Checking for Implicit Assumptions in Data-to-text Generation

Kristýna Onderková and Ondřej Dušek
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University
{onderkova, odusek}@ufal.mff.cuni.cz

1 Introduction

Data-to-text generation is still challenging. Neural natural language generation (NLG) systems, including ones based on large language models (LLMs), continue to produce factual inaccuracies and hallucinations (Thomson et al., 2023), especially for deeper inferences. This is hard to solve, as we lack scalable evaluation for reliable feedback loops. NLG faithfulness metrics, including trained reference-free and LLM judges, have limited reliability, especially for inferences (Onderková et al., 2025). Human evaluation, while most reliable, can still be inconsistent (Sai et al., 2022).

Judging generation faithfulness requires a complete set of premises, which are often unavailable as many tasks are not well defined (Zhou and Shbita, 2026), leaving outputs ambiguous. Consequently, many works rely on benchmarks with subjective interpretations (Raji et al., 2021).

This work-in-progress examines whether human-written benchmark references contain implicit assumptions not grounded in the data. Such assumptions may differ between people (and models vs. people), confounding evaluation. Standard LLMs fail to reliably audit these deep inferences without hallucinating, whereas translating text to formal logic enables strict, auditable verification. Our ultimate goal is designing a "logical grounding" metric to quantify these ungrounded assumptions.

2 Related work

Multiple prior works improve generation faithfulness via code execution (Cheng et al., 2023) or neuro-symbolic verification (Quan et al., 2025). (Kim et al., 2021) show that unverifiable assumptions cause unanswerable questions in question answering; (Dipta and Ferraro, 2025) decompose questions into presupposition-free claims. Vendrow et al. (2025) introduce "platinum benchmarks" by correcting and disambiguating existing ones.

3 Methodology & Experiment

To systematically identify missing assumptions in NLG data, we use a neuro-symbolic pipeline with Prolog and LLM generation to verify gold-standard human references. We focus on the task of insight generation from data tables, though the approach transfers to other data-to-text tasks. The pipeline looks as follows (with examples):

i) **Input tables** to a Prolog knowledge base as cell(Row, Header, Value): *cell(1, day, mon)*
ii) **Generate** Prolog **query** from the claim: *Tuesday is colder than Monday* as *...cell(R2, day, tue), cell(R2, temp, T2), T1 > T2.*
iii) Extract **entities** from the query and **verify** them or try to fix them by generating aliases: *alias(mon, Monday)*
iv) **Verify** the whole **query** with the inference or try to iteratively fix it if an error occurs.
v) If the query is not verified, **generate** possible **assumptions** that could make it correct: *20 > 60 is correct if it is 20°C and 60°F.*
vi) **Verify** the query **with** the added **assumption**.

The pipeline first detects clearly incorrect insights that do not have corresponding entities in the underlying data. Second, we detect if the logical inference fails, which flags an insight for a (possible) missing assumption. Table 1 in the Appendix shows examples of these cases with Prolog queries.

To inform our design, we use the LogicNLG validation set, a standard table-to-text benchmark that includes gold labels (for finding assumptions) with inference operations such as counting, comparisons and aggregations. As an underlying LLM we used

This work 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. It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech MEYS project No. LM2023062).

Qwen 3.5 35B as it was the best performing open weight model of reasonable size. We hand tuned the prompts from a minimal baseline, testing zero-shot, few-shot, chain-of-thought, and specific task constraints. We use the *pyswip* Prolog verifier as it is sufficiently expressive for this task.

4 Results

Based on our experiments, the pipeline verifies approximately 80% of gold claims from LogicNLG. Entity normalization through alias generation improves verification by approximately 15% (from 65%). Iterative Prolog error correction reduces failed queries to approximately 5%. We are currently evaluating query-level critique to address occasional omissions of atomic claims in more complex queries. We are also testing improved data ingestion with beyond symbolic parsing with LLMs-based parsing of complex values such as *'6-3, 6-7 [2-7], 7-6 [7-5]'* (tennis scores).

Assumption generation remains a limitation, particularly for multi-step or compositional cases. For instance, from the claim *"Only one of the artists is named after a state"*, the system identifies the assumption *state('texas')*. However, this alone is insufficient, as the claim also implies that all remaining artists are not named after states.

A manual analysis of 20 potentially missing assumptions shows that ten stem from incorrectly formulated queries, six from erroneous entities in the gold labels, and four from genuinely missing assumptions (reverse counting for BCE years, not specifying AM/FM radio frequencies, missing table headers, and aggregative "total" last row). No logically incorrect claims were identified.

Given that LogicNLG contains 32k examples, an estimate of 20% potentially missing assumptions suggests thousands of incorrect claims and hundreds of missing assumptions. This indicates sufficient scale to support further analysis.

5 Future work

Our next step is to compare pipeline outputs with human-annotated faithfulness judgments to assess the system performance. For this, we plan to use the claims human-annotated for factuality from (Onderková et al., 2025). We will also scale the evaluation to a larger subset of LogicNLG to get more robust results. To improve generality, we will extend the analysis to additional data-to-text benchmarks and models.

We will further improve missing assumption generation through iterative checking in Prolog to construct higher-quality "platinum" labels. This may be further refined by fine-tuning on presuppositions benchmarks (CREPE, PGen), or leveraging abductive reasoning that seeks plausible premises to explain the observed evidence.

These experiments will inform the design of a "logical grounding" metric that quantifies missing atomic assumptions, accounts for multiple interpretations, and rejects invalid logical structures.

References

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 *ICLR*.

Shubhashis Roy Dipta and Francis Ferraro. 2025. If we may de-presuppose: Robustly verifying claims through presupposition-free question decomposition. In **SEM*, pages 253–266.

Najoung Kim, Ellie Pavlick, Burcu Karagol-Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *ACL*, pages 3932–3945.

Kristýna Onderková, Ondřej Plátek, Zdeněk Kasner, and Ondřej Dušek. 2025. Freshtab: Sourcing fresh data for table-to-text generation evaluation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 108–121.

Xin Quan, Marco Valentino, Danilo Carvalho, Dhairya Dalal, and André Freitas. 2025. Peirce: Unifying material and formal reasoning via llm-driven neuro-symbolic refinement. In *ACL Demos*, pages 11–21.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *NeurIPS*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80:101482.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. Do large language model benchmarks test reliability? *Preprint*, arXiv:2502.03461.

Yi Zhou and Basel Shbita. 2026. Evaluating ill-defined tasks in large language models. *Preprint*, arXiv:2603.17067.

A Appendix

error type	table id	title	insight
Correct	2-1458666-4	golf	*Australia and England have same number of wins.*
	Query: *cell(R1, 'nation', 'australia'), cell(R1, 'wins', V1), cell(R2, 'nation', 'england'), cell(R2, 'wins', V2), V1 = V2.* Result: 'R1': 1, 'V1': '5', 'R2': 2, 'V2': '5'		
Incorrect	2-170969-2	charlotte county	*Clarendon has the smallest population of 72.*
	The population is actually 71. Querry *cell(R1, 'population', '72')* is returned empty (no such entity).		
Assumption	2-1228353-1.html.csv	Felice Bonetto	*Felice drove 2 different car.*
	He drove 3 differnt cars: maserati, alfa romeo 159a, alfa romeo 159 m. The claim holds if we assume: *alias(alfa romeo 159a, alfa romeo 159 m)*		

Table 1: Examples from the LogicNLG data with Prolog queries and solutions. Shortened for illustrative purposes.