FreshTab: Sourcing Fresh Data for Table-to-Text Generation Evaluation

Kristýna Onderková, Ondřej Plátek, Zdeněk Kasner and **Ondřej Dušek**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czechia
{onderkova,oplatek,kasner,odusek}@ufal.mff.cuni.cz

Abstract

Table-to-text generation (insight generation from tables) is a challenging task that requires precision in analyzing the data. In addition, the evaluation of existing benchmarks is affected by contamination of Large Language Model (LLM) training data as well as domain imbalance. We introduce *FreshTab*, an on-the-fly table-to-text benchmark generation from Wikipedia, to combat the LLM data contamination problem and enable domain-sensitive evaluation. While non-English table-to-text datasets are limited, *FreshTab* collects datasets in different languages on demand (we experiment with German, Russian and French in addition to English). We find that insights generated by LLMs from recent tables collected by our method appear clearly worse by automatic metrics, but this does not translate into LLM and human evaluations. Domain effects are visible in all evaluations, showing that a domain-balanced benchmark is more challenging.

1 Introduction

Table-to-text generation or insight generation (Liu et al., 2018; Parikh et al., 2020) is a challenging task in natural language generation (NLG), where a NLG system generates insights from a data table. This can provide important support in data analytics and decision making in business or governance. Recent research in insight generation builds on finetuned neural language models (Nan et al., 2022; Zhao et al., 2023a; Kantharaj et al., 2022) or prompted large language models (LLMs) (Zhao et al., 2023b; Bian et al., 2024).

LLMs display excellent performance in various tasks, and unlike prior methods, they do not require costly in-domain training data with human-written references. With few-shot examples and chain-of-thought prompting, they surpass prior methods on insight generation (Zhao et al., 2023b). However, LLMs were also shown to memorize common benchmarks (Oren et al., 2024; Xu et al., 2024), inflating their true performance, and to perform unevenly across domains (Hu et al., 2024; Diao et al., 2025; Zhu et al., 2025).

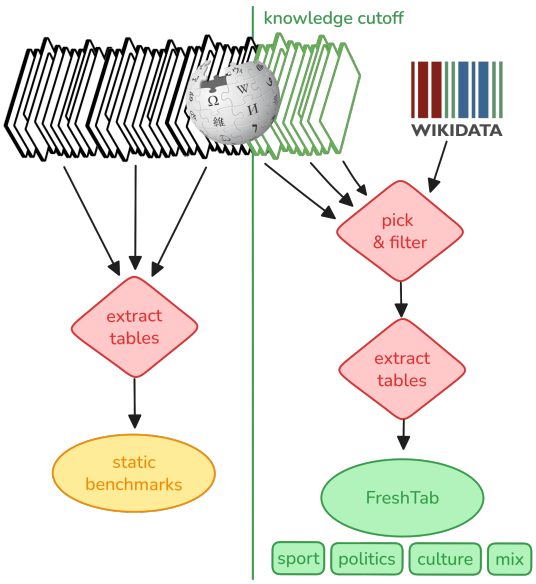


Figure 1: Schema of the FreshTab method

We directly address these problems and present *FreshTab*, an approach for obtaining up-to-date benchmarks for insight generation, following prior work on dynamic dataset construction (Kasner and Dusek, 2024; White et al., 2024). This dataset family, based on Wikipedia tables, is not affected by the problems of LLM memorization and benchmark contamination, as the underlying tables are newer than the LLM's knowledge cutoff date, see Figure 1. We introduce basic domain labels for each table, allowing for domain-specific evaluation insights. The datasets can be generated in any Wikipedia language and configured along multiple parameters.

Our main contributions are as follows:

- We develop *FreshTab* – a method for creating new table-to-text benchmark datasets based on

Proceedings of the 18th International Natural Language Generation Conference, pages 108–121
October 29–November 2, 2025.

recent Wikidata/Wikpedia entries, to avoid LLM memorization. The approach works for any language where a sufficient amount of fresh data is available.

- We include domain information in the process, to allow for domain-specific evaluation.
- In experiments using February-May 2025 tables collected with FreshTab, we show that recent LLMs perform worse than on comparable tables from the earlier LoTNLG/LogicNLG benchmark (Zhao et al., 2023b; Chen et al., 2020) based on automatic metrics. However, this effect is less pronounced in LLM evaluation and absent in human evaluation, indicating a potential metric bias. We show that domain-balanced data are more challenging than the sport-heavy data used by the previous benchmarks. A LLM evaluation of insights for Russian, German and French tables shows similar performance to English.

FreshTab is publicly available and automatically collects a new dataset version each month.[1]

2 Related work

Insight generation Approaches for generating insights from tables have been developed alongside other data-to-text NLG systems for decades (Barzilay and Lapata, 2005). The emergence of neural models brought a lot of research into the area, focusing on end-to-end architectures (Wiseman et al., 2017) that incorporate table-aware training (Liu et al., 2018; Xing and Wan, 2021), use pretrained LMs (Kantharaj et al., 2022), or both (Chen et al., 2020; Andrejczuk et al., 2022). Most recent approaches to table-to-text use LLMs. While Bian et al. (2024) and Li et al. (2023) still focus on finetuning LLMs on tabular tasks, Zhao et al. (2023b) and Pérez et al. (2025) successfully apply chain-of-thought prompting without the need for task-specific training. However, all previous table-to-text approaches focus on fixed benchmarks, making them susceptible to training data contamination (Jacovi et al., 2023; Li and Flanigan, 2024; Oren et al., 2024).

Dynamic benchmarks To counteract the issues of LLM training data contamination, Axelsson and Skantze (2023) propose modifying benchmarks using counter-factual or fictional entities. This partially solves the issue, but the resulting synthetic data are not realistic, and a potential for a repeated leakage remains (hence the non-public release of GEM 2024 test data; Mille et al., 2024). To remove this limitation, dynamic benchmarks emerged recently: White et al. (2024)'s LiveBench represents a set of general questions or problems for LLMs to solve, updated regularly in a manual fashion. Kasner and Dusek (2024) focus specifically on the data-to-text generation task, using open APIs to automatically gather fresh input data in several domains. Our work extends these approaches for the table insight generation task using automatic selection of recent tables from Wikipedia. Furthermore, it adds domain-sensitive evaluation, following (Zhu et al., 2025).

3 Methodology

3.1 Benchmark Format

Unlike previous benchmarks using Wikipedia tables (Chen et al., 2020; Zhao et al., 2023b), our benchmark only includes Wikipedia data tables with no human reference texts as obtaining references on-the-fly is not feasible. Instead, we use reference-free evaluation metrics and human evaluation, following Kasner and Dusek (2024).

In addition to the tables themselves, we include *domain labels*, indicating a broad thematic area (*sport*, *politics*, *culture* or *other*) for each table. Following the *LoTNLG* benchmark, we also include a set of five *logical operation labels* (a subset of *aggregation, all, comparative, count, negation, ordinal, simple, superlative, unique*, see Appendix D), to provide a suggestion for the model on the type of insight to generate.[2]

3.2 Benchmark Production Process

Wikipedia has about 64 million pages,[3] making it non-trivial to identify pages which contain tables added after a specific date. Therefore, we identify a relevant subset of pages heuristically. Our approach proceeds in the following steps:

1. We query Wikidata using SPARQL queries with a handpicked set of concepts and categories, to obtain a list of Wikipedia pages appropriate for scraping. This is done with two distinct multi-step approaches. We follow two strategies for determining if a page is truly new, checking for: (1) pages on events taking place between the

[1] https://github.com/Kristyna-Navitas/FreshTab

[2] Unlike in *LoTNLG* where they were based on references, the logical labels are sampled randomly in *FreshTab*.

[3] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

cutoff date and the present and (2) pages that were newly created after a cutoff date.

2. We scrape these pages for tables, clean them and pick one table per page, based on a pre-set targets on table size in terms of number of rows and columns, as well as non-empty cells.
3. We filter the resulting pages based on configurable domain balance. Each table is also assigned five random logical operations.

The benchmark generation is fully configurable via YAML; more details on the individual steps are included in Appendix A.

4 Experimental Setup

4.1 Benchmark Comparison

To evaluate the usefulness of our method, we compare it to the previous *LoTNLG* benchmark (Zhao et al., 2023b), a subset of the commonly used *LogicNLG* data (Chen et al., 2020), which was available to all LLMs at training time, and is paired with reference insights. Using *FreshTab*, we created several new benchmarks:

- *FreshTab.2-5/25.en.lot* from February-May 2025, after the knowledge cutoff dates for the most recent LLMs. It has 100 English tables with the same domain distribution as the *LoTNLG* benchmark (73 sport, 13 other, 11 culture, and 3 politics tables), to compare the effect of using new data.
- *FreshTab.2-5/25.en.diverse* contains 200 English tables, evenly distributed across the four domains, to evaluate domain-specific performance.
- *FreshTab.2-5/25* variations in six other languages with the most articles on Wikipedia,[4] to assess feasibility of producing non-English datasets.

We set the table size limit to approx. 3k characters, so that all tables comfortably fit into LLMs' context sizes. The table parameters were taken from the *LogicNLG* (Chen et al., 2020) benchmark tables.

4.2 Models Evaluated

We evaluate a broad range of open models for insight generation on both *LoTNLG* and our English *FreshTab.2-5/25.en.{lot/diverse}* data: *Llama 3.3 70B* (Grattafiori et al., 2024), *Qwen 2.5 72B* (Qwen et al., 2025), *Mistral Small 3 24B*[5], *Gemma 3 27B* (Team et al., 2025), and reasoning models *Magistral* (Rastogi et al., 2025) and *DeepSeek R1 Distill Llama 70B* (DeepSeek-AI, 2025) All generations use a temperature of 0.7, in line with Zhao et al. (2023b). We use all models through *Ollama*[6] with 8-bit quantization, to balance our hardware constraints and performance losses due to quantization (Marchisio et al., 2024). We use structured outputs, i.e., constrain the LLM generation to a predefined schema.[7]

4.3 Prompting setups

Following *LoTNLG* (Zhao et al., 2023b), we run two LLM chain-of-thought prompting setups:

- **Direct CoT**. The LLM is given the table and description of one logical operation and asked to generate one insight. This runs five times per table for five logical operations.
- **Choice**. The LLM is given the table and descriptions of all nine logical operations and asked to generate five insights in one go, selecting operations as needed.

4.4 Human Evaluation

We run a crowdsourced human evaluation on a sample of our data (50 tables from each benchmark) with outputs from four LLMs: Llama, DeepSeek, Gemma and Qwen. We recruit annotators on the Prolific platform.[8] We ask the annotators to spot and highlight accuracy errors in the insights on the word level, following Kasner and Dusek (2024)'s setup. We operate with four error categories: *incorrect*, *not checkable*, *misleading*, and *other*. Details of error categories are explained in the annotation interface, shown in Appendix C.

4.5 Automatic Evaluation

We use the standard reference-free automatic metrics for the *LogicNLG* benchmark (Liu et al., 2022a; Zhao et al., 2023b) – trained table entailment metrics *TAPAS* (Herzig et al., 2020) and *TAPEX* (Liu et al., 2022b). We focus on TAPEX in the paper, as we consider output correctness crucial, and TAPEX is the more reliable of the two. TAPAS as well as scores for other generation aspects are given in Appendix B (self-BLEU (Zhu et al., 2018), unique tokens (Li et al., 2016) and Shannon entropy (van Miltenburg et al., 2018) to measure diversity, percentage of failures, and the average output lengths).

[4] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[5] https://mistral.ai/news/mistral-small-3
[6] https://ollama.com/
[7] https://ollama.com/blog/structured-outputs
[8] https://app.prolific.co/

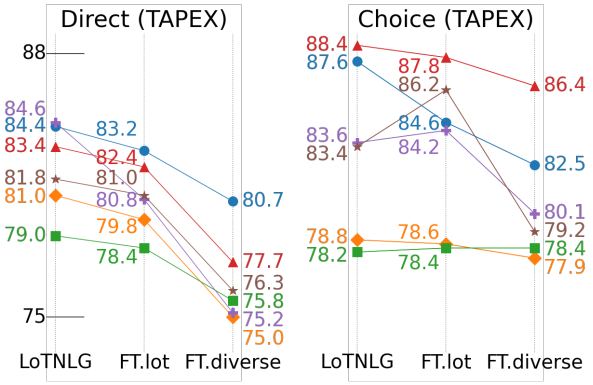


Figure 2: *TAPEX* on *LoTNLG* vs. *FreshTab.2-5/25.en.lot*vs. *FreshTab.2-5/25.en.diverse*

In addition, we ran an LLM-as-a-judge evaluation (Gu et al., 2024) with the Llama 3.3 70B model. We crafted the prompt to be as close as possible to the annotation instructions for human evaluators (see Section 4.4).

5 Results

5.1 TAPEX Performance

Based on TAPEX scores in Figure 2, our *FreshTab* benchmark shows more challenging than *LoTNLG* for both prompting setups and most models, especially in the *diverse* domain distribution. The *diverse* data proves particularly hard for the DeepSeek and Magistral reasoning LLMs, where the chain-of-thought runs into a dead end and does not produce a valid output in 5%-10% cases.

For *Direct CoT*, the performance drop on *FreshTab* is statistically significant for most examined LLMs ($p \leq 0.05$, Z-test for proportions, see Table 4 in Appendix B), with the domain change (*lot* vs. *diverse*) having a stronger effect than the freshness of the tables. The *Choice* experiment consistently outperforms *Direct CoT*, showing that giving the model more freedom in choosing logical operations pays off. Performance drop on new data is statistically significant for Llama and Magistral.

5.2 LLM-as-a-judge Evaluation

Based on the LLM-judge evaluation in Figure 3, the performance drop on new data is not as straightforward. The scores are lower overall and more varied; few differences are statistically significant (Gemma for *Direct*, DeepSeek for *Choice*). In *Direct*, we often see a performance increase on *FreshTab.2-5/25.en.lot* but a subsequent drop on the *diverse* set. We attribute this to the domain balance.

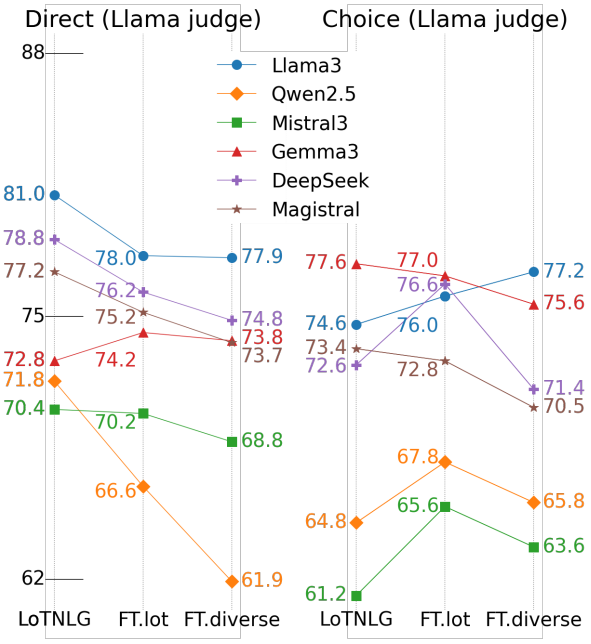


Figure 3: *Llama-as-a-judge* on *LoTNLG* vs. *FreshTab.2-5/25.en.lot* vs. *FreshTab.2-5/25.en.diverse*

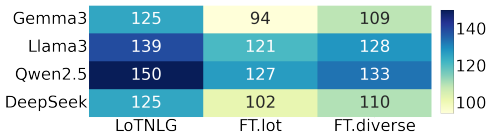


Figure 4: Total number of errors found in human evaluation by model and benchmark

The scores for *Choice* and *Direct* are mostly similar. Differences are probably influenced by logical operation choice – operations picked by LLMs in *Choice* are often different from the ones pre-picked by humans in *Direct* (cf. Figure 7 in the Appendix). Overall, all LLMs except Qwen tend to produce *simple* insights more frequently, and Gemma is the most extreme in this regard, gaining higher scores overall.

5.3 Human Evaluation

Figure 4 shows an overview of our human annotation results (see Table 7 in Appendix B for details). They align better with LLM evaluation than with TAPEX/TAPAS and show an even more consistent trend – the number of errors does not increase on the new data; on the contrary, *FreshTab.2-5/25.en.lot* shows fewer errors overall; the effect is similar in all evaluated LLMs. The drop on the *diverse* set of *FreshTab* compared to the *lot* set is also clearly visible.

The evaluation differences directly translate to correlations: TAPAS and TAPEX show only low

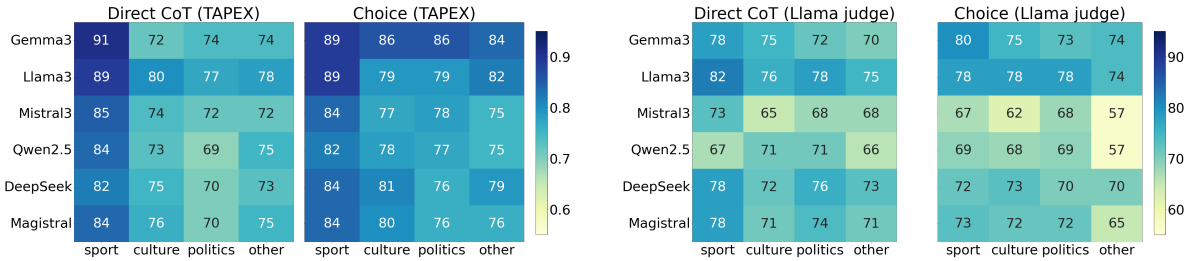


Figure 5: *TAPEX* and *Llama as a judge* on *FreshTab.2-5/25.en.diverse* by domain.

language	en	de	fr	sv	nl	ru	es
total	531	187	177	54	144	106	159
sport	204	86	73	21	94	34	35
politics	142	27	61	1	4	9	19
culture	109	51	25	3	13	29	87
other	48	23	18	29	33	34	18

Table 1: Count of pages with new tables for language variations of *FreshTab.2-5/25* (4 months period).

Pearson correlation with humans (0.12 and 0.11). Based on manual inspection, TAPEX performs better on simple logical operations than on more complex ones. The LLM-as-a-judge with Llama 3.3 70B produces a moderate correlation of 0.53 across models and datasets. We compared all other LLMs in the judge setting on the LotNLG set (see Table 9 in Appendix B); Llama shows as second highest-correlating but without self-bias.

When we analyze the outputs more closely, we can see that the lower number of errors on *FreshTab* is partly due to logical operation choice. On *LotNLG*, models produce more complex insights (e.g. "3 episodes have ratings above 16%.") by using seen patterns. On *FreshTab* data, they play it safer and produce simpler insights (e.g. "France is in qualifying group D."), leading to fewer *aggregation/superlative* insights and thus fewer errors. Errors on *LoTNLG* often concern exact values. With *FreshTab*, models also misinterpret tables (e.g., "Bird [won the most awards] among all the films at Sudbury Film Festival" while the table only lists awards for the Bird movie), column labels, subtables, row/column switches, or unusual formats (e.g. speech transcript). Numerical operations tend to be less accurate. Reasoning models produce empty outputs more frequently. The models also do not shy away from inconsistent claims, e.g., *"Myanmar has the second-highest number of missing persons, equal to the total across all countries affected by the earthquake"*.

5.4 Comparison of Domains

Figure 5 shows that TAPEX and LLM-judge performance varies across domains. With TAPEX, the difference between the *sport* domain and lowest performing domain is statistically significant for all models in the *Direct CoT* experiment and for all except Gemma and Qwen in *Choice* ($p \leq 0.05$, Z-test for proportions). For LLM-judge, the differences are only significant for Mistral and Qwen.

With models constrained by pre-set random logical operations in *Direct CoT*, we see *sport* performing mostly better than other domains. For *Choice*, TAPEX gets more even across domains as models can pick logical operations. LLM-judge reveals that only some models use the larger freedom favorably, with mixed gains and losses.

5.5 Other Languages

Table 1 demonstrates that usably-sized datasets, albeit smaller than English, can be produced in other popular Wikipedia languages using *FreshTab*. We generated insights for three other diverse but high-resource languages from *FreshTab.2-5/25.(de/fr/ru).diverse* and evaluated them with LLM-as-a-judge, as TAPAS and TAPEX cannot be used directly. The scores are mostly consistent across models; slightly lower for German, similar to English for Russian and slightly higher for French. However, this very much depends on the composition of the new data. Full results are in Table 8 in Appendix B.

6 Conclusion

We present *FreshTab*, a method for producing live benchmark datasets for table insight generation from Wikipedia, enabling easy evaluation of LLMs on unseen data and supporting domain balance and non-English languages. Our experiments confirm that LLMs behave differently on the new data. We also found poor performance of automatic metrics, with LLM-judges showing more reliable.

Limitations

We use fairly standard generation LLM parameters shared across all steps and consider our setup to be a reasonable baseline. We adopted the labels for our general ideas from (Chen et al., 2020; Zhao et al., 2023b) but the logical operation categorization is not complete or optimal. However, using nine diverse logical operations allowed us to have some degree of controllability and a known source of diversity. We acknowledge that the current prompting strategy could be refined and optimized, which we consider as future work. Some of the data novelty effect may have been compromised by new articles being only translated from another language. This was only discovered for a single example, but needs to be further evaluated.

Ethics Statement

The human annotation experiment was approved by an internal ethics committee. The annotators were awarded £9 for the annotation task which was estimated to take 60 minutes, in line with the Prolific platform's recommendation.

Acknowledgments

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 project SVV 260 698. It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2023062).

References

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.

Agnes Axelsson and Gabriel Skantze. 2023. Using large language models for zero-shot natural language generation from knowledge graphs. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 39–54, Prague, Czech Republic. 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.

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].

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*, pages 7929–7942, Online. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, and Zhuosheng Zhang. 2025. GuideBench: Benchmarking Domain-Oriented Guideline Following for LLM Agents. In *ACL*, Vienna. arXiv. ArXiv:2505.11368 [cs].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. 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.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA. Association for Computational Linguistics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. 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.

Zdeněk Kasner and Ondrej Dusek. 2024. 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)*, pages 12045–12072, Bangkok, Thailand. Association for Computational Linguistics.

Zdeněk Kasner, Ondrej Platek, Patricia Schmidtova, Simone Balloccu, and Ondrej Dusek. 2024. factgenie: A Framework for Span-based Evaluation of Generated Texts. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 13–15, Tokyo, Japan. Association for Computational Linguistics.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18471–18480.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2023. Table-GPT: Table-tuned GPT for Diverse Table Tasks. ArXiv:2310.09263 [cs].

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 *International Conference on Learning Representations*.

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 AAAI Conference on Artificial Intelligence*, volume 32. Number: 1.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual llms? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947.

Simon Mille, João Sedoc, Yixin Liu, Elizabeth Clark, Agnes Johanna Axelsson, Miruna Adriana Clinciu, Yufang Hou, Saad Mahamood, Ishmael Nyunya Obonyo, and Lining Zhang. 2024. The 2024 GEM shared task on multilingual data-to-text generation and summarization: Overview and preliminary results. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 17–38, Tokyo, Japan. Association for Computational Linguistics.

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.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving Test Set Contamination in Black-Box Language Models. In *ICLR*, Vienna, Austria.

Ankur 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)*, pages 1173–1186, Online. Association for Computational 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 (Volume 1: Long Papers)*, pages 562–582, Albuquerque, New Mexico. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khy-

athi Raghavi Chandu, et al. 2025. Magistral. *arXiv preprint arXiv:2506.10910*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, and Morgane Rivière et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ronald E Walpole, Raymond H Myers, Sharon L Myers, and Keying E Ye. 2010. *Probability and statistics for engineers and scientists*, 9 edition. Pearson, Upper Saddle River, NJ.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv: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.

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.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

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.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26148–26156. Number: 24.

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.

A Data collection details

Further details for the data collection steps:

Data choice (Step 1). We choose the set of concepts and categories by exploration to cover the types of pages that tend to include tables. The tables are picked so that their contents could not have been known before the cutoff date since the page was either non-existent then, or it covers an event (e.g., election, sports competition, book release) that only took place after the cutoff date, and thus its specifics could not have been known before. We check the wikipage's first creation date, to avoid updated entities. We also abstain from getting largely empty tables relating to future events.

Table selection (Step 2). The table selection includes removing noisy, small, and mostly empty tables based on configurable thresholds. The cleaning step shortens very long tables, simplifies multicolumn names, removes references, consolidates non-values, removes unreasonably long text entries, and empty columns and rows.

B Full results

The following tables show our experiments in full: Table 2 for the *Direct CoT* experiment and Table 3 for the *Choice* experiment. P-values for the Z-test for proportions (Walpole et al., 2010) on the TAPEX metric between the individual benchmarks are given in Table 4. Table 5 shows the TAPEX metric separately for each logical operation. In addition to the TAPEX metric (Liu et al., 2022b) reported in the main paper, we report TAPAS (Herzig et al., 2020) in Figure 6. Note that *TAPEX* treats empty results as not-entailed, as opposed to the *TAPAS* metric that treats these as correct.

To measure insights' diversity, we further report self-BLEU, i.e., BLEU when comparing insights against each other (Zhu et al., 2018). Lower self-BLEU means greater diversity. A further measures of diversity are the average number of unique tokens per insight and Shannon entropy (van Miltenburg et al., 2018). We also measure the percentage of empty/failed outputs and the average length of the produced insights (in characters).

The full results of human annotation are in Table 6 showing the percentage of incorrect and misleading insights together, as annotators sometimes used them interchangeably (with gray experiments having too low count to be statistically significant); and

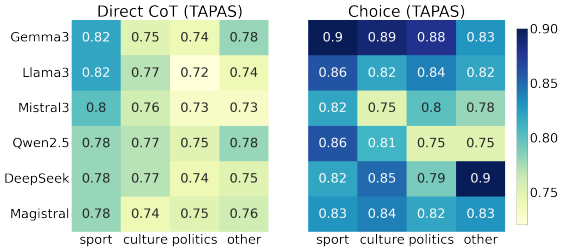


Figure 6: *TAPAS* by domain on *FreshTab.2-5/25.en.diverse*

Table 7 showing the actual counts for all annotated categories.

Figure 7 depicts the counts for specific logical operations picked by the different models in the *Choice* experiment related to the human-picked operations in the LoTNLG dataset (horizontal line).

The complete results for all languages tested are in Table 8. Pearson correlations of different LLM judges with human annotations are in Table 9, where the reasoning models were tested only on one set of data due to the high number of tokens generated and not showing a better correlation for it.

C Human Annotation Details

The examples for the human annotation are sampled randomly while excluding tables with over 120 characters in the header, to fit into the annotation interface without horizontal scrolling.

We use the *Factgenie* annotation tool (Kasner et al., 2024). Each annotator is given 3 tables, each paired with 21 insights – five insights per evaluated model, plus one table-unrelated insight used as an attention check.[9]

Detailed annotation instructions, as shown to the annotators prior to annotation, are given in Figure 8. The annotation interface is shown in Figure 9. Annotators were pre-selected based on their country of residence (UK, U.S., Ireland, Australia, New Zealand), their indicated primary language (English) and good approval rate. We manually checked whether annotators gave meaningful replies to the attention check instances, and if not, their annotations were replaced by an additionally hired annotator.

[9]We sample the attention check insights from insights related to different input tables.

model	empty	TAPAS	TAPEX	self-BLEU4	unique tokens	avg len	entropy
			LoTNLG benchmark				
Gemma 3	0.00	89.8	83.4	0.64	36	86	4.71
Llama 3.3	0.00	87.0	84.4	0.56	42	95	5.23
Mistral	0.00	75.4	79.0	0.28	48	88	5.11
Qwen 2.5	0.01	86.0	81.0	0.54	45	102	5.18
DeepSeek	0.01	85.8	84.6	0.50	39	85	4.76
Magistral	0.02	82.4	81.8	0.41	42	81	4.71
			FreshTab.2-5/25.en benchmark				
Gemma 3	0.00	77.4	82.4	0.39	46	98	5.38
Llama 3.3	0.00	76.6	83.2	0.34	49	99	5.31
Mistral	0.00	81.4	78.4	0.13	49	77	5.33
Qwen 2.5	0.00	78.4	79.8	0.35	53	111	5.62
DeepSeek	0.03	78.2	80.8	0.33	45	89	5.33
Magistral	0.01	79.4	81.0	0.27	48	86	5.50
			FreshTab.2-5/25.en.diverse benchmark				
Gemma 3	0.01	77.3	77.7	0.36	47	101	4.73
Llama 3.3	0.00	76.3	80.7	0.30	51	105	5.21
Mistral	0.00	75.7	75.8	0.25	49	90	5.13
Qwen 2.5	0.01	77.2	75.0	0.30	56	115	5.55
DeepSeek	0.07	75.9	75.2	0.29	47	90	4.69
Magistral	0.05	75.5	76.3	0.24	48	89	4.95

Table 2: Automatic metrics for the *Direct CoT* experiment

model	empty	TAPAS	TAPEX	self-BLEU4	unique tokens	avg len	entropy
			LoTNLG benchmark				
Gemma 3	0.00	87.2	88.4	0.15	52	88	5.35
Llama 3.3	0.00	88.8	87.6	0.18	61	110	5.57
Mistral	0.00	81.8	78.2	0.13	52	82	5.42
Qwen 2.5	0.00	80.0	78.8	0.17	62	103	5.61
DeepSeek	0.01	83.2	83.0	0.14	51	81	5.29
Magistral	0.05	82.2	83.4	0.16	51	79	5.44
			FreshTab.2-5/25.en benchmark				
Gemma 3	0.00	87.0	87.8	0.16	50	83	5.49
Llama 3.3	0.00	83.4	84.6	0.20	60	109	5.89
Mistral	0.00	81.4	78.4	0.13	49	77	5.33
Qwen 2.5	0.00	82.4	78.6	0.19	60	101	5.65
DeepSeek	0.06	83.4	84.2	0.17	47	73	4.23
Magistral	0.03	87.6	86.2	0.16	49	78	5.72
			FreshTab.2-5/25.en.diverse benchmark				
Gemma 3	0.00	87.6	86.4	0.16	53	92	5.41
Llama 3.3	0.00	83.5	82.5	0.18	63	115	5.70
Mistral	0.00	78.9	78.4	0.12	53	87	5.48
Qwen 2.5	0.00	79.5	77.9	0.18	61	106	5.87
DeepSeek	0.10	81.5	80.1	0.16	49	79	6.01
Magistral	0.11	83.0	79.2	0.17	50	80	5.35

Table 3: Automatic metrics for the *Choice CoT* experiment

model	*LoTNLG* vs *Diverse*	*LoTNLG* vs *FreshTab*	*FreshTab* vs *Diverse*
	Direct CoT experiment		
Gemma 3	0.01	0.67	0.03
Llama 3.3	0.08	0.61	0.24
Mistral	0.17	0.82	0.26
Qwen 2.5	0.01	0.63	0.04
DeepSeek	0.00	0.11	0.02
Magistral	0.02	0.75	0.04
	Choice CoT experiment		
Gemma 3	0.28	0.77	0.45
Llama 3.3	0.01	0.17	0.31
Mistral	0.93	0.91	1.00
Qwen 2.5	0.69	0.94	0.76
DeepSeek	0.18	0.61	0.05
Magistral	0.05	0.22	0.00

Table 4: Statistical significance between datasets

model	aggregation	all	comparative	count	negation	ordinal	simple	superlative	unique
				LoTNLG benchmark					
Gemma	80.0	77.8	81.4	71.6	60.7	85.9	87.8	87.1	84.5
Llama	83.3	77.8	79.4	77.3	75.0	89.1	95.1	97.6	77.6
Mistral	86.7	88.9	88.7	84.1	71.4	78.1	95.1	88.2	75.9
Qwen 2.5	90.0	88.9	82.5	86.4	64.3	82.8	85.4	87.1	58.6
DeepSeek	90.0	55.6	78.4	88.6	67.9	89.1	92.7	95.3	72.4
Magistral	89.7	77.8	88.5	77.9	60.7	79.4	95.1	91.6	70.7
				FreshTab.2-5/25.en benchmark					
Gemma 3	89.2	66.7	82.0	88.7	54.2	96.8	86.2	80.9	93.1
Llama 3.3	92.3	75.0	82.0	81.1	59.3	91.9	91.4	89.4	84.5
Mistral	90.8	43.8	76.0	86.8	55.9	83.9	94.8	89.4	84.5
Qwen 2.5	83.1	60.4	82.0	86.8	55.9	91.9	94.8	87.2	74.1
DeepSeek	93.8	64.6	86.0	78.8	66.1	93.5	84.5	84.8	74.1
Magistral	87.5	72.9	84.0	73.1	61.0	93.4	91.4	87.2	81.0
				FreshTab.2-5/25.en.diverse benchmark					
Gemma 3	86.4	68.3	82.2	74.8	49.6	90.4	81.9	85.1	80.9
Llama 3.3	87.3	74.0	85.0	75.7	62.2	89.6	88.8	80.2	82.6
Mistral	84.5	54.8	78.5	77.5	47.9	78.4	92.2	84.2	85.2
Qwen 2.5	79.1	54.8	82.2	76.7	58.0	80.0	86.2	92.1	67.0
DeepSeek	87.3	60.6	78.5	71.8	63.0	90.2	71.6	88.0	67.8
Magistral	79.4	71.2	82.1	76.0	55.5	81.3	88.7	84.2	74.6

Table 5: TAPEX for logical operations for *Direct* CoT experiment.

model	sport	culture	politics	other
		LoTNLG benchmark		
counts	190	20	15	25
Gemma 3	0.35	0.25	0.20	0.48
Llama 3.3	0.44	0.30	0.40	0.44
Qwen 2.5	0.46	0.45	0.27	0.28
DeepSeek	0.35	0.45	0.47	0.36
		FreshTab.2-5/25.en benchmark		
counts	160	50	5	35
Gemma 3	0.29	0.24	0.40	0.17
Llama 3.3	0.39	0.26	0.80	0.17
Qwen 2.5	0.38	0.36	0.60	0.26
DeepSeek	0.24	0.28	0.40	0.11
		FreshTab.2-5/25.en.diverse benchmark		
counts	55	75	65	55
Gemma 3	0.36	0.25	0.22	0.40
Llama 3.3	0.53	0.24	0.35	0.29
Qwen 2.5	0.49	0.25	0.37	0.45
DeepSeek	0.36	0.28	0.32	0.31

Table 6: Percentage of incorrect+misleading insights from human annotation by domains.

model	Incorrect	Misleading	Not checkable	Other
LoTNLG benchmark				
Gemma 3	67	20	20	18
Llama 3.3	83	23	13	20
Qwen 2.5	82	26	21	21
DeepSeek	68	23	15	19
FreshTab.2-5/25.en benchmark				
Gemma 3	49	18	10	17
Llama 3.3	58	28	17	18
Qwen 2.5	73	17	23	14
DeepSeek	46	13	21	22
FreshTab.2-5/25.en.diverse benchmark				
Gemma 3	59	16	11	23
Llama 3.3	64	22	18	24
Qwen 2.5	76	19	12	26
DeepSeek	70	9	14	17

Table 7: Factuality span annotations prevalences from human annotation.

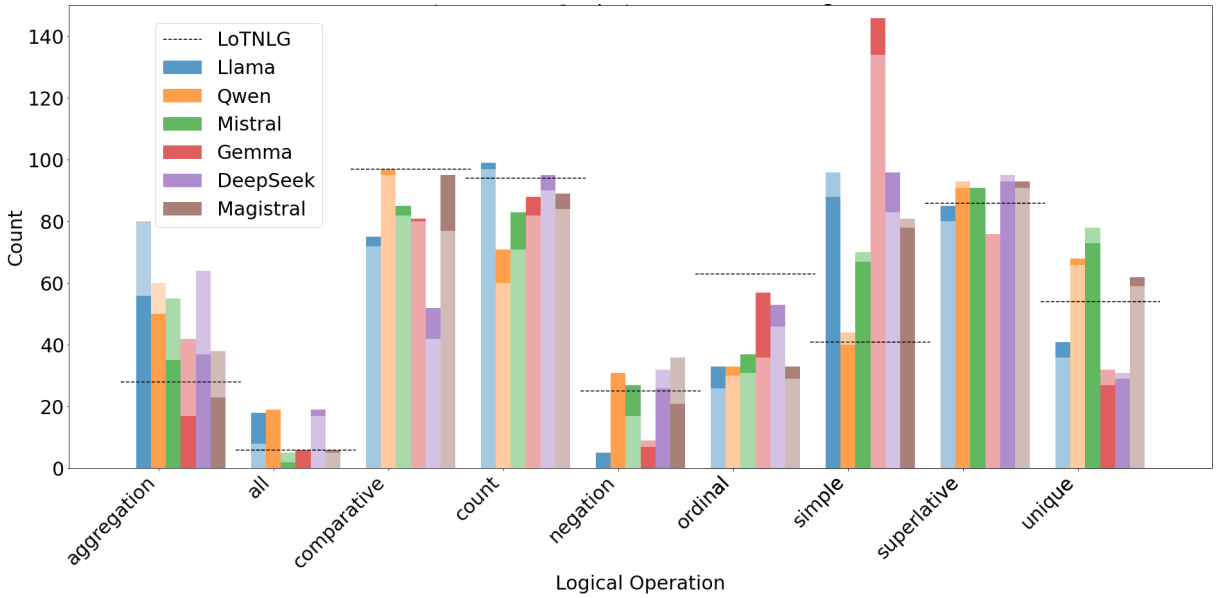
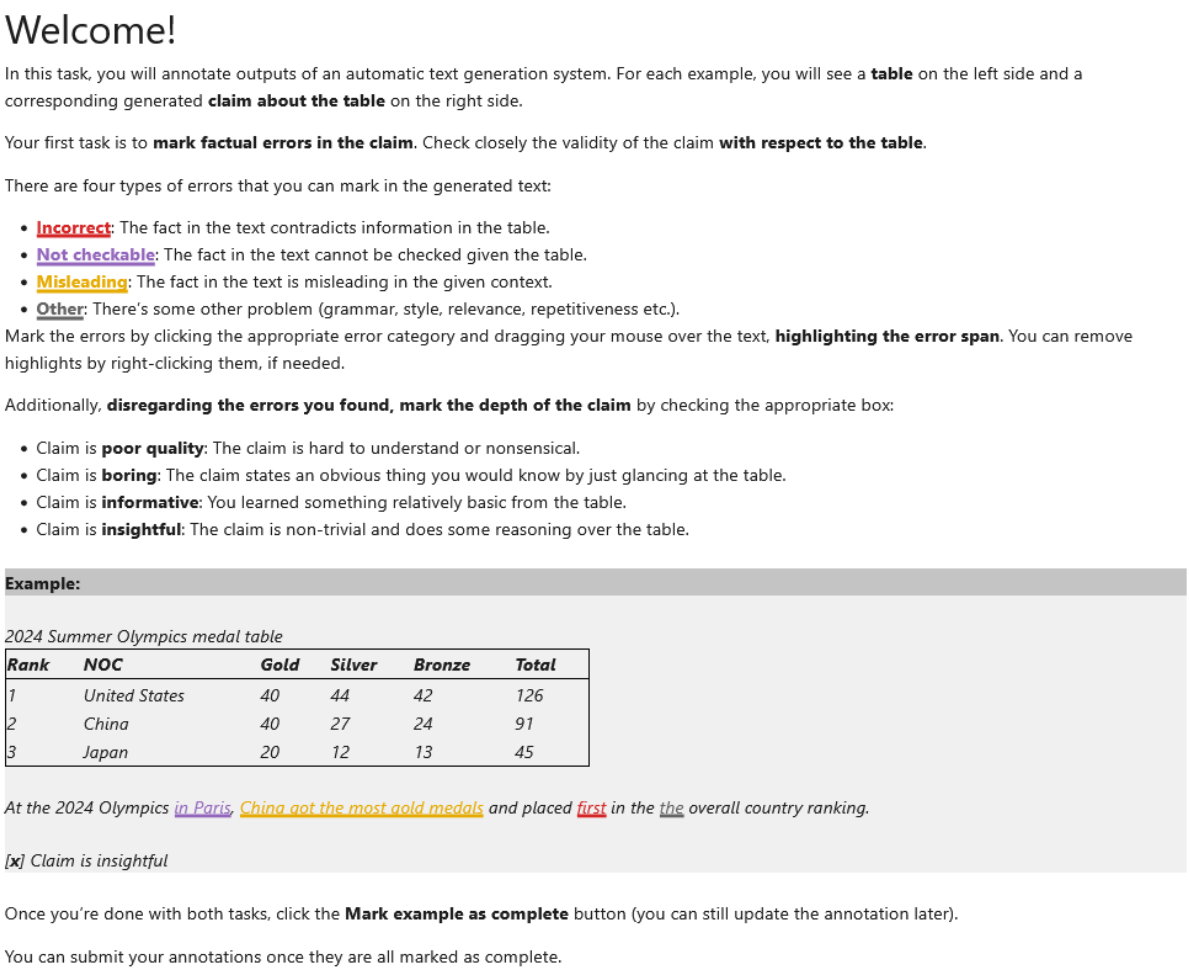


Figure 7: Comparison of logical operation counts. Given from *LoTNLG* (line) and chosen in *Choice* experiment for *LoTNLG* (light) and *FreshTab* (saturated)

D Types of logical inferences

We use the following nine logical operations, proposed by Zhao et al. (2023b):

- *aggregation* – insights that mention aggregate statistics of data such as sums or averages, e.g., average home team score
- *all* – insights where all items share a common property, e.g., all games were played on the same date
- *comparative* – insights that compare different entities on some property, e.g., comparing the scores of two teams
- *count* – knowledge about the number of entities that fulfill some condition, e.g., number of teams that played at a particular venue
- *negation* – formulates a negative claim about an entity, e.g., Team A never played against Team B
- *ordinal* – indicates the ranking of entities on some aspect, e.g., second largest crowd to watch the match at a venue
- *simple* – the sentences which do not involve higher-order operations, e.g., Player X is from country Y.
- *superlative* – data insights about maximum or minimum values, e.g., highest score by any team
- *unique* – insights about distinct values of a column, e.g., the matches were played in different venues



Welcome!

In this task, you will annotate outputs of an automatic text generation system. For each example, you will see a **table** on the left side and a corresponding generated **claim about the table** on the right side.

Your first task is to **mark factual errors in the claim**. Check closely the validity of the claim **with respect to the table**.

There are four types of errors that you can mark in the generated text:

- **Incorrect**: The fact in the text contradicts information in the table.
- **Not checkable**: The fact in the text cannot be checked given the table.
- **Misleading**: The fact in the text is misleading in the given context.
- **Other**: There's some other problem (grammar, style, relevance, repetitiveness etc.).

Mark the errors by clicking the appropriate error category and dragging your mouse over the text, **highlighting the error span**. You can remove highlights by right-clicking them, if needed.

Additionally, **disregarding the errors you found, mark the depth of the claim** by checking the appropriate box:

- Claim is **poor quality**: The claim is hard to understand or nonsensical.
- Claim is **boring**: The claim states an obvious thing you would know by just glancing at the table.
- Claim is **informative**: You learned something relatively basic from the table.
- Claim is **insightful**: The claim is non-trivial and does some reasoning over the table.

Example:

2024 Summer Olympics medal table

Rank	NOC	Gold	Silver	Bronze	Total
1	United States	40	44	42	126
2	China	40	27	24	91
3	Japan	20	12	13	45

At the 2024 Olympics in Paris, China got the most gold medals and placed first in the the overall country ranking.

[x] Claim is insightful

Once you're done with both tasks, click the **Mark example as complete** button (you can still update the annotation later).

You can submit your annotations once they are all marked as complete.

Figure 8: Annotation instructions for the human evaluation campaign.

	en	de	fr	ru
Choice experiment				
Gemma	75.6	74.4	79.6	83.2
Llama	77.2	71.4	76.8	76.2
Mistral	63.6	61.4	71.2	68.2
Qwen	65.8	66.2	69.4	67.4
DeepSeek	71.5	56.0	69.6	62.0
Magistral	70.5	72.0	69.8	64.4
Direct experiment				
Gemma	73.8	74.0	80.8	72.8
Llama	77.9	76.8	81.6	78.8
Mistral	68.8	69.0	70.4	70.0
Qwen	69.1	70.8	74.0	72.8
DeepSeek	74.8	71.8	78.8	76.0
Magistral	73.7	70.6	75.0	75.0

Table 8: Factuality of generations for selected languages with Llama-as-a-judge for *FreshTab.2-5/25.diverse*.

judge / insight	Gemma	Llama	Qwen	DeepSeek
Gemma	0.82	0.52	0.61	0.43
Llama	0.46	0.56	0.61	0.54
Qwen	0.47	0.40	0.42	0.47
Mistral	0.34	0.41	0.33	0.44
DeepSeek	-	0.46	-	-
Magistral	-	0.20	-	-

Table 9: Pearson correlations between LLM-as-a-judge with different LLMs and human evaluation for LoTNLG dataset.

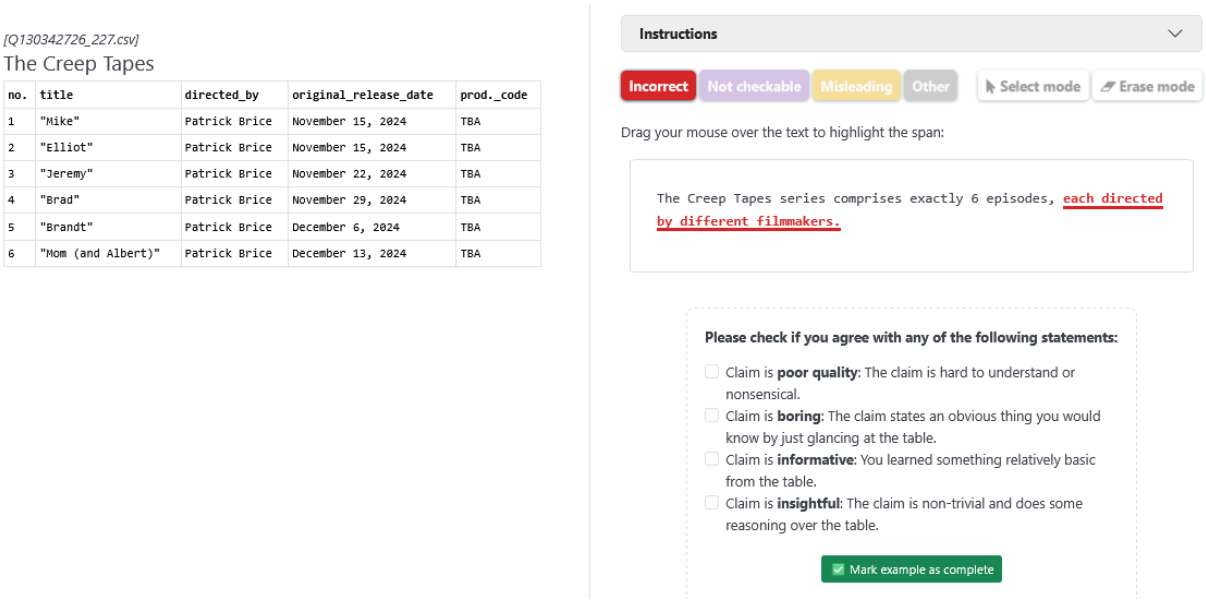


Figure 9: Annotation interface, with the table on the left and the annotation form on the right. Annotators can display the instructions by clicking on the top-right collapsible panel.