eC-Tab2Text: Aspect-Based Text Generation from e-Commerce Product Tables

Luis Antonio Gutiérrez Guanilo[♣] **Mir Tafseer Nayeem**[♦*]
Cristian López[♣] **Davood Rafiei**[♦*]
[♣]University of Engineering and Technology (UTEC) [♦]University of Alberta
{mnayeem, drafiei}@ualberta.ca {luis.gutierrez.g, clopezd}@utec.edu.pe

Abstract

Large Language Models (LLMs) have demonstrated exceptional versatility across diverse domains, yet their application in e-commerce remains underexplored due to a lack of domain-specific datasets. To address this gap, we introduce **eC-Tab2Text**, a novel dataset designed to capture the intricacies of e-commerce, including detailed product attributes and user-specific queries. Leveraging eC-Tab2Text, we focus on text generation from product tables, enabling LLMs to produce high-quality, attribute-specific product reviews from structured tabular data. Fine-tuned models were rigorously evaluated using standard Table2Text metrics, alongside correctness, faithfulness, and fluency assessments. Our results demonstrate substantial improvements in generating contextually accurate reviews, highlighting the transformative potential of tailored datasets and fine-tuning methodologies in optimizing e-commerce workflows. This work highlights the potential of LLMs in e-commerce workflows and the essential role of domain-specific datasets in tailoring them to industry-specific challenges[1].

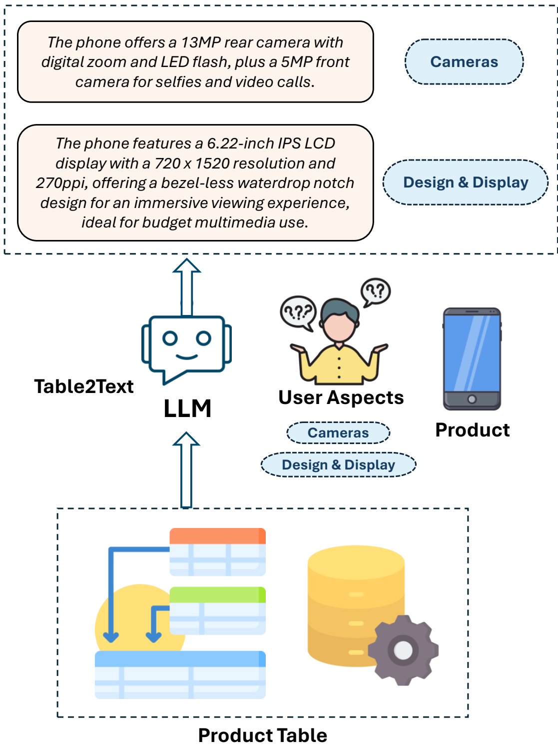

Figure 1: Overview of **eC-Tab2Text**. Illustration of aspect-based text generation from e-commerce product tables, where an LLM generates summaries for user-specific aspects like "Camera" and "Design & Display."

1 Introduction

E-commerce relies heavily on tabular data, such as product details and features, while user interactions, including assistant agents and Q&A, predominantly occur in natural language. This disparity underscores the need for models that can effectively parse tabular data and engage users through coherent, context-aware communication (Zhao et al., 2023b). Table-to-text generation addresses this challenge by transforming structured data into natural language, enabling applications such as product reviews, personalized descriptions, and tailored summaries in e-commerce. Beyond e-commerce, this capacity extends to domains such as healthcare, where structured patient records are converted into concise summaries for doctors (He et al., 2023), and finance, where tabular financial data is transformed into analytical reports (Varshney, 2024). However, generating text that is coherent, contextually relevant, and aligned with user-specific requirements remains a significant challenge, particularly for user- or query-centric tasks that demand domain-specific knowledge. Existing table-to-text datasets often focus on general-purpose applications and lack the depth required for specialized domains. For instance, datasets like QTSumm (Zhao

[*] Corresponding authors.
[1]Our code, dataset, evaluation, model outputs, and other resources are publicly available at eC-Tab2Text.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 849–867
April 30, 2025 ©2025 Association for Computational Linguistics

et al., 2023a) offer tabular summaries unrelated to the product domain, limiting their relevance for generating attribute-specific product reviews. E-commerce text generation requires handling diverse attributes (e.g., battery life, display quality), reasoning across different attributes (e.g., battery life and display size) and adapting to various user intents, such as crafting targeted product reviews (Macková and Pilát, 2024).

While Large Language Models (LLMs) excel in general-purpose text generation (Touvron et al., 2023; Kabir et al., 2024), and fine-tuned models like `LLama2` (Touvron et al., 2023), resulting in `StructLM` (Zhuang et al., 2024) have shown improved performance on table-based datasets, these approaches often struggle with the complexities of product-specific domains. Addressing these intricacies requires tailored datasets to capture the nuanced requirements of attribute-specific text generation. Table-to-text generation has benefited from datasets like WikiTableT (Chen et al., 2021), TabFact (Chen et al., 2020b), and ROTOWIRE (Wiseman et al., 2017). However, these datasets, designed for tasks like Wikipedia table descriptions, fact-checking, and sports summaries, lack the relevance for product-specific applications. Similarly, LogicNLG (Chen et al., 2020a) and ToTTo (Parikh et al., 2020) emphasize logical inferences and refined sentence extraction but fall short in addressing the demands of e-commerce text generation (He and Abisado, 2023).

This paper introduces a tailored table-to-text dataset for the products domain and explores the potential of fine-tuned LLMs to bridge the gap between general-purpose capabilities and domain-specific needs. By leveraging domain-specific datasets and fine-tuning techniques, this work aims to empower e-commerce platforms to generate more precise and engaging product reviews given user aspects and tables (see Figure 1), enhancing customer satisfaction and business outcomes.

Our main contributions are as follows:

- We present **eC-Tab2Text**, a novel domain-specific dataset for table-to-text generation in the e-commerce domain. The dataset features attribute-rich product tables paired with user-specific queries and outputs.
- We fine-tune open-source LLMs on the **eC-Tab2Text** dataset, resulting in significant improvements in text generation performance across various metrics.
- We provide a detailed analysis of domain robustness by comparing models trained on **eC-Tab2Text** with those trained on QTSumm, highlighting the critical need for domain-specific datasets to achieve superior performance in e-commerce applications.

2 Related Work

Table-to-Text Generation Table-to-text generation has advanced through datasets tailored to diverse domains and applications, as summarized in Table 1. Early efforts, such as WikiTableT (Chen et al., 2021), focused on generating natural language descriptions from Wikipedia tables, while TabFact (Chen et al., 2020b) introduced fact-checking capabilities and ROTOWIRE (Wiseman et al., 2017) generated detailed sports summaries. However, these datasets are limited in their relevance to product-specific domains. Later datasets like LogicNLG (Chen et al., 2020a) emphasized logical inference and reasoning, and ToTTo (Parikh et al., 2020) supported controlled text generation by focusing on specific table regions. HiTab (Cheng et al., 2022) extended these capabilities with hierarchical table structures and reasoning operators. Despite these advancements, none of these datasets provide the contextual and attribute-specific depth necessary for e-commerce applications, where generating meaningful descriptions requires reasoning across heterogeneous attributes, such as linking battery capacity to battery life or associating display size with user experience.

Query-Focused Summarization (QFS) Advances in text summarization have improved multi-document summarization through abstractive methods like paraphrastic fusion (Nayeem and Chali, 2017b; Nayeem et al., 2018), compression (Nayeem et al., 2019; Chowdhury et al., 2021), and diverse fusion models (Fuad et al., 2019; Nayeem, 2017), among others (Nayeem and Chali, 2017a; Chali et al., 2017). These approaches lay the groundwork for query-focused summarization (QFS), which tailors summaries to user-specific queries. Initially formulated as a document summarization task, QFS aims to generate summaries tailored to specific user queries (Dang, 2006). Despite its potential real-world applications, QFS remains a challenging task due to the lack of datasets. In the textual domain, QFS has been explored in multi-document settings (Giorgi et al., 2023) and meeting summarization (Zhong et al., 2021). Recent

Dataset	Table Source	# Tables / Statements	# Words / Statement	Explicit Control
		Single-sentence Table-to-Text		
ToTTo (Parikh et al., 2020)	Wikipedia	83,141 / 83,141	17.4	Table region
LOGICNLG (Chen et al., 2020a)	Wikipedia	7,392 / 36,960	14.2	Table regions
HiTab (Cheng et al., 2022)	Statistics web	3,597 / 10,672	16.4	Table regions & reasoning operator
		Generic Table Summarization		
ROTOWIRE (Wiseman et al., 2017)	NBA games	4,953 / 4,953	337.1	✗
SciGen (Moosavi et al., 2021)	Sci-Paper	1,338 / 1,338	116.0	✗
NumericNLG (Suadaa et al., 2021)	Sci-Paper	1,355 / 1,355	94.2	✗
		Table Question Answering		
FeTaQA (Nan et al., 2022)	Wikipedia	10,330 / 10,330	18.9	Queries rewritten from ToTTo
		Query-Focused Table Summarization		
QTSumm (Zhao et al., 2023a)	Wikipedia	2,934 / 7,111	68.0	Queries from real-world scenarios
eC-Tab2Text (*ours*)	e-Commerce products	1,452 / 3,354	56.61	Queries from e-commerce products

Table 1: Comparison between **eC-Tab2Text** (*ours*) and existing table-to-text generation datasets. Statements and queries are used interchangeably. Our dataset specifically comprises tables from the e-commerce domain.

datasets like QTSumm (Zhao et al., 2023a) extend QFS to a new modality, using tables as input. However, QTSumm's general-purpose nature limits its applicability to product reviews, which require nuanced reasoning over attributes and user-specific contexts. Additionally, its queries are often disconnected from real-world e-commerce scenarios. In contrast, our proposed dataset, **eC-Tab2Text**, bridges this gap by providing attribute-specific and query-driven summaries tailored to e-commerce product tables.

3 eC-Tab2Text: Dataset Construction

To address the gap in table-to-text generation for user-specific aspects or queries, such as "Camera" and "Design & Display" (as illustrated in Figure 1), we developed the **eC-Tab2Text** dataset. This dataset comprises e-commerce product tables and is designed to facilitate aspect-based text generation by fine-tuning LLMs on our dataset. The pipeline for creating **eC-Tab2Text** is outlined in Figure 2 and described in detail below.

Data Sources The dataset was constructed using product reviews and specifications (i.e., tables) extracted from the Pricebaba website[2]. Pricebaba provides comprehensive information on electronic products, including mobile phones and laptops. For this study, the focus was exclusively on mobile phone data due to the richness of product specifications (attribute-value pairs) and the availability of detailed expert reviews as summaries. Additionally, the number of samples available for mobile phones is significantly larger than for laptops. Each sample includes feature-specific details such as camera performance, battery life, and display quality.

Data Extraction and Format Data extraction was performed using web scraping techniques, with the extracted data stored in JSON format to serialize the table structure and to ensure compatibility with modern data processing workflows. Two JSON files were generated (Appendix E): one containing aspect-based product reviews and the other containing product specifications. The review JSON file captures user aspects alongside their associated textual descriptions collected from the "Quick Review" section of the website, while the specifications JSON file stores key-value pairs for both key specifications and full technical details. The structures of the sample inputs and outputs are depicted in Figures 3 and 4 in the Appendix.

Data Cleaning, Normalization, and Integration To ensure consistency, usability, and completeness, the extracted data underwent rigorous cleaning, normalization, and integration, similar to previous approaches (Nayeem and Rafiei, 2023, 2024a,b). The process includes **(1)** standardizing all text values to lowercase for uniformity, **(2)** replacing special

[2] https://pricebaba.com, last accessed August 2024.

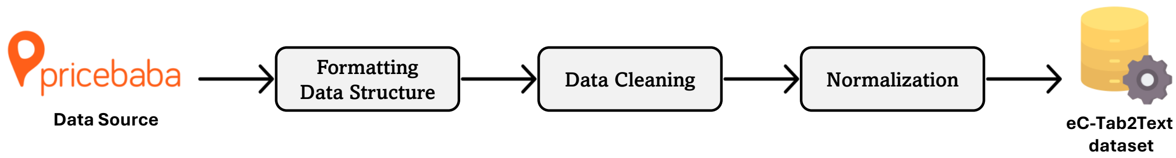

Figure 2: Data collection pipeline for our **eC-Tab2Text** dataset.

characters (e.g., & with "and") to improve readability, and **(3)** normalizing keys to maintain logical and contextual coherence. For example, the key `Display & Design` was transformed into `Design and Display` to improve readability and alignment with naming conventions.

To further enhance the dataset quality, irrelevant and redundant entries were removed through a systematic filtering process: **(1)** reviews lacking textual content in the text field were discarded, **(2)** specifications containing only generic or minimal information (e.g., entries labeled as `General`) were excluded, **(3)** overly simplistic reviews categorized as `Overview` were omited to maintain a focus on detailed and meaningful content.

Finally, the reviews and specifications JSON files were merged into a unified dataset by aligning entries based on their unique product URLs. This integration consolidated each product's reviews and specifications into a single, cohesive record, creating a streamlined and comprehensive dataset for downstream applications.

Metric	Value
Input	
# Tables	1,452
Avg # Attribute-Value Pairs	59.8
Max # Attribute-Value Pairs	68
Output	
# Queries	3,354
Avg # queries/table	2.31
Avg # words/query	56.61

Table 2: Statistics of our **eC-Tab2Text** dataset.

Our **eC-Tab2Text** dataset provides a comprehensive resource for table-to-text generation tasks based on user queries, as summarized in Table 2. The input JSON files contain attribute-rich product specifications, averaging 59.8 attribute-value pairs per table, with the largest entries containing up to 68 pairs. The dataset includes 3,354 queries, averaging 2.31 queries per table, with concise outputs averaging 56.61 words per query. This design supports query-specific training and evaluation of LLMs, enabling precise and contextually relevant text generation tailored to user queries.

4 eC-Tab2Text: Models

This section outlines the methodology for table serialization and provides details on the selection and fine-tuning of LLMs using our dataset.

Table Serialization The representation of tabular data in machine learning has been addressed through various serialization techniques, including markdown format, comma-separated values (CSV), HTML (Fang et al., 2024; Singha et al., 2023), and LaTeX (Jaitly et al., 2023). However, for our specific problem involving semi-structured tables with nested structures, we adopt JSON serialization. This approach effectively addresses two critical needs: **(1)** representing the nested structures inherent in product tables and **(2)** enabling query-specific generation and evaluation (Gao et al., 2024).

In our eC-Tab2Text dataset, both input tables and query-specific outputs are serialized using JSON. The input JSON captures structured product specifications, while the output JSON aligns queries (e.g., "Design and Display" or "Battery") as keys and their corresponding generated texts as values. This unified representation facilitates efficient querying and maintains alignment between inputs and outputs, ensuring consistency across the dataset. Additional implementation details can be found in Appendix D (Listing 7 prompt).

Model Selection and Characteristics To evaluate the effectiveness of the eC-Tab2Text dataset, we fine-tuned three open-source LLMs: **LLaMA 2-Chat 7B** (Touvron et al., 2023), **Mistral 7B-Instruct** (Jiang et al., 2023), and **StructLM 7B** (Zhuang et al., 2024). These models were selected due to their distinct pretraining paradigms, which address diverse data modalities and tasks. Detailed descriptions of these models are provided in Appendix B and summarized below.

- **LLaMA 2-Chat 7B**[3]: This model, pretrained on 2 trillion tokens of publicly available text data, is fine-tuned on over one million human-annotated examples. It excels in general-purpose conversational and language understanding tasks (Touvron et al., 2023).

- **Mistral 7B-Instruct**[4]: Leveraging a mix of text and code during training, this model demonstrates strong performance in tasks that require natural language understanding and programming-related reasoning (Jiang et al., 2023).

- **StructLM 7B**[5]: Pretrained on structured data, including databases, tables, and knowledge graphs, StructLM is optimized for structured knowledge grounding, making it particularly effective for domain-specific tasks (Zhuang et al., 2024).

Fine-Tuning Process The fine-tuning process adapts these models to the e-commerce domain using the eC-Tab2Text dataset. This dataset focuses on attribute-specific and context-aware text generation tailored to user queries, such as detailed reviews of "Camera" or "Design & Display." The fine-tuning process follows best practices in instruction tuning and domain-specific dataset alignment (Zhang et al., 2023; Chang et al., 2024). Optimization of hyperparameters ensured computational efficiency while maintaining high-quality performance, as detailed Table 4.

By leveraging these diverse models and aligning them with the eC-Tab2Text dataset, this work aims to advance the state-of-the-art in domain-specific language generation for e-commerce applications.

Column Name	Data Type	Description
`table`	Dictionary	Contains structured data with headers and rows.
`example_id`	String	Unique identifier for each dataset example.
`query`	String	Textual description or query related to the dataset.
`summary`	String	Summary or explanation generated in response to the query.
`row_ids`	Sequence of Integers	Row indices corresponding to the data referenced in the `table` column.

Table 3: Structure of the QTSUMM Dataset.

Hyperparameter	Value
Learning Rate	2×10^{-4}
Batch Size	2
Epochs	1
Gradient Accumulation Steps	1
Weight Decay	0.001
Max Sequence Length	900

Table 4: Hyperparameter settings for fine-tuning.

Hyperparameter	Value
`bnb_4bit_compute_dtype`	float16
`bnb_4bit_quant_type`	nf4
`use_nested_quant`	False

Table 5: Quantization settings used for fine-tuning.

5 Evaluation

In this section, we evaluate the performance of the eC-Tab2Text models described in Section 4 along with several closed-source models, including GPT-4o-mini and Gemini-1.5-flash. The evaluation follows standard metrics commonly used in table-to-text generation, as outlined in (Zhao et al., 2023a). These metrics include BLEU (Reiter, 2018), the F-1 scores of ROUGE-1 and ROUGE-L (Ganesan, 2018), METEOR (Dobre, 2015), and BERTScore (Zhang* et al., 2020), following (Akash et al., 2023; Shohan et al., 2024). To assess the correctness, faithfulness, and fluency of the generated text, we employ PROMETHEUS 2 (Kim et al., 2024) and an open-source LLM-based evaluator as an alternative to the closed-source G-Eval (Liu et al., 2023). Our objective is to benchmark the performance of various LLMs under both zero-shot and fine-tuned settings using the proposed eC-Tab2Text dataset.

Experimental Setup The fine-tuning process was conducted on a NVIDIA RTX 4070 Ti Super GPU with 16GB of VRAM, ensuring efficient training while managing memory-intensive operations. The AdamW optimizer (Loshchilov and Hutter, 2019) was configured with a learning rate of

[3] Llama-2-7b-chat-hf
[4] Mistral-7B-Instruct-v0.3
[5] StructLM-7B

Mode	Models	BLEU	METEOR	ROUGE-1	ROUGE-L	BERTScore	Correctness	Faithfulness	Fluency
Zero-Shot	`Llama2`	1.39	3.59	5.57	4.09	66.49	32.18	37.68	32.47
	`StructLM`	6.21	11.96	20.09	15.34	82.56	64.30	70.08	63.10
	`Mistral`	4.19	9.55	25.64	18.99	82.12	77.02	81.16	76.5
	`GPT-4o-mini`	7.14	16.12	29.44	19.47	83.75	**80.89**	**83.92**	**80.81**
	`Gemini-1.5-flash`	8.8	15.18	30.38	21.51	84.05	78.79	83.04	78.54
Fine-tuned	`Llama2`	29.36	40.2	48.36	39.25	90.05	61.38	63.78	61.47
	`StructLM`	31.06	42.3	49.42	40.58	90.9	69.70	72.46	69.93
	`Mistral`	**38.89**	**49.43**	**56.64**	**48.32**	**92.18**	73.07	76.63	73.03

Table 6: Evaluation results of zero-shot and fine-tuned models on the **eC-Tab2Text** dataset. The best results are highlighted in **bold**, and the second-best results are underlined.

Dataset Trained	Dataset Tested	Models	BLEU	METEOR	ROUGE-1	ROUGE-L	BERTScore	Correctness	Faithfulness	Fluency
QTSumm	QTSumm (`In-domain`)	`Llama2`	13.32	32.38	26.3	19.22	86.47	51.09	57.30	48.98
		`StructLM`	6.6	22.04	13.52	10.04	84.5	41.14	48.92	39.68
		`Mistral`	10.1	28.57	20.7	15.51	85.65	49.99	57.73	50.71
	eC-Tab2Text (`Out-of-domain`)	`Llama2`	17.47	40.2	35.69	21.14	85.41	63.98	71.40	64.07
		`StructLM`	3.73	17.42	10.41	6.77	82.91	36.69	60.81	37.03
		`Mistral`	13.97	26.88	28.58	17.08	84.83	58.35	69.81	58.95
eC-Tab2Text	QTSumm (`Out-of-domain`)	`Llama2`	6.5	22.77	7.79	16.59	81.93	48.42	48.66	48.55
		`StructLM`	10.15	30.59	30.59	23.04	85.13	58.71	56.60	58.26
		`Mistral`	10.39	18.11	30.27	24.24	84.23	64.83	61.14	64.51
	eC-Tab2Text (`In-domain`)	`Llama2`	29.4	40.21	48.43	39.25	90.05	61.38	63.78	61.47
		`StructLM`	31.06	42.3	49.42	40.58	90.9	69.70	72.46	69.93
		`Mistral`	**38.89**	**49.43**	**56.64**	**48.32**	**92.18**	**73.07**	**76.63**	**73.03**

Table 7: Robustness evaluation results on our **eC-Tab2Text** dataset and the QTSumm dataset (Zhao et al., 2023a). The best results on our dataset, including both in-domain and out-of-domain scenarios, are highlighted in **bold**, while the best results on the QTSumm dataset, both in-domain and out-of-domain, are underlined.

2×10^{-4}, chosen for its effectiveness in maintaining stability and convergence during training. To optimize resource usage, the *bitsandbytes* library[6] was employed for 4-bit quantization, reducing VRAM requirements without significant performance loss. Table 5 outlines the key parameters used, including '`float16`' for computation data type and '`nf4`' for quantization type. The '`use_nested_quant`' option was set to 'False' to ensure compatibility across models.

Detailed information on the evaluation metrics is included in Appendix A. Our eC-Tab2Text dataset was divided into training and testing subsets, using an 80%-20% split. This ensures a sufficient volume of data for training while preserving a reliable subset for evaluation.

5.1 Robustness Evaluation

We evaluate the robustness of the models under domain differences, focusing on their performance with in-domain and out-of-domain training data. The primary objective is to analyze how models perform when fine-tuned on data from different domains and to emphasize the importance of our proposed eC-Tab2Text dataset for the e-commerce product domain. For this evaluation, we compare the performance of models fine-tuned on the QTSumm dataset (Zhao et al., 2023a), which contains Wikipedia tables with queries, against those fine-tuned on our eC-Tab2Text dataset, which consists of product tables with user-specific queries.

QTSumm Dataset Details The QTSumm dataset, obtained from Hugging Face[7] provides structured data that facilitates query-specific text summarization tasks. The detailed structure of QTSumm is outlined in Table 3. This dataset's structure ensures a systematic alignment between the input queries, the corresponding structured data, and the generated summaries, making it a valuable benchmark for fine-tuning and evaluating the performance of LLMs in handling structured data. Its focus on query-specific summarization provided an excellent foundation for testing the robustness and adaptability of the proposed methodologies.

For fine-tuning, we utilized the same models described in Section 4, employing identical hyperparameters. The models were trained using prompts structured consistently with those designed for the

[6] `https://github.com/bitsandbytes-foundation/bitsandbytes`

[7] `https://huggingface.co/datasets/yale-nlp/QTSumm`

eC-Tab2Text dataset. However, in the QTSumm setup, the prompts included row-level content tailored to the dataset's structure, as outlined in Appendix D (Listing 8). This alignment ensured methodological consistency while accounting for the unique characteristics of the QTSumm dataset. By highlighting these differences, our evaluation underscores the critical need for domain-specific datasets, such as eC-Tab2Text, to achieve robust and accurate performance in the product domain.

5.2 Results & Analysis

Our experimental results, illustrated in Table 6, demonstrate that fine-tuning open-source 7B models on our dataset leads to substantial performance improvements. These fine-tuned models significantly outperform major proprietary models, such as GPT-4o-mini and Gemini-1.5-flash, across text-based metrics, including BLEU, ROUGE-1, ROUGE-L, METEOR, and BERTScore, while achieving competitive results in model-based metrics like faithfulness, correctness, and fluency, narrowing the gap with proprietary counterparts. This is significant given the relatively small size of our dataset compared to the much larger datasets used for training many proprietary models. Notably, Mistral_Instruct, fine-tuned on our dataset, excels by achieving the highest scores across all metrics, surpassing both zero-shot and fine-tuned models.

As highlighted in Table 7, the robustness of our dataset is further evidenced by comparing it against the QTSUMM dataset; models trained with our dataset consistently outperform those trained on QTSUMM across both in-domain and out-of-domain tasks, with Mistral_Instruct leading, followed closely by StructLM. Although both datasets share similar task objectives, the domain differences significantly affect the models' performance.

Outputs generated by different open-source models are presented in Mistral (Listing 11), StructLM (Listing 14), and Llama2 (Listing 15), as well as by closed-source models GPT-4o-mini (Listing 13) and Gemini1.5-flash (Listing 12). Notably, the closed-source models tend to produce longer outputs compared to the open-source models, with their outputs often containing nested keys and detailed information.

6 Discussion and Future Directions

This section highlights the need for better numerical reasoning in table-to-text generation and improved evaluation methods.

Numerical Reasoning Product tables, with their semi-structured and nested attributes (e.g., battery capacity in mAh, display size in inches), demand advanced numerical reasoning to generate meaningful text. Models must analyze relationships, such as how battery life depends on capacity and display size, or how display dimensions impact user experience. Unlike Wikipedia tables (Zhao et al., 2023a; Nahid and Rafiei, 2024), which focus on factual text generation, our eC-Tab2Text dataset challenges models to integrate numerical reasoning with qualitative text generation (Islam et al., 2024). This unique focus enables LLMs to synthesize structured data into nuanced, human-readable summaries, providing a benchmark for evaluating and improving reasoning capabilities in real-world applications (Naeim abadi et al., 2023; Akhtar et al., 2023; Zhao et al., 2024). Future work could explore pushing the boundaries of LLMs capabilities in numerical and qualitative reasoning using our dataset.

Evaluation Although we evaluated the correctness, faithfulness, and fluency of the generated text using PROMETHEUS 2 (Kim et al., 2024), attribute-specific text evaluation against product tables requires a more nuanced approach. Future evaluations could involve extracting attribute-value pairs from the generated text (Shinzato et al., 2023; Brinkmann et al., 2024), verifying their correctness and contextual relevance, and comparing them with the corresponding values in the source tables to enable more fine-grained and precise assessments.

7 Conclusion

This work introduces **eC-Tab2Text**, a novel dataset for table-to-text generation in the e-commerce domain, addressing the limitations of existing general-purpose datasets. By fine-tuning open-source LLMs, we demonstrate substantial improvements in generating attribute-specific, contextually accurate product reviews. Our evaluation highlights the robustness of **eC-Tab2Text**, outperforming comparable datasets like QTSumm, and underscores the importance of domain-specific datasets for advancing LLM performance in industry-specific applications. This study lays the groundwork for future research in expanding dataset scope, evaluation methodologies, and enhancing numerical reasoning in e-commerce workflows.

Limitations

In this work, we evaluated our proposed methods using a selection of both open-source and closed-source LLMs. We intentionally focused on cost-effective yet efficient closed-source models and open-source models deployable on consumer-grade hardware, considering the constraints of *academic settings*. The performance of more powerful, large-scale models remains unexplored; however, we encourage the broader research community to benchmark these models using our dataset. To support future research, we make our code, dataset, evaluation, model outputs, and other resources publicly available[8].

This study faced several system and resource constraints that shaped the methodology and evaluation process. For example, VRAM limitations required capping the maximum token length at 900 for the Mistral_Instruct model to ensure uniform hyperparameter settings across all models. While this standardization enabled consistent comparisons, it may have limited some models' ability to generate longer and potentially more nuanced outputs.

Our dataset focused exclusively on mobile phone data due to the richness of product specifications (attribute-value pairs) and the availability of detailed expert reviews as summaries. Future work could expand the dataset to include other domains, such as laptops, home appliances, and wearable devices, to assess the generalizability of the LLMs in e-Commerce domains.

Finally, the development of eC-Tab2Text has been exclusively centered on the **English language**. As a result, its effectiveness and applicability may differ for other languages. Future research could explore multilingual extensions to broaden its usability across diverse linguistic and cultural contexts.

Ethics Statement

The data scraping process for this research was conducted with strict adherence to ethical guidelines and solely for non-commercial research purposes, under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0)[9]. To minimize potential harm to the source website, measures were implemented to ensure controlled and responsible scraping practices. These safeguards were designed to avoid undue strain on the website's infrastructure, such as preventing Distributed Denial of Service (DDoS) attacks, thereby maintaining the integrity and functionality of the site.

[8] https://github.com/Luis-ntonio/eC-Tab2Text

[9] https://creativecommons.org/licenses/by-nc-sa/4.0/

Acknowledgements

We thank all the anonymous reviewers and the meta-reviewer for their valuable feedback and constructive suggestions. This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Additionally, Luis Antonio Gutiérrez Guanilo is supported by the Emerging Leaders in the Americas Program (ELAP), and Mir Tafseer Nayeem is supported by a Huawei PhD Fellowship.

References

Abu Ubaida Akash, Mir Tafseer Nayeem, Faisal Tareque Shohan, and Tanvir Islam. 2023. Shironaam: Bengali news headline generation using auxiliary information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 52–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405, Singapore. Association for Computational Linguistics.

Alexander Brinkmann, Roee Shraga, and Christian Bizer. 2024. Extractgpt: Exploring the potential of large language models for product attribute value extraction. *Preprint*, arXiv:2310.12537.

Yllias Chali, Moin Tanvee, and Mir Tafseer Nayeem. 2017. Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 418–424, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2021. WikiTableT: A large-scale data-to-text dataset for generating Wikipedia article sections. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 193–209, Online. Association for Computational Linguistics.

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 *International Conference on Learning Representations*.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland. Association for Computational Linguistics.

Radia Rayan Chowdhury, Mir Tafseer Nayeem, Tahsin Tasnim Mim, Md. Saifur Rahman Chowdhury, and Taufiqul Jannat. 2021. Unsupervised abstractive summarization of Bengali text documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2612–2619, Online. Association for Computational Linguistics.

Hoa Trang Dang. 2006. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48–55, Sydney, Australia. Association for Computational Linguistics.

Iuliana Dobre. 2015. A comparison between bleu and meteor metrics used for assessing students within an informatics discipline course. *Procedia - Social and Behavioral Sciences*, 180:305–312.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. Large language models (LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*.

Tanvir Ahmed Fuad, Mir Tafseer Nayeem, Asif Mahmud, and Yllias Chali. 2019. Neural sentence fusion for diversity driven abstractive multi-document summarization. *Computer Speech & Language*, 58:216–230.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *Preprint*, arXiv:1803.01937.

Chang Gao, Wenxuan Zhang, Guizhen Chen, and Wai Lam. 2024. Jsontuning: Towards generalizable, robust, and controllable instruction tuning. *Preprint*, arXiv:2310.02953.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Wang, and Arman Cohan. 2023. Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8177–8199, Singapore. Association for Computational Linguistics.

Aixiang He and Mideth B. Abisado. 2023. Review on sentiment analysis of e-commerce product comments. *2023 IEEE 15th International Conference on Advanced Infocomm Technology (ICAIT)*, pages 398–406.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Preprint*, arXiv:2310.05694.

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sukriti Jaitly, Tanay Shah, Ashish Shugani, and Razik Singh Grewal. 2023. Towards better serialization of tabular data for few-shot classification with large language models. *Preprint*, arXiv:2312.12464.

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

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2024. BenLLM-eval: A comprehensive evaluation into the potentials and pitfalls of large language models on Bengali NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2238–2252, Torino, Italia. ELRA and ICCL.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4).

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Kateřina Macková and Martin Pilát. 2024. Promap: Product mapping datasets. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II*, page 159–172, Berlin, Heidelberg. Springer-Verlag.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2022. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1731–1751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ali Naeim abadi, Mir Tafseer Nayeem, and Davood Rafiei. 2023. Product entity matching via tabular data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4215–4219, New York, NY, USA. Association for Computing Machinery.

Md Nahid and Davood Rafiei. 2024. TabSQLify: Enhancing reasoning capabilities of LLMs through table decomposition. 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 5725–5737, Mexico City, Mexico. Association for Computational Linguistics.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Mir Tafseer Nayeem. 2017. *Methods of Sentence Extraction, Abstraction and Ordering for Automatic Text Summarization*. Universtiy of Lethbridge, Department of Mathematics and Computer Science.

Mir Tafseer Nayeem and Yllias Chali. 2017a. Extract with order for coherent multi-document summarization. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 51–56, Vancouver, Canada. Association for Computational Linguistics.

Mir Tafseer Nayeem and Yllias Chali. 2017b. Paraphrastic fusion for abstractive multi-sentence compression generation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2223–2226, New York, NY, USA. Association for Computing Machinery.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2019. Neural diverse abstractive sentence compression generation. In *Advances in Information Retrieval (ECIR)*, pages 109–116, Cham. Springer International Publishing.

Mir Tafseer Nayeem and Davood Rafiei. 2023. On the role of reviewer expertise in temporal review helpfulness prediction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1684–1692, Dubrovnik, Croatia. Association for Computational Linguistics.

Mir Tafseer Nayeem and Davood Rafiei. 2024a. KidLM: Advancing language models for children – early insights and future directions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4813–4836, Miami, Florida, USA. Association for Computational Linguistics.

Mir Tafseer Nayeem and Davood Rafiei. 2024b. Lfosum: Summarizing long-form opinions with large language models. *Preprint*, arXiv:2410.13037.

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.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Keiji Shinzato, Naoki Yoshinaga, Yandi Xia, and Wei-Te Chen. 2023. A unified generative approach to product attribute-value identification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6599–6612, Toronto, Canada. Association for Computational Linguistics.

Faisal Tareque Shohan, Mir Tafseer Nayeem, Samsul Islam, Abu Ubaida Akash, and Shafiq Joty. 2024. XL-HeadTags: Leveraging multimodal retrieval augmentation for the multilingual generation of news headlines and tags. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12991–13024, Bangkok, Thailand. Association for Computational Linguistics.

Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. Tabular representation, noisy operators, and impacts on table structure understanding tasks in LLMs. In *NeurIPS 2023 Second Table Representation Learning Workshop*.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1451–1465, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Tanay Varshney. 2024. Build an llm-powered data agent for data analysis.

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.

Yuekun Yao and Alexander Koller. 2024. Predicting generalization performance with correctness discriminators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11725–11739, Miami, Florida, USA. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *ArXiv*, abs/2308.10792.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Ruizhe Chen, Xiangru Tang, Yumo Xu, Dragomir Radev, and Arman Cohan. 2023a. QTSumm: Query-focused summarization over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1157–1172, Singapore. 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.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Alex Zhuang, Ge Zhang, Tianyu Zheng, Xinrun Du, Junjie Wang, Weiming Ren, Wenhao Huang, Jie Fu, Xiang Yue, and Wenhu Chen. 2024. StructLM: Towards building generalist models for structured knowledge grounding. In *First Conference on Language Modeling*.

Supplementary Material: Appendices

A Evaluation Metrics

- **BLEU (Bilingual Evaluation Understudy)**: Commonly used in machine translation and natural language generation, BLEU measures the overlap of n-grams between generated and reference texts. Despite its popularity, BLEU has limitations, particularly in capturing semantic similarity and evaluating beyond exact matches (Reiter, 2018).

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: Focuses on recall-oriented evaluation by comparing the overlap of n-grams, word sequences, and word pairs between generated summaries and reference texts. It is highly effective for summarization tasks (Ganesan, 2018).

- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)**: Incorporates stemming, synonymy, and flexible matching, providing a more nuanced evaluation than BLEU. It strongly correlates with human judgments, especially in translation tasks (Dobre, 2015).

- **BERTScore**: Leverages contextual embeddings from pre-trained transformer models to measure semantic similarity between generated and reference texts. Unlike n-gram-based metrics, BERTScore captures meaning and context, offering a robust evaluation for text generation tasks (Zhang* et al., 2020).

The reliability and faithfulness of generated text, particularly in applications requiring high accuracy, such as medical or financial domains is crucial. To identify inaccuracies, hallucination detection was conducted using Prometheus 2, a robust evaluation model designed for analyzing outputs of Large Language Models (LLMs) (Kim et al., 2024). This framework helps evaluate three critical dimensions:

- **Faithfulness**: Ensures that the generated content aligns with the source data and avoids unsupported claims (Madsen et al., 2022; Jacovi and Goldberg, 2020).

- **Correctness**: Measures factual accuracy and checks for logical consistency in the output (Yao and Koller, 2024; Kim et al., 2024).

- **Fluency**: Evaluates the readability and linguistic quality of the text, ensuring it adheres to natural language norms (Suadaa et al., 2021; Lee et al., 2023).

B Models for Fine-tuning

- **LLaMA 2-Chat 7B** (Touvron et al., 2023): LLaMA 2-Chat 7B is a fine-tuned variant of the LLaMA 2 series, optimized for dialogue applications. It employs an autoregressive transformer architecture and has been trained on a diverse dataset comprising 2 trillion tokens from publicly available sources. The fine-tuning process incorporates over one million human-annotated examples to enhance its conversational capabilities and alignment with human preferences for helpfulness and safety.

- **StructLM 7B** (Zhuang et al., 2024): StructLM 7B is a large language model fine-tuned specifically for structured knowledge grounding tasks. It utilizes the CodeLlama-Instruct model as its base and is trained on the SKGInstruct dataset, which encompasses a mixture of 19 structured knowledge grounding tasks. This specialized training enables StructLM to effectively process and generate text from structured data sources such as tables, databases, and knowledge graphs, making it robust in domain-specific text generation tasks.

- **Mistral 7B-Instruct** (Jiang et al., 2023): Mistral 7B-Instruct is an instruction fine-tuned version of the Mistral 7B model, designed to handle a wide array of tasks by following diverse instructions. It features a 32k context window and employs a Rope-theta of 1e6, without utilizing sliding-window attention. This configuration allows Mistral 7B-Instruct to perform effectively in multi-modal and domain-adapted text generation scenarios, achieving state-of-the-art performance in various benchmarks.

C Prometheus Evaluation

To evaluate model-based metrics, the Prometheus framework (Kim et al., 2024) was employed, utilizing structured prompts for three key evaluation criteria: fluency, correctness, and faithfulness. The primary framework leverages an Absolute System Prompt, which defines the role of the evaluator and ensures objective, consistent assessments based on established rubrics. This Absolute System Prompt, shown in Listing 1, forms the foundation for all evaluations across metrics.

Listing 1: Absolute System Prompt

```
You are a fair judge assistant tasked
    with providing clear, objective
    feedback based on specific criteria,
     ensuring each assessment reflects
    the absolute standards set for
    performance.
```

The task descriptions for evaluating fluency, correctness, and faithfulness share a similar structure, as shown in Listing 2,3. These instructions define the evaluation process, requiring detailed feedback and a score between 1 and 5, strictly adhering to a given rubric.

Listing 2: Task description used for evaluation of faithfulness

```
###Task Description:
An instruction (might include an Input
    inside it), a response to evaluate,
    a reference answer that gets a score
     of 5, and a score rubric
    representing a evaluation criteria
    are given.
1. Write a detailed feedback that assess
     the quality of the response
    strictly based on the given score
    rubric, not evaluating in general.
2. After writing a feedback, write a
    score that is an integer between 1
    and 5. You should refer to the score
     rubric.
3. The output format should look as
    follows: "Feedback: (write a
    feedback for criteria) [RESULT] (an
    integer number between 1 and 5)"
4. Please do not generate any other
    opening, closing, and explanations.
5. Only evaluate on common things
    between generated answer and
    reference answer. Don't evaluate on
    things which are present in
    reference answer but not in
    generated answer.
```

C.1 Instructions for Evaluation

Prometheus prompts are customized for each evaluation metric. Below are the specialized structures and rubrics for fluency, faithfulness, and correctness.

Faithfulness This metric ensures the generated response aligns with both the provided context and reference answers. The evaluation structure incorporates specific rubrics for relevance and information consistency.

Listing 3: Task description used for evaluation of fluency and correctness

```
###Task Description:
An instruction (might include an Input
    inside it), a response to evaluate,
    a reference answer that gets a score
     of 5, and a score rubric
    representing a evaluation criteria
    are given.
1. Write a detailed feedback that assess
     the quality of the response
    strictly based on the given score
    rubric, not evaluating in general.
2. After writing a feedback, write a
    score that is an integer between 1
    and 5. You should refer to the score
     rubric.
3. The output format should look as
    follows: "Feedback: (write a
    feedback for criteria) [RESULT] (an
    integer number between 1 and 5)"
4. Please do not generate any other
    opening, closing, and explanations.
```

Listing 4: Prompt structured correctness

```
###The instruction to evaluate:
Evaluate the fluency of the generated
    JSON answer.
###Context:
{Prompt}
###Existing answer (Score 5):
{reference_answer}
###Generate answer to evaluate:
{response}
###Score Rubrics:
"score1_description":"If the generated
    answer is not matching with any of
    the reference answers and also not
    having information from the context
     .",
"score2_description":"If the generated
    answer is having information from
    the context but not from existing
    answer and also have some irrelevant
     information.",
"score3_description":"If the generated
    answer is having relevant
    information from the context and
    some information from existing
    answer but have additional
    information that do not exist in
    context and also do not in existing
    answer.",
"score4_description":"If the generated
    answer is having relevant
    information from the context and
    some information from existing
    answer.",
"score5_description":"If the generated
    answer is matching with the existing
     answer and also having information
    from the context."}
###Feedback:
```

Fluency This metric evaluates the grammatical accuracy and readability of the generated response.

Listing 5: Prompt structured fluency

```
###The instruction to evaluate: Evaluate
the fluency of the generated JSON answer
###Response to evaluate: {response}
###Reference Answer (Score 5):
{reference_answer}
###Score Rubrics:
"score1_description":"The generated JSON
     answer is not fluent and is
    difficult to understand.",
"score2_description":"The generated JSON
     answer has several grammatical
    errors and awkward phrasing.",
"score3_description":"The generated JSON
     answer is mostly fluent but
    contains some grammatical errors or
    awkward phrasing.",
"score4_description":"The generated JSON
     answer is fluent with minor
    grammatical errors or awkward
    phrasing.",
"score5_description":"The generated JSON
     answer is perfectly fluent with no
    grammatical errors or awkward phrase
###Feedback:
```

Correctness This metric assesses the logical accuracy and coherence of the generated response compared to the reference.

Listing 6: Prompt estructured correctness

```
###The instruction to evaluate:
Your task is to evaluate the generated
    answer and reference answer for the
    query: {Prompt}
###Response to evaluate:
{response}
###Reference Answer (Score 5):
{reference_answer}
###Score Rubrics:
"criteria": "Is the model proficient in
    generate a coherence response",
"score1_description": "If the generated
    answer is not matching with any of
    the reference answers.",
"score2_description": "If the generated
    answer is according to reference
    answer but not relevant to user
    query.",
"score3_description": "If the generated
    answer is relevant to the user query
     and reference answer but contains
    mistakes.",
"score4_description": "If the generated
    answer is relevant to the user query
     and has the exact same metrics as
    the reference answer, but it is not
    as concise.",
"score5_description": "If the generated
    answer is relevant to the user query
     and fully correct according to the
    reference answer.

###Feedback:
```

D Fine-tuning models

The prompts outlined below utilized for training eC-Tab2Text models (Listing 7) and for the QTSumm dataset (Listing 8).

Listing 7: Prompt structure for eC-Tab2Text

```
"Given following json that contains
    specifications of a product,
    generate a review of the key
    characteristics with json format.
    Follow the structure on Keys to
    write the Output:

### Product: Product for JSON
    specifications

### Keys: Combination of the keys of the
     JSON reviews

### Output: reviews for JSON reviews
    accordingly to the keys"
```

Listing 8: Prompt structure for QTSumm

```
"Given following json that contains
    specifications of a product,
    generate a review of the key
    characteristics with json format.
    Follow the structure on Keys to
    write the Output:
### Product: Column table of JSON
    specifications
### Keys: Column query of the dataset
### Output: Column summary of the
    dataset"
```

E eC-Tab2Text Data Formats

Listing 9: JSON Data Format Product specification

```
{
  "url": {
    "keys_specifications": [],
    "full_specifications": [
      "Launch Date": "Launch Date",
      "General": {
        "subcategories1": [
            "value1" ...
            ],
        "subcategories2": [
            "value1" ...
            ], ...
      },
      "Characteristic1": {
        "subcategories1": [
            "value1" ...
            ],
        "subcategories2": [
            "value1" ...
            ], ...
      },
      "Characteristic2": {
        "subcategories1": [
```

```
                "value1" ...
                ],
            "subcategories2": [
                "value1" ...
                ], ...
          }, ...
        ]
    },
}
```

Listing 10: JSON Data Format reviews

```
{
  "url": {
    "text": {
      "Characteristic1": ["Description1"
          ],
      "Characteristic2": ["Description2"
          ], ...
    }
  }
}
```

OnePlus Nord 3 5G Quick Review

Design and Display

The OnePlus Nord 3 5G could feature a 6.43 inch Fluid AMOLEDdisplay with a resolution of 1080 x 2400 pixelsand a pixel density of 409ppi. The display is said to come with a Punch-hole design and an aspect ratio of 20.4:9. The device will come with 90Hz refresh rate .

Cameras

The OnePlus Nord 3 5G is said to come with a triple camera system on the back with a powerful 50MP wide angle primary sensor, a 12MP wide angle sensor, a 5MP sensor, and an LED flash. On the front, The device will probably get a 32MP wide angle selfie cam. Auto Flash, Auto Focus, Bokeh Effect, Continuos Shooting, Exposure compensation, Face detection, Geo tagging, High Dynamic Range mode (HDR), ISO control, Touch to focus, White balance presets are some of the many features that the camera is likely to support.

Battery and Performance

The OnePlus Nord 3 5G is said to be embedded with a MediaTek Dimensity 1200 processor and a Mali-G77 MC9GPU. The RAM and internal memory of the device could possibly be 8GB and 128GB respectively. A large 4299mAh Li-Polymer battery could come with the device. It is said to have wrap charging too.

Software and Connectivity

OnePlus Nord 3 5G is likely to come with Android out of the box. The smartphone could get connectivity options like 5G ,dual sim , Wi-Fi 802.11, b/g/n, GPS, and Bluetooth 5.2. In terms of ports selection, the smartphone will probably be getting a USB Type-C port, and an on-screen fingerprint scanner.

Figure 3: An illustration of sample output texts generated for user-specific queries based on structured input from product tables.

OnePlus Nord 3 5G Full Features & Specifications

Report error on this page

Launched in: July 2023

Note: Scores are assigned in comparison to similarly priced products

General	
Operating System	Android 13
Custom UI	Oxygen OS
Dimensions	162.6mm x 75.1mm x 8.1mm
Weight	193.5g

Display & Design 8 / 10	
Size	6.74 inches (17.12 cm)
Resolution	1240 x 2772 pixels
Pixel Density	451ppi
Touch Screen	Yes, Capacitive Touchscreen, Multi-touch
Type	Super Fluid AMOLED, Auto-Brightness, Blue light filter, HDR 10+
Screen To Body Ratio	93.5 %
Aspect Ratio	20.1:9
Refresh Rate	120Hz
Design	Punch-hole display
Colour Options	Misty Green, Tempest Gray
Water Resistance	IP54, Splash proof

Hardware 9 / 10	
Chipset	MediaTek Dimensity 9000 MT6893
CPU	1 x 3.05GHz Cortex X2 3 x 2.85GHz Cortex A710 4 x 1.8GHz Cortex A510
GPU	Mali-G710 MC10
Architecture	64-bit
RAM	8 GB
Internal Storage	128 GB
MicroSD Card Slot	No

Main Camera 8 / 10	
Number of Cameras	Triple
Resolution	50 MP f/1.8 Wide Angle main camera PDAF, EIS, OIS, 20x Digital Zoom 8 MP f/2.2 ultra-wide camera 2 MP f/2.4 macro sensor
Flash	LED Flash
Video	3840x2160@30fps, 1920x1080@30fps
Features	AF Phase Detection, Artificial Intelligence, Auto Flash, Auto Focus, Bokeh Effect, Continuous Shooting, Electronic Image Stabilization (EIS), Exmor-RS CMOS

Figure 4: An example of a product specification table structure.

Listing 11: Output generated with Mistral Instruct trained with eC-Tab2Text.


```
{
    "Software and Connectivity": "Tecno Spark Go 2020 will run Android Go UI on top
        of Android 10 out of the box. For the connectivity, the smartphone will
        support features like Dual sim VoLTE (Micro-Micro), Wi-Fi 802.11, b/g/n, 3.5
        mm jack, Bluetooth 4.2, 10-pin charging, GPS, and USB 2.0.",
    "Design and Display": "The Tecno Spark Go 2020 will reportedly showcase a 6.52
        inch TFTdisplay with a resolution of 720 x 1600 pixelsand a pixel density of
         269ppi. The display will supposedly have a waterdrop notch with minimum
        bezels.",
    "Battery and Performance": "The Tecno Spark Go 2020 will be embedded with a
        MediaTek Helio A20 chipset and PowerVR GE8300 GPU. The device will come with
         2GB RAM and 32GB of internal memory, which can be expanded up to 256GB.
        Fuelling the device will be a huge 5000mAh battery."
}
```


Listing 12: Output generated with Gemini1.5-flash.


```
{
    "Cameras": {
        "Main Camera": {
            "Description": "Single 13 MP main camera with LED flash, Auto Flash, and
                 Auto Focus. Offers digital zoom.",
            "Video": "Specifications not provided."
        },
        "Front Camera": {
            "Description": "Single 8 MP main camera.",
            "Video": "Specifications not provided."
        }
    },
    "Battery and Performance": {
        "Battery": {
            "Capacity": "4000 mAh Li-ion battery",
            "Charging": "No fast charging or wireless charging."
        },
        "Performance": {
            "Processor": "Quad-core 1.3GHz processor.  Specific chipset and GPU
                information not provided.",
            "RAM": "4GB RAM",
            "Storage": "64GB internal storage, expandable up to 64GB via microSD
                card."
        }
    },
    "Design and Display": {
        "Display": {
            "Size": "6.53 inches (16.59 cm) IPS LCD with 720 x 1560 pixels
                resolution (263 ppi).",
            "Design": "Waterdrop notch display."
        },
        "Other": {
            "Colors": "Available in Blue and Purple.",
            "Fingerprint Sensor": "No fingerprint sensor."
        }
    }
}
```

Listing 13: Output generated with GPT-4o-mini.

```
{
    "Software and Connectivity": {
        "Operating System": "Android 10 (Q) with Android Go custom UI",
        "Network Compatibility": "Supports 2G, 3G, and 4G networks, making it
            suitable for use across various locations",
        "SIM Slots": "Dual SIM (Nano + Nano), allowing for flexible connectivity
            options",
        "Wi-Fi and Bluetooth": "Equipped with Wi-Fi 802.11 b/g/n and Bluetooth 4.2
            for seamless wireless connections",
        "GPS": "Includes A-GPS and Glonass for accurate location tracking",
        "FM Radio and Charging Port": "Features FM Radio and uses a microUSB 2.0
            charging port"
    },
    "Design and Display": {
        "Dimensions": "165.6mm x 76.3mm x 9.1mm, making it a reasonably sized
            smartphone",
        "Weight": "193g, providing a solid feel without being overly heavy",
        "Display Size": "6.52 inches with a TFT display that offers a resolution of
            720 x 1600 pixels",
        "Pixel Density": "269ppi, delivering decent clarity for everyday usage",
        "Design": "Waterdrop notch design adds a modern touch; available in colour
            options of Aqua Blue and Ice Jadeite"
    },
    "Battery and Performance": {
        "Battery Capacity": "5000 mAh, ensuring long-lasting performance between
            charges",
        "Battery Type": "Li-ion, which is standard for smartphones in this range",
        "Charging Options": "No wireless charging and no fast charging feature, but
            provides adequate battery life throughout the day",
        "Hardware": {
            "Chipset": "MediaTek Helio A20",
            "RAM": "2 GB, sufficient for basic tasks and everyday applications",
            "Internal Storage": "32 GB, expandable up to 256 GB via microSD card,
                allowing for additional storage for apps and media"
        }
    }
}
```

Listing 14: Output generated with StructLM trained with eC-Tab2Text.

```
{
    "Software and Connectivity": "Tecno Spark Go 2020 will run on Android Go UI on
        top of Android 10 out of the box. The smartphone will offer connectivity
        options like Dual SIM VoLTE, Wi-Fi 802.11, b/g/n, GPS, and Bluetooth 4.2.
        The device will also come with a rear-mounted fingerprint sensor and a 3.5mm
         headphone jack.",
    "Design and Display": "The Tecno Spark Go 2020 is expected to feature a 6.52
        inch (16.56 cm) display with a screen resolution of 720 x 1600 pixels. The
        display will come with a waterdrop notch design and a screen-to-body ratio
        of 40.97 %. The device will come in Aqua Blue and Ice Jade colors.",
    "Battery and Performance": "The Tecno Spark Go 2020 will be powered by a
        MediaTek Helio A20 processor clocked at 2GHz and a PowerVR GE8300 GPU. The
        device will come armed with 2GB RAM and an internal storage capacity of 32GB
        , which can be expanded upto 256GB via a microSD card."
}
```

Listing 15: Output generated with Llama2 trained with eC-Tab2Text.


```
{
    "Battery and Performance": "The Tecno Spark Go 2020 is anticipated to be powered
         by MediaTek\u2019s latest entry-level Helio A20quad-core chipset coupled
        with a PowerVR GE8300 GPU. That will again reportedly be paired with 2 GB
        RAM and an internal storage capacity of only 32 GB, which can further be
        expanded up to 256GB. Further, the device will reportedly pack in a 5000mAh
        Li-ion battery but won\u2019t support fast charging.",
    "Cameras": "The Tecno Spark Go 2020 is expected to house a dual-camera setup on
        the back with a 13MP (Digital Zoom) camera as the primary sensor. Besides,
        there will also be a 2MP depth sensor onboard. On the front, the smartphone
        will supposedly feature an 8MP selfie shooter. There will also likely be a
        bunch of camera features such as Artificial Intelligence,Auto Flash,Auto
        Focus,Bokeh Effect,Continuos Shooting,Exposure compensation,Face detection,
        Geo tagging,High Dynamic Range mode (HDR),ISO control,Touch to focus,White
        balance presets.",
    "Design and Display": "The Tecno Spark Go 2020 will reportedly feature a 6.52
        inch TFT panel tipped with a resolution of 720 x 1600 pixels. The pixel
        density will supposedly max out at 269ppi. The bezel-less display is further
         anticipated to boast a waterdrop notch design to furnish an immersive
        viewing experience."
}
```