Facts-and-Feelings: Capturing both Objectivity and Subjectivity in Table-to-Text Generation

Anonymous ACL submission

Abstract

Table-to-text generation, a long-standing challenge in natural language generation, has remained unexplored through the lens of subjectivity. Subjectivity here encompasses the comprehension of information derived from the table that cannot be described solely by objective data. Given the absence of pre-existing datasets, we introduce the **Ta2TS dataset** with **3849 data instances**. We perform the task of fine-tuning sequence-to-sequence models on the linearized tables and prompting on popular large language models. We analyze the results from a quantitative and qualitative perspective to ensure the capture of subjectivity and factual consistency. The analysis shows the fine-tuned LMs can perform close to the prompted LLMs. Both the models can capture the tabular data, generating sound texts with **85.15%** BERTScore and **26.28%** Meteor score. To the best of our knowledge, we provide the first kind of dataset on tables with multiple genres and subjectivity included and present the first comprehensive analysis and comparison of different LLM performances on this task.

1 Introduction

In the contemporary era of big data, humongous volumes of information are being generated and archived in numerous different formats. Among these, tables stand out as a very useful way of storing structured data. Using the potential of natural language to comprehend tabular data holds the promise of enhancing human efficiency across diverse applications. Text generation from tables has seen some contributions previously (See Section 2). These efforts have predominantly centred on either the generation of text from relatively simple tabular structures or the translation of numerical data into natural language, devoid of nuanced interpretation. In this research, we propose a subjective point of view to look at the table-to-text generation problem statement.

Table:

Particulars	*2019*	*2020*	*2021*	*2022*	*2023*
Revenue	2,462	2,365	3,634	4,316	4,197
Expenses	2,084	2,080	2,088	2,701	2,955
EBITDA	918.7	819.37	1,936	2,139	1,896
EBIT	894.75	779.66	1,889	2,091	1,838
Profit Before Tax	377.9	285.19	1,458	1,616	1,242
Net Profit	285.25	215.4	1,203	1,311	932.82
Financial Ratios		None	None	None	None
Operating Profit Margin (%)	37.32	34.64	53.28	49.55	45.18
Net Profit Margin (%)	11.59	9.11	33.10	30.37	22.23
Earnings Per Share (Diluted)	19.67	12.2	84.21	88.38	62.66
Dividends Per Share	8.5	4	10	10	10

Reference Text:

The company's income statement shows its financial performance in the past 5 years with the revenue showing a **minor dip** in 2020 touching 2,365 Cr. Rs. and then **increasing steadily** until 2022 to 4,316 Cr. Rs. The year 2023 has witnessed another **slight decline** in revenues, throwing light on the company's slow growth. Expenses have increased by **greater margins** 2022 onwards reaching 2,701 Cr. Rs. and thus, resulting in decreased profit margin. Positive EPS in 2021-2022 **signals profitability** per share which is **crucial for shareholders**. The company's commitment to maintain a **constant** value of 10.00 Rs. of dividends per share demonstrates an **overall balance** in the company. Thus, the company statistics reflect a **steady growth** with recent slowdown which hints at a **decline of performance** and operational efficiency. Investors with future prospects with this company should hold cautious **optimism**.

Figure 1: Generating text with subjectivity from a table: The table contains the income statement report of a company over 5 years. The reference text below describes the tabular information where the bold phrases refer to the infused subjectivity. This subjectivity is the interpretation of the data from a human perspective. (Due to space constraints, shown in figure 4)

1.1 Problem Statement

We introduce a novel problem statement with the aim of text generation from tabular data, devoted to infusion of subjectivity into the generated text. Subjectivity, in the context, refers to the nuanced interpretation of the data, beyond the realm of numeric or objective representation.

The table depicted in Figure 1 presents a com-

pany's income statement table with 10 features such as Revenue, Expenses, EBIDTA, EBIT, Profit before taxes, Net profit, Financial ratios, Operating profit margin, Net profit margin, Earnings per share and Dividends per share respectively. Within the accompanying narrative of the table, several expressions are employed, which do not find direct representation in the objective data. Instead, these phrases refer to interpreting the underlying sentiment associated with the data.

The phrase ***minor dip*** comes from the idea that having a downfall of 200 Cr. rs. is **small** compared to a growth of 2000 Cr. rs. over the last few years. Whereas, ***signals profitability*** comes from the understanding that positive EPS refers to positive growth [1].

1.2 Problem Formulation

The problem of table-to-text generation takes a table T as input where $T = \{t_{1,1}, .., t_{i,j}, .., t_{m,n}\}$ where $t_{i,j}$ refers to the token of ith row and jth column with at most m rows and n columns. A pre-processed input document Y is produced from the table where Y can be written as, $Y = y_1 y_2 .. y_l$. Here y_i is the ith token in Y which has l tokens and $l \geq mn$. In an auto-regressive approach text $S = s_1 s_2 ... s_k$ is generated from Y which can be defined as, $s_i = argmax P(s_i|Y, s_1, s_2, ..., s_{i-1}; \theta)$ where θ is model parameters.

1.3 Motivation

Text generation from tables can be useful in various social applications where a mere description of the numerical values proves to be insufficient.

- News or Blog writings on a Tournament points table, Pricing table, or Voting results table.
- Reports on a Match summary, Business details, Sales detail of a company.
- Explaining Healthcare reports, Weather reports, or understanding Legal documents, etc.

In recent years Table-to-text generation has seen significant contributions (section 2). However, the domain of exploration has been very limited to sports and Wikipedia tables mostly. Along with the perspective of subjectivity, which can add a human touch to the generations; we explore different genres of tables, which is necessary to make comprehensive systems for this task.

The advancement of LLMs has contributed largely to text generation, but not so much in the case of tabular data. A study revealing the potential possibility of using LLMs for this task is also necessary in the current scenario.

1.4 Our Contributions

Our contributions are:

1. the formulation of a problem statement (Section 1.1) of subjective text generation from tables, enriched with social significance. To the best of our knowledge, this is the first-of-its-kind effort.
2. the **Ta2TS** (pronounced as Tattoos and expanded as Table-2-Text with Subjectivity and Objectivity) dataset for this task comprising **3849 table instances** sourced from sports, financial statements, and weather forecast domains (Section 3).
3. analysis and comparison of the **performance** of different **LLMs** on the dataset, enabling insight into LLMs' capability of maintaining factual consistency, coverage, and semantic similarity with reference for table-to-text generation task (Section 4.4).
4. showing that regular LMs perform close to LLMs. This we do by developing a T5-based fine-tuned sequence-to-sequence learner which has the highest Bleu-4 and BERTScore on test Ta2TS dataset, **beating GPT3.5** and the other LLMs (Section 5.1).

2 Related Works

The table-to-text problem contains several different approaches based on the type of table. The **Wikibio-infobox** problem (Lebret et al., 2016) involves generating textual descriptions from tables with just one column extracted from Wikipedia infoboxes. Versatile neural language models and encoders were introduced to tackle this challenge by Liu et al. (2017) and Rebuffel et al. (2022).

The **ToTTo** dataset (Parikh et al., 2020a) represents another prominent objective of text generation from a single highlighted row of the table. Gehrmann et al. (2021) harnessed the T5 model to address this challenge. Subsequently, sequence-to-sequence techniques and structure-aware frameworks were introduced by Parikh et al. (2020b) and Wang et al. (2022). The current state-of-the-art on this dataset has been held by Kale and Rastogi (2020), who leveraged the T5 model with 3 billion parameters to achieve the highest scores. A very similarly aligned problem, **Wiki-table-to-text**, has

[1] The table example is taken from Groww official website

also been worked on by Chen et al. (2021).

Among the previously discussed problem statements, none encompass the task of comprehending complicated tables with multiple rows and columns. The **Rotowire** dataset (Wiseman et al., 2017) is expressly designed to address this challenge. The concepts of macro-planning, content selection, and planning have been significant contributions to this field by Puduppully and Lapata (2021), Puduppully et al. (2019). Li et al. (2021) achieved state-of-the-art results by introducing a Record encoder, a Reasoning module, and a Decoder equipped with Dual attention. Additionally, alternative approaches have also been explored by Rebuffel et al. (2019), (Gong et al., 2019) and Choi et al. (2021). It is worth noting that the RotoWire tables exhibit a notable degree of similarity to the tables of our interest. The flattened tables derived from data instances within RotoWire can contain up to 17,000 tokens. Most of the current language model architectures do not support such a huge context length. Thus, using pre-trained models for this task is very challenging. While training a model from the ground up may address this challenge, there are concerns regarding its comprehensive understanding of the language. Zhao et al. (2023) explored the **performances of LLMs** in table question-answering tasks with comparative analysis. Similar studies in text generation from tables could be useful for prospects.

3 Dataset

We contribute to building a novel dataset for this problem statement which is named **Ta2TS Dataset** (Table-to-Text with Subjectivity), pronounced as *Tattoos*.

3.1 Data Collection and Filtration

The dataset encompasses three types of tables, such as financial statement tables, weather forecast tables, and sports tables. We collected the tables using various web-based platforms that are entitled to showcase different data tables. Five sites were selected as sources of the tables such as Groww, Indian Meteorological Dept., ESPN Cric Info, IPL, and Goal (see A.2 for sources). The tables are scraped using Python Programming language. A total of 4200 tables were collected from the sources by web scraping. The collected data has been filtered manually. Tables with fewer rows and columns than a threshold are removed. Also, data instances having very low cell values are discarded. Such as, some of the collected tables had zero or NaN values. After filtration, the number of instances in the dataset stands at 3849. The complete dataset statistics are in Table 2. The dataset contains arguably an unbiased distribution between different genres, shown in Figure 6.

3.2 Annotation Details

We employ three annotators for this task who are proficient in English. One of the annotators is a graduate of Computer science and engineering while the others are PhD students and postgraduates in English literature respectively. They were provided with detailed annotation guidelines with rules, objectives, and multiple examples. Each instance in the set was annotated only once. The distribution among the annotators was 1150, 1300, and 1399 instances. An example annotation is shown in figure 3 (See A.4). 100 samples were annotated by all three annotators which are used to compute the **Inter-Annotator Agreement** (IAA). The annotators were asked to mark the other annotators' annotations [2] based on the correctness of factual information and appropriate use of subjectivity out of 10 (1 being the lowest score and 10 being the highest). Also, pairwise BERT-based semantic similarity has been captured to emphasize the similarity in the annotation tasks. BERTScore ensures the contextual meaning of the sentences. We present the harmonic mean of these two metrics as the Inter-Annotator Agreement. We have also evaluated pairwise Bleu-2 score for all the annotators which is mentioned in the appendix (section A.4).

	A	**B**	**C**	**D**
A		8.74	8.64	8.43
B	8.56		8.99	8.42
C	8.69	8.84		8.67
D	8.33	8.69	8.99	

Table 1: Inter-Annotator Agreement: We compute the IAA using two metrics. Pairwise scoring of each annotation based on factual consistency and capture of subjectivity and calculation of pairwise average BERTScore. The table represents the harmonic mean of these two metrics.

[2] Annotator A is asked to mark annotators B and C's annotations for the given features.

Genre	No. of Instances	Avg. no. of rows	Avg. no. of columns
Finance			
Income statement	500	11	5
Balance Sheet	499	11	5
Cash flow	499	11	5
Weather forecast			
District-wise tables	1031	5	10
Sports			
Tournament Points Table	409	9.4	7.2
Series Form Table	941	5	6
Total	**3849**		

Table 2: Genres in the Ta2TS Dataset: The Ta2TS dataset comprises three primary genres, namely, financial tables, weather forecast tables and sports tables. In this table number of tables from each domain and different types are shown.

4 Experimental Setup

We performed experiments on different versions of the T5 model (Raffel et al., 2020). T5 is a sequence-to-sequence model trained on C4 data with prefixes. We understand that the metadata associated with each table (company, team or district name, forecast range etc.) can be an essential component of the text generation. We fine-tuned the T5 models with different sets of prefixes as a part of our ablation study to understand the importance of table context for the model.

4.1 Linearization of Tables

Transformer-based sequence-to-sequence models are capable of understanding linearized information with long context. A closer examination of the reference texts reveals that the tabular content contains two distinct types of information, namely, row-wise and column-wise comparison. For example, the income-statement table depicts phrases that express the relative growth of the company in time which requires row-wise comparison and relation between revenue-EBIDTA or other features which requires column-wise comparison. Hence, the application of row-major flattening along with a column separator token is required. In this linearization, each cell is separated by a **<sep>** token and rows start and end with **<row>** and **</row>** tokens respectively. An example of linearization is shown in section A.5.

4.2 Sequence-to-sequence Model

We use the T5 (Raffel et al., 2020) models to perform the task. T5 is a transformer-based sequence-to-sequence learner. Multiple versions of the T5 model such as T5-base, T5-large, and T5-3b are used for the experiments. We used three different setups with respect to prefixes,

1. Without any table context or prefix.
2. With the prefix *"Summarize the table: "*
3. With contextual prefix such as *"Summarize the table about Kolkata district of West Bengal state, during the time range 01-Jan-2024 to 06-Jan-2024: "*

The overall architecture with linearization, prefix and model usage is shown in figure 2.

4.3 Training Details

The T5 models are fine-tuned on different settings. The dataset has been divided into an 80-10-10 train-test-validation split. Using five-fold cross-validation we trained the T5-base, T5-large and T5-3b models on 30, 24 and 20 epochs respectively. The number of epochs has been chosen by minute observation of the change in loss and metrics over the epochs. We used a learning rate of $2e-5$ with a batch size of 4. On average, the whole fine-tuning for the base model took around 6 hours, for large around 10 hours and for 3b for 17 hours. Adafactor optimizer has been used with a learning rate of $1e-3$ and with a decay rate of 0.8.

4.4 Prompting on LLM

Along with T5 results, we perform prompting experiments on various state-of-the-art Large language models. Comparing and analysing the results we try to establish a baseline on this dataset. With the experiments on LLM, we fundamentally try to answer three different questions.

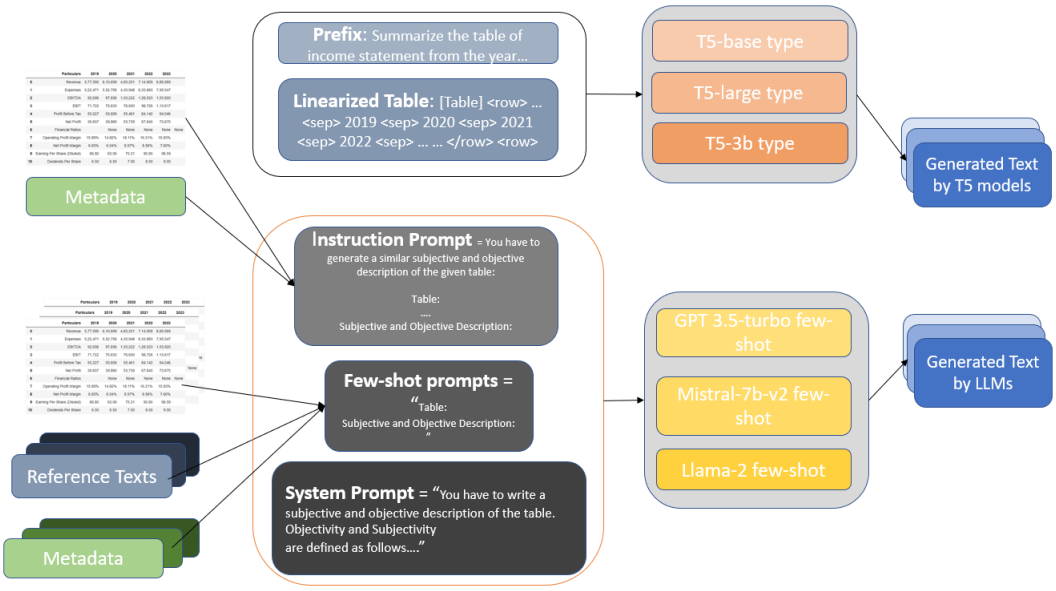

Figure 2: Architecture: The diagram shows two parallel methods of experimentation we adopted. In the top experiment, the tables are linearized (section 4.1). Metadata is the genre of the table, company or district name, time and additional details. Different types of prefixes are given based on the context (section 4.2). Different T5 models are used with different prefix types. In the bottom experimentation, some popular LLMs are prompted. A system prompt is generated by formally defining the task. Multiple examples are given in few-shot methods along with the metadata and reference text. Finally generated texts by all of these methods are compared and analysed.

1. How well do LLMs perform in understanding human perception behind the tabular information?
2. Can LLMs maintain factual consistency through the generation tasks?
3. What is the future of table-to-text generation through generic LLMs?

Generic LLM here refers to the models which are generally used for a wide variety of tasks and not specifically designed for a particular task, such as ChatGPT, Mistral etc. To answer the questions raised above, we have prompted on four Language models, namely, **GPT-3.5-turbo** (Speed, 2024), **Mistral-7b** (Jiang et al., 2023), and **Llama-2** (Touvron et al., 2023). The experiments include 0-shot, 1-shot and few-shot (up to 7 shots).

We experimented with different prompts with changing definitions of subjectivity and instruction. We divide the prompt into three sections. *System prompt* is the part where we define the task and provide the necessary definitions to the LLM (see the figure 2). In the *few-shot prompt* section, we add the example instances with notation of table and description. In the last *instruction prompt* section, the input table with instructions to write description text is given. Our experiments focused on the system prompt section mostly. The tables are given in input with separator tokens and proper indentation. An example of a table and prompt is given in Figure 8 (section A.6).

5 Results

We measured the generated texts of the fine-tuned models and LLMs through automatic and human evaluation methods. The automatic metrics are explained below.

5.1 Quantitative Analysis

The three different prefixes are denoted by the following notations.

1. **w/o-pf**: Without any table context or prefix.
2. **pf-w/o-ct**:With the prefix *"Summarize the table: "*
3. **pf-ct**: With contextual prefix such as *"Summarize the table about Kolkata district of West Bengal state, during the time range 01-Jan-2024 to 06-Jan-2024: "*

5.1.1 Evaluation Metrics

Parent-FScore Parent-Fscore is F1 score that assesses the overall generation in the domain of table-to-text (Dhingra et al., 2019). It uses entailment-based precision and recall to analyse the factual consistency in the generated texts.

Model	Parent-Fscore	Bleu-4	Meteor	Rouge-L	BertScore
T5-base, w/o-pf	0.44	2.89	22.17	21.11	80.43
T5-base, pf-w/o-ct	0.44	3.22	24.60	22.97	84.14
T5-base, pf-ct	0.45	**3.25**	24.98	23.45	**85.15**
T5-large, w/o-pf	0.47	2.81	23.54	21.98	81.43
T5-large, pf-w/o-ct	0.47	2.94	23.65	22.31	82.22
T5-large, pf-ct	0.48	3.20	24.11	**23.67**	84.37
T5-3b, w/o-pf	0.41	2.57	20.11	21.39	81.58
T5-3b, pf-w/o-ct	0.42	2.63	21.77	22.64	82.00
T5-3b, pf-ct	0.44	2.67	22.03	22.78	82.28
GPT3.5, 0-shot	0.51	2.64	25.3	21.9	81.70
GPT3.5, 1-shot	0.52	3.08	**26.28**	23.02	83.87
GPT3.5, 3-shot	**0.54**	2.98	25.97	23.57	84.78
GPT3.5, 7-shot	0.51	2.48	25.29	23.24	84.42
Mistral-7b, 0-shot	0.42	2.21	18.09	13.65	30.01
Mistral-7b, 1-shot	0.42	2.48	18.76	18.11	40.83
Mistral-7b, 3-shot	0.43	2.69	19.26	18.76	29.64
Mistral-7b, 7-shot	0.42	2.08	20.17	13.92	29.87
Llama-2, 0-shot	0.45	2.17	17.44	19.32	56.71
Llama-2, 1-shot	0.46	2.31	18.13	19.45	59.00
Llama-2, 5-shot	0.46	2.84	19.65	19.88	63.14

Table 3: Metrics Table: We computed Parent-Fscore (Dhingra et al., 2019), Bleu-4 score (Papineni et al., 2002), Meteor score (Banerjee and Lavie, 2005), Rouge-L score (Lin, 2004) and BERTScore (Zhang et al., 2020) to analyse the generations. The highest scores in each metric are highlighted in bold and the second highest is highlighted with underline. We can see that the results of our T5 model and comparable with GPT3.5 and they together outperformed the rest. Parent score is shown out of 1 and the rest of the metrics are shown out of 100.

Bleu-4 Score It is an n-gram precision-based metric (Papineni et al., 2002). B-4 calculates precision up to 4 grams and represents the lexical match between reference and machine-generated text.

Meteor Score It is an alignment-based metric which uses lexical matching, stemming-based matching, synonyms etc (Banerjee and Lavie, 2005). It also penalizes unaligned words. Meteor represents a good characteristic for judging content and fluency.

Rouge-L Score It is a recall-based metric which measures the longest common sub-sequence between the machine-generated text and reference text (Lin, 2004). It emphasizes on how much content is covered from the reference text.

BERTScore This metric is based on contextual embeddings of the BERT model (Zhang et al., 2020). BERTScore represents the similarity in the reference text and generated text.

The results are showcased in table 3. We can see that T5 and GPT3.5 outperformed all the other models. The T5-base (pf-ct) achieved the highest Bleu-4 score of **3.25** and the highest BERTScore of **85.15**. On the other hand, T5-large (pf-ct) achieved the highest Rouge-L score of **23.67** and the second-highest Bleu-4 of **3.20**. BERTScore signifies the similarity of generated text in the embedding space and B-4 represents the lexical content overlap.

As a part of the ablation study, we fine-tuned the T5 models with different prefix settings. As we can see from table 3, the *pf-ct* setting (section 5.1) had the highest performances in both T5-base and T5-large. Hence, we conclude that adding the context in prefixes helps generate more relevant texts to the ground truth. For T5-3b, we estimate that it is overfitting on the training data. The generated texts from the 3b model lacked coverage significantly and sometimes hallucinating too.

GPT3.5 significantly outperformed all other large language models. With a 3-shot setting, it achieved the highest Parent-fscore and the second-highest scores in Meteor, Rouge-L and BERTScore. The generations covered all necessary information from the tables, had fluency, and coherence and provided significant similarity with the ground truth too. Mistral-7b and LLama-2 performed poorly on all the settings with similar prompts. For simple text generation, we understand that 3-shot prompt-

ing outperforms 7-shot. The shorter context helps the model understand the task and semantics of ground truth soundly. The B-4 scores are significantly low yet the generated texts represent a good summary of the table. We explain this by the property of synonyms. The T5 models and LLMs understand the language phenomena soundly. Hence, it can use different words or expressions to express similar meanings. Hence, the lexical overlap can be low. However, a similar meaning refers to close enough values of context vectors in BERT embedding space, which in turn is reflected in high BERTScores. An example of generated text is shown in Figure 3.

5.2 Qualitative Analysis

Qualitative Evaluation has been performed over 100 randomly chosen samples. The evaluators are asked to mark the generations out of 10 (1 being the lowest and 10 being the highest), based on four parameters, namely, coherence, coverage, accuracy, and subjectivity capture.

Coherence is defined as the logical flow of information in the text, where each sentence or phrase follows the previously generated phrase.

Coverage is defined as ensuring the inclusion of all the necessary information from the table. Necessary, although depends on different perspectives, we have specifically defined to the evaluators. There can be information in the table which is co-related to some other information (can be inferred from other information), we call it dependent information. For instance, when we say that a team has played 16 matches and won 10 while losing 4, it is obvious that 2 matches were tied. The evaluators are asked to ignore the dependent information and judge based on only the independent ones.

Accuracy is defined as the factual correctness of the objective information present in the table. It is noteworthy that the accuracy metric will not penalize coverage, hence it will judge only the information in the generated text.

Subjectivity capture is defined as the correctness of subjective phrases generated in the text and the inclusion of subjective phrases for every objective information.

The human evaluation scores on GPT3.5, 3-shot are shown in Table 4. The evaluators analysed each instance on the four features described above. To summarise scores, we report the harmonic mean of all the features as it represents a more refined differentiation than the arithmetic mean. The scores showcase that in all of the given features, the generations reflect soundness and quality. The relatively lower coherence is caused due to some unattended co-reference resolution in the text, which reduces the flow of reading to some extent. The high subjectivity capture score reflects on use of correct adjectives at the right places throughout the text.

	Coh.	Cov.	Acc.	Sub. Cap.
A	8.7	9.0	8.8	9.1
B	8.4	8.8	8.9	9.0
C	8.9	9.0	9.1	9.1
H-Mean	8.66	8.93	8.93	9.06

Table 4: Human evaluation on GPT3.5 Generated Instances: the average score of all 200 generated samples over the features of Coherence (Coh.), Coverage (Cov.), Accuracy (Acc.) and Subjectivity Capture (Sub. Cap.). We report the summary of scores by harmonic mean of all the evaluators.

The human evaluation scores of T5-large (pf-ct) generated instances, as shown in Table 5, indicate a lower quality compared to texts generated by GPT. Specifically, the coverage score for T5-generated texts is significantly lower than that of GPT-generated texts. Upon closer examination, it becomes apparent that many of the generated samples leave some objective facts unattended, which could have enhanced their impact on readers. Despite this, T5-generated texts achieve relatively higher scores in Accuracy and Subjectivity capture, indicating a factually consistent presentation of objective data and an appropriate emotional interpretation of the objective information context. However, some issues with co-reference resolution persist in these texts like GPT, which results in lower coherence scores.

	Coh.	Cov.	Acc.	Sub. Cap.
A	8.3	8.2	8.7	8.8
B	8.1	8.0	8.5	9.0
C	8.4	8.3	8.8	9.0
H-Mean	8.26	8.16	8.66	8.93

Table 5: Human evaluation on T5-large-pf-ct Generated Instances: The average score of all 200 generated samples over the features of Coherence (Coh.), Coverage (Cov.), Accuracy (Acc.) and Subjectivity Capture (Sub. Cap.). We report the summary of scores by harmonic mean of all the evaluators.

6 Conclusion and Future Work

In this paper, we present a novel perspective on the domain of table-to-text generation by introducing

Particulars	2019	2020	2021	2022	2023
Revenue	249.79	339.87	558.00	1,153	1,488
Expenses	189.99	260.65	410.08	908.05	1,206
EBITDA	70.71	93.98	171.19	284.31	342.38
EBIT	60.61	80.65	151.56	251.87	292.96
Profit Before Tax	59.80	79.22	147.92	244.80	281.55
Net Profit	48.82	65.52	135.04	214.69	245.47
Operating Profit Margin	28.35%	28.16%	33.13%	26.29%	23.88%
Net Profit Margin	19.57%	19.63%	26.13%	19.85%	17.12%
Earning Per Share (Diluted)	20.10	26.13	10.59	16.18	18.43
Dividends Per Share	0.00	0.00	0.00	0.00	0.00

Reference Text

The table depicts the highs and lows of the company in the past 5 years in terms of its financial numbers. It has been performing remarkably well in maintaining its operational efficiency over the years which hints at the fact that the company has been judiciously investing on its strategies and management plans. Though the operating profit margin is on the rise, it has not been able to maintain an equilibrium growth in its net profit margin too. In the year 2023, net profit margin faced a severe low of 17.12%. As a result, it is currently declining and needs to be making gradual improvements at the same pace to ensure the stability of the company and to gain a good trust of its prospective investors.

T5-large (pf-ct) Generated Text

The financial report of the company shows a significant increase in revenue from 249.79 million to 1,488 million over the years. Expenses also increased from 260.65 billion to 1.206 million, while EBITDA ranged from 306.31 million. Ebit and profit before tax exhibited some fluctuations, with the highest values in the third year. Net profit margin grew from 19.57% to 26.13%, indicating efficient cost management. The earnings per share (diluted) increased significantly from 26.13 to 18.43, reflecting positive growth for shareholders. Overall, the financial performance showed both positive and negative trends but remained above 2021 and 2022 respectively.

GPT3.5 (3-shot) Generated Text

The table presents financial data for the years 2019 to 2023, showing a significant increase in revenue from 249.79 to 1,488. Expenses have also grown, resulting in positive EBITDA and EBIT figures. The profit before tax and net profit have shown a consistent growth trend, with net profit reaching 245.47. The operating profit margin fluctuated over the years, reaching a peak in 2021 at 33.13%. The net profit margin has also varied, but overall showing an improvement. Earnings per share have fluctuated, indicating some variability in profitability per share. It is noteworthy that no dividends were distributed during these years, suggesting reinvestment of profits back into the business.

Figure 3: Example of texts generated by T5-large (pf-ct) and GPT 3.5 (3-shot): In the figure the given table and reference text (ground truth) are taken from the Ta2TS dataset and at the bottom the generated texts are shown.

the element of subjectivity, an unexplored dimension of this field for several different table genres. To facilitate our goal, we curate the **Ta2TS** Dataset with tables and respective subjective-objective ground truth texts. We conduct fine-tuning and prompting experiments to assess the effectiveness of various models in capturing subjectivity and factual objectivity. The evaluations indicate that the models exhibit a robust understanding of word knowledge, with subjective phrases appropriately reflected in the generated text. Our fine-tuned T5 models perform comparably with the GPT3.5-turbo (few-shot prompted) model. Also, we provide a to-the-best-of-our-knowledge first comprehensive analysis of LLM performances in the table-to-text generation task. We proposed three questions regarding the use of LLMs in table-to-text generation (section 4.4) which we try to answer here. The detailed quantitative and qualitative evaluations reveal that the generated texts are sound in terms of most of the evaluation parameters. LLMs do understand the structure of tabular data and the relation between different cells. The texts remain consistent factually throughout, indicating reliability in the objective information. Also, it has higher coverage than the fine-tuned models. Hence, we conclude the effectiveness of LLMs in table-to-text generation.

7 Limitations

Our dataset contains tables from three genres. Subjectivity in the same type of table can be influenced by similar semantics, which may not be the case for a more diverse dataset. Hence, building a robust system to handle all kinds of tables will be more feasible with an even more diverse or large dataset. Moreover, we developed a table flattening technique to fit tables in sequence-to-sequence models which are trained on normal text. For prospects, we propose ideas to fuse pre-trained table encoders with decoders and fine-tune our dataset to get even better metric scores.

8 Ethics Statement

All of our collected data were present in open-source mediums and did not contain personal, restricted, or illegal data. None of the generated texts or annotated texts were intended to promote or derogate any team or entity.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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.

Sanghyuk Choi, Jeong in Hwang, Hyungjong Noh, and Yeonsoo Lee. 2021. May the force be with your copy mechanism: Enhanced supervised-copy method for natural language generation.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Li Gong, Josep Crego, and Jean Senellart. 2019. Enhanced transformer model for data-to-text generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 148–156, Hong Kong. Association for Computational Linguistics.

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

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

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

Liang Li, Can Ma, Yinliang Yue, and Dayong Hu. 2021. Improving encoder by auxiliary supervision tasks for table-to-text generation. 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 5979–5989, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2017. Table-to-text generation by

structure-aware seq2seq learning. In *AAAI Conference on Artificial Intelligence*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

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

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text Generation with Macro Planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Clement Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2022. Controlling hallucinations at word level in data-to-text generation. *Data Min. Knowl. Discov.*, 36(1):318–354.

Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2019. A hierarchical model for data-to-text generation.

Ann Speed. 2024. Assessing the nature of large language models: A caution against anthropocentrism.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Fei Wang, Zhewei Xu, Pedro Szekely, and Muhao Chen. 2022. Robust (controlled) table-to-text generation with structure-aware equivariance learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5037–5048, Seattle, United States. Association for Computational Linguistics.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. 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.

A Appendix

A.1 Examples of Subjective Text Generation from Tables

Figure 4 and figure 5 showcase some more examples of data instances from the **Ta2TS dataset**. These tables are collected from Sports and Weather forecast domains respectively.

A.2 Data Collection

The tables of Ta2TS Dataset have been scraped from 5 sites namely, Groww [3], Indian Meteorolog-

[3] https://groww.in/

Table:

TEAMS	M	W	L	T	PT	Series Form
Delhi Daredevils	16	11	5	0	22	WLWLW
Kolkata Knight Riders	16	10	5	0	21	WWLLW
Mumbai Indians	16	10	6	0	20	WLWWL
Chennai Super Kings	16	8	7	0	17	LWWWL
Royal Challengers Bangalore	16	8	7	0	17	LWLWW
Kings XI Punjab	16	8	8	0	16	LWLWW
Rajasthan Royals	16	7	9	0	14	LLWLW
Deccan Chargers	16	4	11	0	9	WWLLL
Pune Warriors	16	4	12	0	8	LLLLL

Reference Text:

In the league, Delhi Daredevils have showcased remarkable performance, grabbing 11 wins out of 16 matches and proving dominance with 22 points. Kolkata Knight Riders follow closely with 10 wins and 21 points. Mumbai Indians secured 10 wins as well, with a slightly fluctuating form. The competition has been intense for the teams like Chennai Super Kings and Royal Challengers Bangalore, both securing 17 points each. Kings XI Punjab, Rajasthan Royals and Deccan Chargers also fail to display promising performance with 16, 14 and 9 points each. Pune Warriors faced an uphill battle with only 8 points and securing bottom spot.

Figure 4: Generating text with subjectivity: The table shows different teams and their relative performances in a tournament. The table is described by the reference text, enriched with subjectivity. It is an example instance table from the sports domain of the **Ta2TS dataset**. In the table the abbreviations denote the following; M:Matches; W:Wins; L:Losses; T:Ties; PT:Points.

ical Dept. [4], ESPN Cric Info [5], IPL [6] and Goal [7] which are mentioned in the section 3.1.

A.3 Ta2TS Dataset Statistics

The dataset contains three different kinds of tables such as financial-statement tables, weather forecast tables, and sports tables. The financial statements tables are of different companies that are listed on the National Stock Exchange, India. There are three kinds of financial tables for each company, namely its income statement, balance sheet, and cash flow tables. On the other hand, the weather forecast data is collected from the official site of Govt. of India. These tables present data for each district in each state. Statistics of the dataset have been shown in table 2 and figure 6. As mentioned in section 3.1, a total of 3849 tables are collected with an average of nearly 9 rows and nearly 7 features.

A.4 Dataset Annotation

The data instances have been manually annotated which is mentioned in section 3.2. The annotators were given a detailed instruction set to follow and were paid a total of 1.20$ per data instance for annotation. The annotation task was conducted through regular monitoring over two months, following a fixed set of instructions. We asked the annotators to write at most 150 words for each instance. We try to keep the length of the reference text invariable for a variable number of rows and columns.

We define the annotation task in three steps: (1) Information extraction from the table, (2) Subjectivity infusion, and (3) Text generation. In step 1, the annotators are expected to extract the objective information in the table and understand its dynamics, changes, and relations. Furthermore, they should understand the emotions lying with the variables in the data and infuse subjectivity into the information. In the last step, they should write a text combined with the previous semantics which describes the table. The annotation details and guidelines are described in section 3.2. The pairwise Bleu-2 scores are shown in table 6.

	A	B	C	D
A		27.13	32.49	29.33
B			23.53	34.18
C				28.58
D				

Table 6: Bleu-2 for Inter-Annotator Agreement: Pairwise Bleu-2 scores are calculated for annotators.

A.5 Table Linearization

Linearization of tables is crucial when it is given as input in seq-to-seq models. We used a tagged concatenation of delimiter-separated row-major ordering to generate a flattened sequence of tables. The row-major orders are initiated with tags **<rows>** and end with **</rows>**. Whereas, each column in row-major order is separated by **<sep>** token. The intuition behind the linearization is described in section 4.1.

A row-major flattening of the table in figure 7 can be written as, *[TABLE] <row> Teams <sep> Matches <sep> Wins <sep> Losses <sep> ... <sep> Series Form </row> <row> Nottinghamshire <sep> ... <SEP> ... LLWLL </rows>*.

A.6 Prompting on LLMs

Large language models are prompted to generate objective and subjective text from the given tables. We use three different sections in the whole prompt. The first section is **system prompt** which refers to

[4] https://mausam.imd.gov.in/imd_latest/contents/agromet/advisory/indiadistrictforecast.php

[5] https://www.espncricinfo.com/

[6] https://www.iplt20.com/

[7] https://www.goal.com/en-in

Table:

Forecast Date	Valid From	Valid Till	Rainfall (mm)	TempMax (deg C)	TempMin (deg C)	Humidity-I (%)	Wind Speed (m/s)	Wind Direction (deg)	Cloud Cover (octa)
0	2024-02-07	2024-02-06	2024-02-07	0	28	20	38	10	248
1	2024-02-08	2024-02-07	2024-02-08	0	27	19	23	11	335
2	2024-02-09	2024-02-08	2024-02-09	0	26	17	17	12	9
3	2024-02-10	2024-02-09	2024-02-10	0	25	16	14	12	347
4	2024-02-11	2024-02-10	2024-02-11	0	26	16	16	11	342

Reference Text:

The data from the table shows the variation in weather forecast for consecutive 5 days in the month of February. There has been no rainfall recorded between 7th of February to 11th of February. While both the maximum and minimum degree of temperature is dropping, the region is slowly pacing towards colder days and nights. The humidity pattern demonstrates a mixed variation with a straight fall from 70% to 29% in the first three days, indicating even minor chances of rain. The weather is overall comfortable and pleasing. People in this region witness clear skies and an overall amiable weather. Significant cloud cover was observed on 11th of February which warns of the advent of scattered drizzles. The area experiences moderate breezes making the environment more conducive to the people. On a general note, the weather is fairly enjoyable, but it would be good if precautionary measures are adopted to counteract sudden showers or temperature drip.

Figure 5: Generating text with subjectivity: The table contains data on the weather forecast in the district of Kolkata from 6th February 2024 to 11th February 2024. The subjective reference text describes the table. This example data instance is taken from the **Ta2TS dataset**.

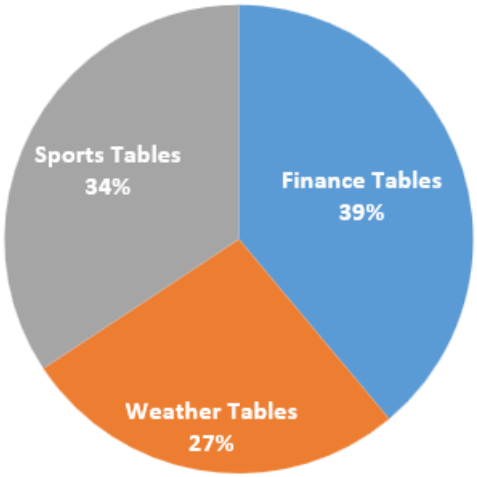

Figure 6: Distribution of Genres in Ta2TS Dataset: The chart depicts the distribution of different genres in the dataset. 34% of the tables are from the sports genre, 39% are from the finance genre and 27% are from the weather genre.

	0	1	2	3	4	6	7	8
0	TEAMS	M	W	L	T	PT	NRR	Series Form
1	Nottinghamshire	10	7	1	0	16	1.315	WLWWW
2	Lancashire	10	5	3	0	12	-0.250	LLWWL
3	Leicestershire	10	4	3	0	11	-0.180	WWLLW
4	Durham	10	4	5	0	9	0.421	LWWWW
5	Yorkshire	10	3	5	0	8	0.297	WLLLL
6	Derbyshire	10	1	7	0	4	-1.583	LLWLL

Figure 7: Table flattening: extracting information from row-major and column-major ordering

the initial description of the task along with necessary definitions. Followed by **few-shot prompt**, where multiple examples of tables and description texts are given. This section can be repeated several times depending on the shot numbers. In the last section, i.e., **instruction prompt**, we add the input table along with a little instruction about what has to be performed. An example of prompt has been shown in the figure 8.

You have to write a subjective and objective description of the table. Objectivity and Subjectivity are defined as follows.
Objectivity: It refers to the factual information given in the table, like numerical values or data.
Subjectivity: It refers to the human interpretation of the data. Subjective texts have proper use of adjectives to denote the trends in data or impact of the data on reader.

Here are a few examples of tables and corresponding descriptions.

Table:

Particulars|2019|2020|2021|2022|2023|
|---|---|---|---|---|---|---|
Revenue|5,77,595|6,10,699|4,83,251|7,14,909|8,89,569|
Expenses|5,22,471|5,52,756|4,33,948|6,33,883|7,95,547|
EBITDA|92,656|97,836|1,03,222|1,28,523|1,53,920|
EBIT|71,722|75,633|76,650|98,726|1,13,617|
Profit Before Tax|55,227|53,606|55,461|84,142|94,046|
Net Profit|39,837|39,880|53,739|67,845|73,670|
Financial Ratios||||||
Operating Profit Margin|15.89%|14.82%|18.11%|16.21%|15.83%|
Net Profit Margin|6.83%|6.04%|8.97%|8.56%|7.60%|
Earning Per Share (Diluted)|66.80|63.06|75.21|90.85|98.59|
Dividends Per Share|6.50|6.50|7.00|8.00|9.00|

Subjective and Objective Description:

The overall summary of the income mentioned above shows a consistent increase in revenue over the years, reaching a peak of 7,89.569. Expenses have also increased but at a slower rate, resulting in a positive EBITDA of 1,53.920. EBIT and profit before tax have shown steady growth, with net profit reaching 73.670. The operating profit margin has remained relatively stable around 15-18%, while the net profit margin has improved slightly over the years. Earnings per share have also increased, indicating positive performance. Dividends per share have also increased steadily, reflecting the company's profitability.

.....

You have to generate a similar subjective and objective description of the given table:

Table:

Particulars|2019|2020|2021|2022|2023|
|---|---|---|---|---|---|---|
Revenue|548.39|707.45|756.15|865.68|1,166|
Expenses|494.47|494.38|364.13|462.99|756.74|
EBITDA|58.05|235.85|412.08|407.79|410.50|
EBIT|53.92|214.70|396.02|395.84|379.43|
Profit Before Tax|53.92|211.42|389.35|390.44|371.28|
Net Profit|20.04|147.38|279.77|297.62|283.83|
Financial Ratios||||||
Operating Profit Margin|11.44%|36.92%|61.54%|54.12%|41.66%|
Net Profit Margin|3.95%|23.07%|41.78%|39.50%|28.80%|
Earning Per Share (Diluted)|7.61|50.24|94.58|96.84|92.64|
Dividends Per Share|0.00|10.00|15.00|2.00|20.00|

Subjective and Objective Description:

Figure 8: Prompt Example: An example prompt to the LLMs is given in the figure. The three sections represent system prompt, few-shot prompt and instruction prompt explained in section 4.4. The few-shot section is repeated several times depending on the shot numbers.