Prolog queries; and (iii) the resulting queries are symbolically verified.

For example, the insight *"Australia and England have same number of wins."* can be expressed as: *"cell(R1, 'nation', 'australia'), cell(R1, 'wins',*

V1), cell(R2, 'nation', 'england'), cell(R2, 'wins', V2), V1 = V2". Prolog provides a flexible representation for logical operations and table structure, although the pipeline requires further refinement. In a preliminary analysis of LogicNLG instances flagged as containing presuppositions, a manual inspection of 20 cases identified two with missing assumptions and two with misleading claims.

As a complementary approach, I prompted LLM to spot and generate the missing assumptions. This approach now flags 29% of insights as missing assumptions, which is too high and needs improving. I plan to combine both methods and iteratively generate possible missing assumptions and verify them in Prolog, until it finds a valid solution with fitting atomic assumptions. This could improve benchmark labels to "platinum" (Vendrow et al., 2025), supporting more automated analysis for *RQ1*.

3.6 Future Work

I plan to participate in the SciClaimEval shared task (Ho et al., 2026) on verification of scientific claims from articles against their tables and figures. I expect our work on verification of table data through neuro-symbolic methods will be useful and the task will provide new types of data to explore to make our work more general. The task also contains a hierarchy of contexts that are needed for the claim disambiguation, which I hope will be helpful for our work on presuppositions.

Interestingness To address *RQ3*, I model interestingness as information gain relative to a reader's prior presuppositions. This is motivated by my analysis of current model failures, which shows a tendency to favor superficial insights. I hypothesize that perceived interestingness depends on the reader's prior knowledge. For example, an expert who expects a typical number of yellow cards in a football match may find insights about deviations from this expectation interesting, whereas a non-expert may not. Conversely, non-experts may find simpler observations more interesting due to the lack of such priors.

I will analyze existing data sources, including Wikipedia tables and their corresponding articles. With a way to identify assumptions, I will analyze the human annotated data we had already collected from human annotation for the FreshTab. Finally, I will design and conduct a human evaluation to quantify these relationships and assess how assumptions influence perceived interestingness.

4 Deeper Inference

Building on a clearer problem formulation, *RQ2* investigates how models can generate deeper insights with non-trivial inference over data. Prior work focused on applying predefined logical operations over structured inputs, primarily to improve controllability and diversity of generated insights (Chen et al., 2020c; Zhao et al., 2023a,c). While specific formulations vary, these operations typically include aggregation (e.g., computing sums or averages), comparison (contrasting values between entities), and counting (determining how many items meet a condition). We will build on these operators while exploring complementary forms of inference, especially qualitative reasoning through agentic approaches, which may better support insight generation in underspecified settings.

To study *RQ3*, i.e., how depth of inferences influence perceived interestingness (*PI*), we define inference complexity (*IC*) as:

$$IC = \sum_{o \in \mathcal{O}} g(o), \quad IC \in \mathbb{R}_{\geq 0}, \tag{3}$$

where $\mathcal{O}$ denotes the set of inference operations and $g(o)$ their respective contribution to complexity. We hypothesize that both overly simple (directly seen from data) and overly complex (too niche) inferences reduce *PI*, suggesting a non-monotonic relationship. Additionally, we hypothesize that different types of operations and their compositions contribute differently to the perceived interestingness. We will analyze existing data and conduct a human evaluation to estimate the relationship between logical structure and perception.

Finally, we model total *PI* as a function of inference complexity and user's information gain (*IG*):

$$IG = D_{KL}(P_{post} \parallel P_{prior}), \tag{4}$$

where P_{prior} and P_{post} denote distributions over interpretations of the data under the reader's assumptions before and after observing the insight.

4.1 Logical Reasoning

Logical reasoning provides formal frameworks for deriving conclusions from premises, while *deduction* derives conclusions that are logically entailed by given premises, *induction* generalizes rules from observed instances, and *abduction* seeks the most plausible explanation that accounts for available (and also hypothetical) evidence and rules (Walton, 2014). Unlike deduction, induction and abduction

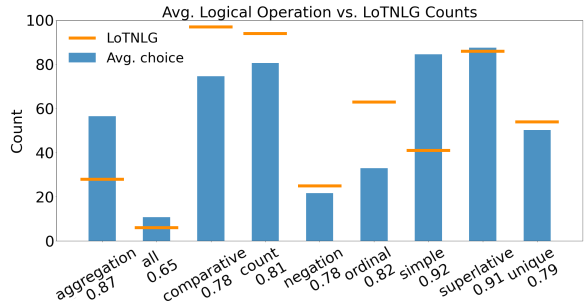

Figure 3: Comparison of logical operation counts. Human choices = orange line, LLM choices = blue bars.

are forms of defeasible inference; even with valid premises, the conclusion may not be deductively valid (Walton, 2014). Abduction is especially relevant for insight generation, as it is the only form of reasoning among these that can introduce novel explanatory hypotheses rather than recombining existing information (Bhagavatula et al., 2020). Prior work shows abductive reasoning can extract latent knowledge from LLMs (Jung et al., 2022), which may help surface plausible assumptions. Abductive formalisms are also used in automated planning to infer prior events from goals (Shanahan, 2000).

4.2 Qualitative Reasoning

Qualitative reasoning (Forbus, 1997) is motivated by human cognition, where people rely on coarse abstractions, heuristics, and mental models rather than precise numerical computation. This is useful when exact measurements are unavailable, unnecessary, or infeasible to obtain. For example, predicting that water on a lit stove will eventually boil and evaporate requires no knowledge of exact temperatures or heat transfer rates (Iwasaki, 1997).

It is built on principles that preserve essential behavioral properties rather than exact quantities:

- Discretization: continuous values represented symbolically (increasing, decreasing, stable)
- Relevance: focused on relevant states or thresholds (boiling and freezing point)
- Ambiguity: allows multiple outcomes

These properties make qualitative reasoning well suited to incomplete, noisy, or underspecified settings. In data-to-text generation, it naturally engages with factuality by using knowledge of world.

4.3 Published Work

Abductive Reasoning Tasks My master's thesis studied the performance of language models on abductive reasoning tasks (Onderková and Nickles, 2023). I evaluated then-prominent open models on the ART dataset (Bhagavatula et al., 2020). Only Flan-T5 (Chung et al., 2024) achieved an accuracy above random choice, while few-shot prompting did not yield improvements over zero-shot settings.

I further explored abductive reasoning for content selection and discourse planning by prompting Flan-T5 on synthetic weather data. While the models are somewhat effective at selecting relevant content (mapping data to outcomes such as snow, rain or clear weather), discourse planning proved a significant challenge. I might revisit abductive inference with more capable contemporary LLMs.

Inference Operations In FreshTab (see Section 3.4), we studied table-to-text insight generation using the same set of logical operations and two-shot chain-of-thought prompting setup as Zhao et al. (2023c). The model directly generates an insight given a table and a specified logical operation. According to an LLM judge, models adhered to the selected logical operation in 81% cases.

I additionally introduced a more flexible setup where the model, provided with descriptions of available logical operations, selects which operation to apply for a given table. This allowed us to analyze model preference for logical operations relative to gold human annotations. LLMs strongly favor *simple* (e.g., *player A gets 2 points*), surface-level outputs, while more complex reasoning types such as *ordinal* (e.g., *player B has the second most points*) and *comparative* (e.g., *player C has more points than player A*) operations are underused (Figure 3). Moreover, automatic metrics tend to be much more reliable for *simple* surface-level generations than for inference operations such as *ordinal* and *comparative*. Those findings inform our *RQ2*.

Content effect We participated in SemEval 2026 Task 11 on *Disentangling Content and Formal Reasoning in Language Models* (Valentino et al., 2026), which studies syllogistic reasoning under plausible and implausible (contradicting world knowledge) premises. Our submission uses a neuro-symbolic pipeline with a small LLM and a first-order logic prover. Even a small Qwen 3 4B model (Yang et al., 2025) shows a strong capability in translating natural language into first-order logic, especially when prompted to produce LaTeX representations.

I developed an alternative pipeline based on Aristotelian reasoning, highlighting differences from modern first-order logic. To assess whether mod-

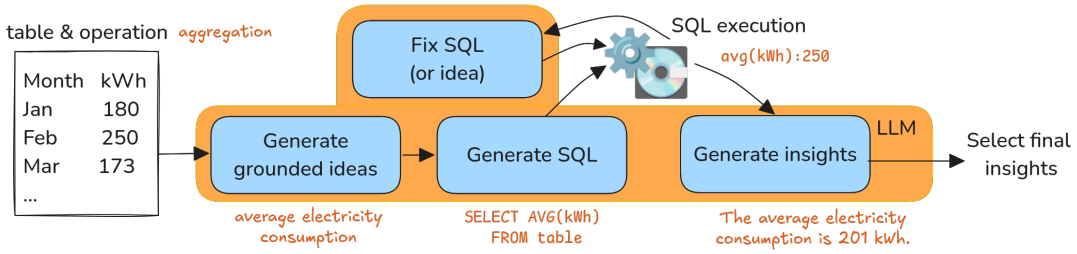

Figure 4: Overview of the *SQuirreL* approach (with an illustrative example shown in orange).

els adhere to the specified logical framework or instead default to learned patterns, I compare this method with zero-shot approach. While the models reliably recall Aristotelian rules, they often fail to apply them for inference and default to first-order logic. When rules are explicitly specified in the prompt, the model is able to follow them. This invites analysis of alternative logical frameworks.

4.4 Ongoing Work

Beyond the prompting-based methods evaluated in Section 3.4, we developed *SQuirreL*, a neuro-symbolic pipeline with SQL code execution to explore whether code generation can support deeper, more reliable insight generation. The pipeline comprises four stages, see Figure 4: (1) insight *idea generation* using an LLM over the table and a set of logical operations, (2) SQL *query generation* for data retrieval, (3) execution of the SQL query, and (4) final insight generation grounded in the query results. When the SQL execution fails, a reflection-and-retry mechanism attempts to correct the query or fall back to a simplified idea.

To identify failure points, we evaluated several ablations: (i) removing reflection and retry to assess its contribution, (ii) a setup without idea generation, (iii) replacing SQL execution with LLMs retrieval, and (iv) a truncated-data setting where only the first three rows of a table are provided to test generalization and scalability to large tables.

While *SQuirreL* reduces incorrect insights compared to prompting, it produces more misleading insights, resulting in comparable overall quality to direct prompting baselines. This is primarily due to noisy real-world tables and limitations in text-to-SQL generation. Performance degrades notably under table truncation. However, preliminary human evaluation suggests that *SQuirreL* produces fewer trivial summaries and more genuinely insightful and interesting outputs than the baseline.

To address *RQ2*, I am moving beyond SQL and numerical calculation toward queries over knowledge bases and logic programming languages like Prolog or Isabelle, aligning with work on structured critique and formal verification of natural language explanations (Quan et al., 2025, 2024). As these formalisms are underrepresented in typical training data, we anticipate a need for task-specific fine-tuning, potentially via instruction-tuning CoT demonstrations (Ranaldi and Freitas, 2024).

4.5 Future Work

I plan to explore the qualitative reasoning for NLG, using "softer" CoT prompting for quasi-symbolic abstractions (Ranaldi et al., 2025) and qualitative reasoning over data or causal sketches. These formalisms should enable more human-like qualitative reasoning, better suited for noisy and incomplete environments. The qualitative reasoning seems suitable for using agentic frameworks such as multi-agent debate (Smit et al., 2024) to improve inference depth and factuality.

As for *RQ3*, I will analyze insight complexity by identifying the number of logical operations in existing data, using both neural models and complementary regex approaches. The most decisive will be a well-planned human evaluation, for which I will systematically craft insights using rule-based methods and templates to comprehensively cover various logical operations and their combinations.

5 Conclusion

This thesis proposal reframes data-to-text generation as an assumption-sensitive inference problem, rather than a data-driven transformation task. By intentionally surfacing tacit assumptions and exploring structured reasoning strategies, we aim to move beyond surface-level insights toward deeper inference. We presented completed and ongoing work on presuppositions and logical inference, and outlined a research agenda dedicated to improving the reliability, interestingness, interpretability, and controllability of insight generation.

Limitations

This work focuses primarily on data-to-text generation, which may restrict the direct generalizability of findings to other tasks. Most experiments rely on open-weight language models, which currently lag behind closed models in some capabilities. The evaluation of deeper inference, presuppositions, and interestingness is subjective and depends on the user. Automatic evaluations have specific preferences and biases that may not be generalizable across diverse users and domains. The reasoning patterns explored, while extending beyond standard logical operators, still represent a constrained subset of possible strategies.

Acknowledgements

This research was co-funded by the European Union (ERC, NG-NLG, 101039303), the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO, and Charles University projects GAUK 506526 and SVV 260 821.

Artificial intelligence tools were used according to the *ARR AI writing assistance policy* for better recall in the literature search and as an assistance with the language in the paper.

References

Iñigo Alonso and Eneko Agirre. 2024. Automatic Logical Forms improve fidelity in Table-to-text generation. *Expert Syst. Appl.*, 238(Part D):121869.

Ewa Andrejczuk, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yasemin Altun. 2022. Table-to-text generation and pre-training with TabT5. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6758–6766, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 331–338, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Stine Lyngsø Beltoft and Lukas Galke. 2025. Not Everything That Counts Can Be Counted: A Case for Safe Qualitative AI. *CoRR*, abs/2511.09325.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A Systematic Review of Reproducibility Research in Natural Language Processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Stickland, Tomek Korbak, and Owain Evans. 2024. The reversal curse: Llms trained on "a is b"fail to learn "b is a". In *International Conference on Learning Representations*, volume 2024, pages 18623–18642.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive Commonsense Reasoning. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia. OpenReview.net.

Junyi Bian, Xiaolei Qin, Wuhe Zou, Mengzuo Huang, Congyi Luo, Ke Zhang, and Weidong Zhang. 2024. HeLM: Highlighted Evidence Augmented Language Model for Enhanced Table-to-Text Generation. ArXiv:2311.08896 [cs].

Richard Brath and Craig Hagerman. 2021. Automated Insights on Visualizations with Natural Language Generation. In *25th International Conference Information Visualisation, IV 2021*, pages 278–284, Sydney, Australia. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

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

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020b. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia. OpenReview.net.

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020c. Logic2Text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2096–2111, Online. Association for Computational Linguistics.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding Language Models in Symbolic Languages. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, Kigali, Rwanda. OpenReview.net.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Irving Copi, Carl Cohen, and Daniel Flage. 2016. *Essentials of Logic*. Routledge.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, Kigali, Rwanda. OpenReview.net.

Liam Cripwell, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, William Soto-Martinez, and Craig Thomson. 2023. The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 55–66, Prague, Czech Republic. Association for Computational Linguistics.

Robert Dale. 2023. Navigating the text generation revolution: Traditional data-to-text NLG companies and the rise of ChatGPT. *Natural Language Engineering*, 29(4):1188–1197.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2022. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, 23(226):1–61.

Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019. Handling Divergent Reference Texts when Evaluating Table-to-text Generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, and Zhuosheng Zhang. 2025. GuideBench: Benchmarking Domain-oriented Guideline Following for LLM Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025*, pages 11361–11399, Vienna, Austria. Association for Computational Linguistics.

Shubhashis Roy Dipta and Francis Ferraro. 2025. If We May De-Presuppose: Robustly Verifying Claims through Presupposition-free Question Decomposition. *CoRR*, abs/2508.16838.

Ondrej Dusek, David M. Howcroft, and Verena Rieser. 2019. Semantic Noise Matters for Neural Natural Language Generation. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

Kenneth D Forbus. 1997. Qualitative Reasoning.

Maria Francis, Julius Steuer, Dietrich Klakow, and Volha Petukhova. 2024. Who did you blame when your project failed? Designing a corpus for presupposition generation in cross-examination dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17564–17574, Torino, Italia. ELRA and ICCL.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

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

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.

Liqiang Geng and Howard J. Hamilton. 2006. Interestingness Measures for Data Mining: A Survey. *ACM Comput. Surv.*, 38(3):9.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A Survey on LLM-as-a-judge. *CoRR*, abs/2411.15594.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-R1: Incentivizing Reasoning Capability in LLMS via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Wookje Han, Jinsol Park, and Kyungjae Lee. 2023. PreWoMe: Exploiting Presuppositions as Working Memory for Long Form Question Answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 8312–8322, Singapore. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020a. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020b. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 4320–4333, Online. Association for Computational Linguistics.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, Andre Greiner-Petter, and Akiko Aizawa. 2026. SciClaimEval: Cross-modal Claim Verification in Scientific Papers. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026)*, Palma de Mallorca, Spain. ELRA Language Resources Association.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Large Language Models Are Cross-lingual Knowledge-free Reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque*, pages 1525–1542, New Mexico, USA. Association for Computational Linguistics.

Shahzad Inayat, Ahtisham Younas, Sergi Fàbregues, and Parveen Ali. 2024. Premature Closure of Analysis in Qualitative Research: Identifying Features and Mitigation Strategies. *International Journal of Qualitative Methods*, 23:16094069241234187.

Y. Iwasaki. 1997. Real-world Applications of Qualitative Reasoning. *IEEE Expert*, 12(3):16–21.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. 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*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic Prompting: Logically Consistent Reasoning With Recursive Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq 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)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Zdenek Kasner and Ondrej Dusek. 2024a. Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 12045–12072, Bangkok, Thailand. Association for Computational Linguistics.

Zdenek Kasner and Ondrej Dusek. 2024b. Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 12045–12072, Bangkok, Thailand. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-improving Pipelines. *CoRR*, abs/2310.03714.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 3932–3945, Virtual Event. Association for Computational Linguistics.

Benjamin Kuipers. 1994. *Qualitative Reasoning: Modeling and Simulation With Incomplete Knowledge*. MIT press.

Po-Ming Law, Alex Endert, and John T. Stasko. 2020. Characterizing Automated Data Insights. In *31st IEEE Visualization Conference, IEEE VIS 2020 - Short Papers*, pages 171–175, Virtual Event, USA. IEEE.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin Chen-Chuan Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023a. Can LLM Already Serve as A Database Interface? A BIg Bench for Large-scale Database Grounded Text-to-SQLs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin Chen-Chuan Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023b. Can LLM Already Serve as A Database Interface? A BIg Bench for Large-scale Database Grounded Text-to-SQLs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2023c. Table-GPT: Table-tuned GPT for Diverse Table Tasks. *CoRR*, abs/2310.09263.

Zhenyu Li, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2023d. Toward a Unified Framework for Unsupervised Complex Tabular Reasoning. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA*, pages 1691–1704, USA. IEEE.

Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han, and Dongmei Zhang. 2022a. PLOG: Table-to-logic pretraining for logical table-to-text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5531–5546, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022b. TAPEX: Table Pre-training via Learning a Neural SQL Executor. In *The Tenth International Conference on Learning Representations, ICLR 2022*, Virtual Event. OpenReview.net.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text Generation by Structure-aware Seq2seq Learning. 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)*, pages 4881–4888, New Orleans, Louisiana, USA. AAAI Press.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 1906–1919, Online. Association for Computational Linguistics.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. *CoRR*, abs/2402.06196.

Linyong Nan, Lorenzo Jaime Flores, Yilun Zhao, Yixin Liu, Luke Benson, Weijin Zou, and Dragomir Radev. 2022. R2D2: Robust data-to-text with replacement detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6903–6917, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*, pages 432–447, Online. Association for Computational Linguistics.

Saad Obaid ul Islam, Iza Škrjanec, Ondrej Dusek, and Vera Demberg. 2023. Tackling hallucinations in neural chart summarization. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 414–423, Prague, Czechia. Association for Computational Linguistics.

Kristýna Onderková, Ondrej Plátek, Zdenek Kasner, and Ondrej Dusek. 2025. FreshTab: Sourcing Fresh Data for Table-to-text Generation Evaluation. *CoRR*, abs/2510.13598.

Kristýna Onderková, Mateusz Lango, Patrícia Schmidtová, and Ondrej Dusek. 2025. ReproHum #0669-08: Reproducing sentiment transfer evaluation. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 601–608, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Kristýna Onderková and Matthias Nickles. 2023. Exploring abductive reasoning in language models for data-to-text generation. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, pages 1–4.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving Test Set Contamination in Black-box Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria. OpenReview.net.

Chinonso Cynthia Osuji, Thiago Castro Ferreira, and Brian Davis. 2024. A Systematic Review of Data-to-text NLG. *CoRR*, abs/2402.08496.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions With Human Feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, New Orleans, LA, USA.

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

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020b. ToTTo: A Controlled Table-To-text Generation Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 1173–1186, Online. Association for Computational Linguistics.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Panupong Pasupat and Percy Liang. 2015. Compositional Semantic Parsing on Semi-structured Tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Volume 1: Long Papers*, pages 1470–1480, Beijing, China. The Association for Computer Linguistics.

Alberto Sánchez Pérez, Alaa Boukhary, Paolo Papotti, Luis Castejón Lozano, and Adam Elwood. 2025. An LLM-based Approach for Insight Generation in Data Analysis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque*, pages 562–582, New Mexico, USA. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text Generation with Content Selection and Planning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on*

Educational Advances in Artificial Intelligence, EAAI 2019, pages 6908–6915, Honolulu, Hawaii, USA. AAAI Press.

Xin Quan, Marco Valentino, Danilo Carvalho, Dhairya Dalal, and Andre Freitas. 2025. PEIRCE: Unifying material and formal reasoning via LLM-driven neuro-symbolic refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 11–21, Vienna, Austria. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise A. Dennis, and André Freitas. 2024. Verification and Refinement of Natural Language Explanations through LLM-symbolic Theorem Proving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL*, pages 2933–2958, USA. Association for Computational Linguistics.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*.

Leonardo Ranaldi and André Freitas. 2024. Aligning Large and Small Language Models via Chain-of-thought Reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, and André Freitas. 2025. Improving Chain-of-thought Reasoning via Quasi-symbolic Abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025*, pages 17222–17240, Vienna, Austria. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 1997. Building Applied Natural Language Generation Systems. *Natural language engineering*, 3(1):57–87.

Swarnadeep Saha, Xinyan Yu, Mohit Bansal, Ramakanth Pasunuru, and Asli Celikyilmaz. 2023. MURMUR: Modular multi-step reasoning for semi-structured data-to-text generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11069–11090, Toronto, Canada. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.*, 55(2):26:1–26:39.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Findings of ACL, pages 10776–10787, Singapore. Association for Computational Linguistics.

Murray Shanahan. 2000. An Abductive Event Calculus Planner. *J. Log. Program.*, 44(1-3):207–240.

Claude Elwood Shannon. 1948. A Mathematical Theory of Communication. *The Bell system technical journal*, 27(3):379–423.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *CoRR*, abs/2506.06941.

Judith Sieker, Clara Lachenmaier, and Sina Zarrieß. 2025. LLMs Struggle to Reject False Presuppositions when Misinformation Stakes are High. *arXiv preprint arXiv:2505.22354*.

Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going MAD? A Look at Multi-agent Debate Strategies for LLMs. In *Forty-first International Conference on Machine Learning, ICML 2024*, Proceedings of Machine Learning Research, pages 45883–45905, Vienna, Austria. PMLR / OpenReview.net.

Ian Soboroff. 2025. Don't Use LLMs to Make Relevance Judgments. *Inf. Retr. Res. J.*, 1(1):29–46.

Maja Stahl, Nick Düsterhus, Mei-Hua Chen, and Henning Wachsmuth. 2023. Mind the Gap: Automated Corpus Creation for Enthymeme Detection and Reconstruction in Learner Arguments. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Findings of ACL, pages 4703–4717, Singapore. Association for Computational Linguistics.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating Factual Accuracy in Complex Data-to-text. *Comput. Speech Lang.*, 80:101482.

Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026. SemEval-2026 Task 11: Disentangling Content and Formal Reasoning in Large Language Models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human Evaluation of Automatically Generated Text: Current Trends And

Best Practice Guidelines. *Comput. Speech Lang.*, 67:101151.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. Do Large Language Model Benchmarks Test Reliability? *CoRR*, abs/2502.03461.

Douglas Walton. 2014. *Abductive Reasoning*. University of Alabama Press.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A Survey on Large Language Model Based Autonomous Agents. *Frontiers Comput. Sci.*, 18(6):186345.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P. Xing, and Zhiting Hu. 2024b. PromptAgent: Strategic Planning with Language Models Enables Expert-level Prompt Optimization. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria. OpenReview.net.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent Abilities of Large Language Models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024. LiveBench: A Challenging, Contamination-free LLM Benchmark. *CoRR*, abs/2406.19314.

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

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Xinyu Xing and Xiaojun Wan. 2021. Structure-aware pre-training for table-to-text generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2273–2278, Online. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria. OpenReview.net.

Bowen Xu. 2025. Hallucination is Inevitable for LLMs with the Open World Assumption. *CoRR*, abs/2510.05116.

Cheng Xu, Shuhao Guan, Derek Greene, and M. Tahar Kechadi. 2024. Benchmark Data Contamination of Large Language Models: A Survey. *CoRR*, abs/2406.04244.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain Question Answering with False Presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Yilun Zhao, Zhenting Qi, Linyong Nan, Lorenzo Jaime Flores, and Dragomir Radev. 2023a. LoFT: Enhancing faithfulness and diversity for table-to-text generation via logic form control. In *Proceedings of the*

17th Conference of the European Chapter of the Association for Computational Linguistics, pages 554–561, Dubrovnik, Croatia. Association for Computational Linguistics.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023b. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023c. Investigating Table-to-text Generation Capabilities of Large Language Models in Real-world Information Seeking Scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Yi Zhou and Basel Shbita. 2026. Evaluating Ill-defined Tasks in Large Language Models. *CoRR*, abs/2603.17067.

Qiming Zhu, Jialun Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Shing-Chi Cheung. 2025. DOMAINEVAL: An Auto-constructed Benchmark for Multi-domain Code Generation. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025*, pages 26148–26156, Philadelphia, PA, USA. AAAI Press.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.