A Comparative Study of Language Models for Chart Summarization

An Chu[1,2], **Thong Huynh**[1,2], **Long Nguyen**[1,2,*], **Dien Dinh**[1,2]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
Correspondence: nhblong@fit.hcmus.edu.vn

Abstract

This paper investigates the potential of state-of-the-art Large Language Models— Mistral, Starling-LM, Gemma-1.1, Llama-2 and its variant Llama-3 —in the context of chart summarization. We evaluate their performance on established datasets supplemented by our datasets from Our World in Data designed to address potential gaps. Methodologically, we delve into the architecture of each baseline model and any task-specific modifications. The experimental setup covers training processes, hyperparameter tuning, and specific configurations used for evaluation. Results highlight the models' performances across our datasets, offering insights into their strengths and weaknesses. The discussion interprets findings, exploring implications for real-world applications. This study concludes by emphasizing the pivotal role of these models in advancing chart summarization, providing valuable insights for practitioners, and suggesting promising directions for future research.

1 Introduction

Chart Summarization stands at the intersection of natural language processing and visual data comprehension, playing a critical role in extracting meaningful insights from visual representations like charts and graphs (Hoque et al., 2022). In an era where data-driven decision-making is paramount (Kim et al., 2020), understanding and querying information presented in visual formats have become integral across various domains (Hoque et al., 2017).

While advancements in Natural Language Processing (NLP) have led to the development of powerful models (Masry et al., 2022; Ishwari et al., 2019; Namazifar et al., 2021; Demszky et al., 2018), applying these techniques to the unique challenges posed by chart summarization remains an ongoing research frontier. Previous studies have addressed a variety of tasks, yet challenges persist in adapting state-of-the-art NLP models to effectively summarize based on visual data.

Central to the progress of chart summarization are the datasets employed for model training and evaluation. Datasets not only provide the foundation for model development but also serve as a benchmark for gauging the performance of different approaches. Understanding the intricacies of these datasets is crucial for uncovering the potential of state-of-the-art models in handling the nuances of chart-based queries.

Previous research has made notable strides in chart summarization, yet significant gaps persist. The focus of many studies has been on enhancing optical character recognition (OCR), neglecting the broader challenges posed by diverse datasets and varied chart types. This paper addresses these gaps by emphasizing the importance of comprehensive datasets and shedding light on the challenges faced by previous studies. By doing so, we aim to contribute valuable insights that go beyond OCR enhancements.

In our exploration, we leverage state-of-the-art baseline models, including Mistral-7B (Jiang et al., 2023), Starling-LM-7B (Zhu et al., 2023), Gemma-1.1-7B (Team et al., 2024), Llama-2-7B (Touvron et al., 2023), and Llama-3-8B (AI@Meta, 2024). Each of these models was selected for their unique strengths and capabilities. Mistral-7B is recognized for its superior performance and efficiency, leveraging grouped-query attention (GQA) and sliding window attention (SWA) for faster inference and handling sequences effectively, making it a highly efficient model (Jiang et al., 2023). Starling-LM-7B excels with an 8.09 MT Bench score, backed by the robust Nectar dataset and RLAIF techniques, enhancing its helpfulness and safety (Zhu et al., 2023). Gemma-1.1-7B stands out for its compact size and remarkable performance, utilizing grouped-query attention and sliding window attention to outperform larger models

in reasoning, math, and code generation (Team et al., 2024). Llama-2-7B, with its large parameter count and training on a diverse corpus, excels in language understanding, generation, and reasoning benchmarks (Touvron et al., 2023). Llama-3-8B showcases advancements in performance, safety, and helpfulness, with extensive training on over 15 trillion tokens, outperforming previous Llama models and ensuring enhanced helpfulness and reduced false refusals (AI@Meta, 2024). By understanding how these models navigate the challenges posed by diverse datasets, we hope to provide a nuanced perspective on their potential in overcoming the hurdles presented by various chart types and data representations. The code and dataset used in this study are available at https://github.com/chuducandev/ChartQA.

2 Related Works

The landscape of Chart Summarization has evolved significantly in recent years, reflecting the broader advancements in NLP and visual data comprehension. Early studies in chart summarization focused on foundational challenges, including optical character recognition (OCR) (Kim et al., 2022; Kavehzadeh, 2023) and basic question interpretation (Kim et al., 2020; Masry et al., 2022). However, as the field matured, researchers recognized the need for more sophisticated approaches to handle the complexities of diverse chart types and data representations (Li and Tajbakhsh, 2023).

Early efforts in chart summarization predominantly revolved around planning-based architecture (Mittal et al., 1998; Ferres et al., 2013) and two stage approach that applied content selection using different statistical tools in the first step followed by generating summaries using pre-defined templates (Reiter, 2007; Zhu et al., 2021). Nevertheless, despite their focus on elucidating the critical insights communicated by the chart, these systems often fall short in furnishing lucid instructions for interpretation.

In previous years, both commercial platforms and academic projects have significantly advanced the field of Chart Summarization. Notable examples include Narrative Science Quill and Automated Insights Wordsmith (Caswell and Dörr, 2018), alongside research initiatives, e.g., (Cui et al., 2019) and (Srinivasan et al., 2018), which have all made strides in extracting and presenting key data insights through the computation of statistical measures such as extrema and outliers. Similarly, the work (Demir et al., 2012) stands out for its innovative approach to generating bar chart summaries. This method employs a bottom-up strategy that intricately weaves together discourse and sentence structures, effectively summarizing data trends. Moreover, a pioneering approach (Chen et al., 2019) leverages the ResNet architecture (He et al., 2016) to encode chart images. This process is complemented by an LSTM-based decoder that meticulously crafts captions, showcasing the integration of deep learning techniques to enhance data visualization interpretation.

In the realm of Chart-To-Text summarization, the field has progressively moved from template-driven methods towards more nuanced data-driven approaches, underscored by the introduction and evolution of several pivotal datasets. The sequence began with the Chart2Text dataset (Obeid and Hoque, 2020), offering an initial collection of 8,305 chart samples from Statista. This dataset, although groundbreaking, was limited by its size, posing challenges for the training of comprehensive data-driven models. Subsequently, (Spreafico and Carenini, 2020) deployed an LSTM-based encoder-decoder model on a smaller dataset of 306 chart summaries, a step that, while innovative, still did not fully leverage the visual aspects of charts. Furthermore, efforts to diversify and enrich the data landscape saw the introduction of the SciCAP dataset (Hsu et al., 2021) focused on chart image captioning, and the AutoChart dataset (Zhu et al., 2021) which utilized predetermined templates for generating chart descriptions. These advancements highlighted the constraints of fixed templates, such as reduced variability and insight in the generated summaries.

In recent advancements, our work aligns with significant contributions such as ChartSumm (Rahman et al., 2023) and Chart-To-Text (Kantharaj et al., 2022), focusing on advancing interpretability through summarization methodologies. While ChartSumm focuses on automatic chart-to-text summarization, catering primarily to visually impaired individuals and facilitating precise insights of tabular data in natural language, Chart-To-Text contributes a large-scale dataset with chart images, metadata, and corresponding human-written descriptions, addressing the task of generating textual descriptions from visual data. In contrast, our work diverges by concentrating on fine-tuning state-of-the-art models and enriching datasets to enhance

chart understanding and interpretation. Through this approach, we aim to advance interpretability, leveraging sophisticated techniques tailored to handle diverse chart types and data representations. By contextualizing our contributions within this framework, we seek to bolster the repertoire of NLP techniques for deriving insights from visual data.

3 Methodology

3.1 Dataset Construction

To conduct our research on fine-tuning large language models for chart summarization, we curated a comprehensive dataset from Our World in Data (Roser et al., 2015). This platform provides empirical evidence on global issues such as poverty, health, and education. We manually collected charts and their corresponding summaries and metadata, focusing on relevant countries and structuring the information into a comprehensive data table. Each chart was then accompanied by a concise and informative summary generated using the GPT-4 (OpenAI, 2023) language model. To ensure the quality and accuracy of the summaries, a team of human annotators reviewed each output, verifying the correctness of facts and numbers, and assessing the coherence and clarity of the summaries (Huang, 2012).

Through this comprehensive data collection and curation process, we have successfully generated a dataset consisting of 5,166 charts, each accompanied by a concise and accurate summary. This dataset, derived from the authoritative Our World in Data platform, covers a wide range of subjects and provides a solid foundation for our research on fine-tuning large language models for chart summarization. By leveraging this carefully constructed dataset, we aim to advance the state-of-the-art in automated chart analysis and contribute to the development of more effective tools for understanding and communicating complex data (Lai et al., 2020).

The distribution of chart types within the Our World in Data dataset showcases a predominance of line charts, accounting for 60.2% of the total charts. Bar charts follow as the second most common chart type, representing 20%. Additionally, the dataset includes bubble charts (0.4%), scatter plots (9.6%), and area charts (9.6%), highlighting a variety of visualization techniques employed.

The topic distribution in the Our World in Data dataset covers a broad range of global issues, with

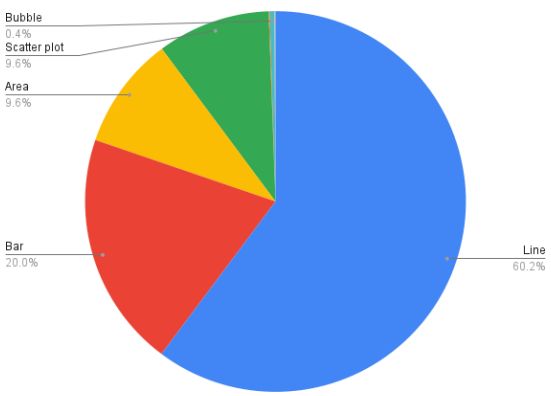

Figure 1: Chart type distribution of Our World in Data dataset

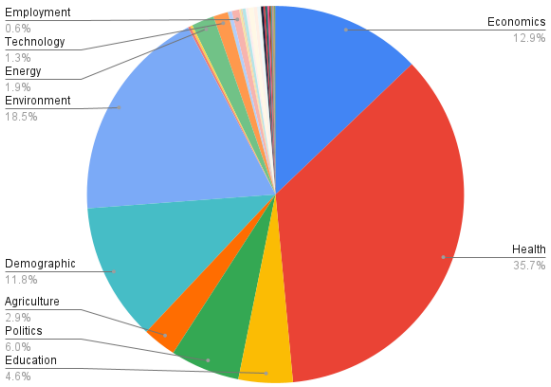

Figure 2: Topic distribution of Our World in Data dataset

health being the most prominently represented theme at 35.7%. This is followed by environment (18.5%), demographics (11.8%), economy (12.9%), politics (6.0%), education (4.6%), technology (1.3%), and energy (1.9%). This dataset serves as an invaluable resource for researchers and practitioners interested in exploring and understanding various global trends and patterns, providing insights into key areas such as health, environment, and economy.

3.2 Language Model Fine-tuning Process

In this segment, we introduce the foundational models employed to assess performance within our designated dataset, followed by an outline of the fine-tuning procedure.

3.2.1 Baseline Models

We provide an overview of the state-of-the-art language models used as baselines.

LLAMA-2 (Touvron et al., 2023) - 7B, developed

by Meta AI, excels in language understanding and reasoning with 7 billion parameters. It's open-sourced, enabling wide exploration and innovation in AI.

LLAMA-3 (AI@Meta, 2024) - 8B, advances performance and safety using SFT and RLHF techniques. It outperforms Llama-2 models, with a focus on safety and helpfulness in AI interactions.

STARLING-LM (Zhu et al., 2023) achieves high scores in MT Bench, leveraging the Nectar dataset and RLHF. It enhances the reliability and performance of fine-tuned models.

MISTRAL (Jiang et al., 2023) - 7B, developed by Anthropic, outperforms larger models like Llama-2 with its efficient architecture, excelling in reasoning, math, and code generation.

GEMMA-1.1 (Team et al., 2024) model by Google DeepMind, a compact 7B model, surpasses larger models in reasoning, math, and code generation, showcasing Google's commitment to responsible AI.

3.2.2 Fine-tuning Process

In this study, we explore the fine-tuning process for several state-of-the-art language models, including Llama-2, Llama-3, Starling-LM, Mistral, and Gemma-1.1. The fine-tuning methodology is crucial for adapting these models to the specific task of summarization with chart data.

For all five models, we employed a consistent fine-tuning approach using 3 epochs of training. This decision aimed to ensure a fair comparison across the models and maintain a balance between performance and computational efficiency.

The fine-tuning process involved the use of specific prompts tailored to each model. These prompts, fully demonstrated in Appendix A.2, were designed to guide the models in understanding the task at hand and generating appropriate responses based on the chart content. By incorporating these prompts into the fine-tuning process, we aimed to provide the models with clear instructions and context for generating accurate and relevant summaries based on the chart content.

Through the fine-tuning process, we sought to leverage the pre-trained knowledge of these language models while adapting them to the specific task of summarization with chart data. By carefully tuning the models on our curated dataset and utilizing tailored prompts, we aimed to enhance their ability to understand and generate accurate responses based on the visual information presented in charts.

The fine-tuning methodology employed in this study serves as a critical component in optimizing model performance for the task at hand. By dedicating computational resources and implementing a consistent training approach across all models, we strive to unlock the full potential of these state-of-the-art language models in the context of chart summarization.

4 Evaluation

In this section, we present a comprehensive evaluation of the fine-tuned models' performance on the chart summarization task. Our evaluation methodology encompasses two key components: automated benchmarks and human evaluation. The automated benchmarks provide quantitative measures of the models' performance, while the human evaluation offers qualitative insights into the generated summaries' quality and coherence. By combining these two approaches, we aim to deliver a holistic assessment of the models' capabilities and limitations in the context of chart summarization.

4.1 Evaluation Metrics

In assessing the quality of our automated summarization, we employ a comprehensive set of evaluation metrics to capture various aspects of the generated summaries. Our evaluation framework encompasses the following key metrics:

BLEU Score (Bilingual Evaluation Understudy) evaluates the overlap of n-grams between the model-generated summaries and the reference texts (Post, 2018). We compute BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores to capture different levels of n-gram overlap, providing insights into the linguistic fidelity and structural alignment of the generated summaries with the references.

BLEURT Score (Bilingual Evaluation Understudy with Representations from Transformers) is a model-based metric designed to assess the fluency and semantic fidelity of generated text (Sellam et al., 2020). Leveraging BLEURT-base-128, we evaluate the grammatical correctness and semantic alignment of the machine-generated summaries with respect to the reference documents.

PPL (Perplexity) serves as a metric to quantify the predictive performance of language models (Rad-

ford et al., 2019). Lower perplexity scores means more coherence and contextual relevance of the generated summaries.

4.2 Automated Benchmarks

Our study evaluates the performance of several state-of-the-art language models on the task of summarization with chart data. Table 1 summarizes the experimental results obtained from these models across various evaluation metrics. Notably, the Gemma-1.1 model leads in BLEU-1 with a score of 54.15, while the Starling-LM model performs slightly lower with a BLEU-1 score of 54.12 but surpasses in BLEU-2, achieving the highest score of 37.98. The Llama-3 model stands out with the highest BLEURT score of 0.1832, indicating superior semantic similarity, and also has the lowest perplexity (PPL) at 7.7889, suggesting it generates the most fluent and coherent summaries among the evaluated models.

Overall, the experimental results highlight the competitive performance of the Gemma-1.1 model in terms of BLEU-1, indicating its ability to generate summaries with high unigram precision. The Starling-LM model achieves the highest BLEU-2 score, demonstrating its strength in generating summaries with high bigram precision. Both models exhibit identical performance for BLEU-3 and BLEU-4. The Llama-3 model stands out with the highest BLEURT score and the lowest PPL value, suggesting its superiority in generating semantically similar and fluent summaries.

These results provide valuable insights into the strengths and weaknesses of each model in the task of summarization with chart data. The Gemma-1.1 and Starling-LM models demonstrate strong performance in terms of n-gram precision, while the Llama-3 model excels in semantic similarity and fluency. Further analysis and experimentation may be necessary to investigate the factors contributing to these differences in performance and to validate the findings across different datasets and chart types.

4.3 Human Evaluation

To complement the automated benchmarks, we conducted a human evaluation to assess the quality of summaries generated by different models. This evaluation involved a total of 750 pair-wise comparisons across 50 samples randomly selected from the test dataset. Four human annotators evaluated the summaries based on three criteria: **factual correctness**, **coherence**, and **fluency** (Kantharaj et al., 2022).

After collecting the results, we used the Elo rating system to comprehensively evaluate the models' performance. The Elo rating system calculates the expected score E_A for a model with rating R_A when matched against an opponent with rating R_B using the formula:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}. \quad (1)$$

The model's new rating R'_A is then updated based on the match outcome using the following formula:

$$R'_A = R_A + K \cdot (S_A - E_A), \quad (2)$$

where K is the K-factor (a constant determining the sensitivity of the rating system), S_A is the actual score from the comparison (1 for a win, 0.5 for a draw, and 0 for a loss), and E_A is the expected score as calculated earlier (Elo, 1978). In our study, we adapted the Elo rating system with a K-factor of 4 and an initial rating of 1000, providing a clear comparative analysis across the three criteria. The results are summarized in Table 2.

Among the models, Llama-3 consistently achieved the highest Elo ratings across all factors, making it the strongest performer in our evaluation. Its particularly high ratings in coherence and fluency indicate its ability to generate summaries that are both logically consistent and readable, closely approaching the quality of reference summaries.

On the other end of the spectrum, Starling-LM and Llama-2 demonstrated the weakest performance, with the lowest ratings in coherence and factual correctness, respectively. Starling-LM's struggles across multiple dimensions suggest a need for further optimization, while Llama-2's low factual accuracy points to potential challenges in interpreting the data correctly.

These Elo ratings highlight the varying strengths and weaknesses of each model, emphasizing the competitive performance of advanced models like Llama-3, while also indicating areas where other models require further improvement. Detailed pairwise comparison results are included in Appendix A.1 for additional context.

4.4 Factual Correctness Analysis

In addition to the overall human evaluation, we conducted a specific analysis focused on Factual

Models	BLEU-1 (↑)	BLEU-2 (↑)	BLEU-3 (↑)	BLEU-4 (↑)	BLEURT (↑)	PPL (↓)
LLAMA-2	53.22	36.79	27.03	20.00	0.1355	7.9861
MISTRAL	53.5	37.42	27.78	20.78	0.1252	8.0027
Starling-LM	54.12	**37.98**	**28.29**	**21.23**	0.1380	8.0097
GEMMA-1.1	**54.15**	37.95	**28.29**	**21.23**	0.1434	7.9000
LLAMA-3	53.61	37.64	28.12	21.15	**0.1832**	**7.7889**

Table 1: Model performance comparison based on BLEU, BLEURT, and PPL

Models	Correctness	Coherence	Fluency
Gold Text	1163	1048	1035
LLAMA-3	1034	1021	1026
GEMMA-1.1	934	1015	982
MISTRAL	983	970	1006
Starling-LM	981	965	964
LLAMA-2	906	982	987

Table 2: Elo ratings for models based on human evaluation

Correctness to assess how accurately each model represents the information in the charts. This analysis was based on the factual correctness factor from the human evaluation results, where we assumed the gold texts were entirely accurate in terms of factual information, as they were carefully annotated during the data construction phase. For GPT-4, the result was derived by counting the number of items in the dataset that were validated as fully factually correct during the dataset construction step.

Models	Correctness (%)
GPT-4	86.2
LLAMA-3	63.5
GEMMA-1.1	57.5
MISTRAL	45.5
STARLING-LM	47.5
LLAMA-2	43.0

Table 3: Percentage of entirely factually correct summaries generated by each model.

The results, as shown in the table above, indicate a notable gap between the performance of the fine-tuned models. Among these, Llama-3 achieved the highest correctness rate at 63.5%, outperforming the other open models such as Gemma-1.1 at 57.5%, and Mistral and Starling-LM at 45.5% and 47.5%, respectively. Llama-2 had the lowest percentage of factually correct summaries at 43.0%.

While there is a clear gap between these models, further work is needed to explore potential improvements and optimizations. The relatively strong performance of Llama-3 highlights its potential as a leading model in this category, although there is still room for enhancing factual correctness across all open models.

Future work could focus on closing the gap between these models, refining their ability to generate factually accurate summaries, and bringing them closer to the performance exhibited by proprietary models.

4.5 Alignment Between Automated Benchmarks and Human Evaluation

This subsection examines the alignment between the automated benchmarks and human evaluation results, providing a clearer picture of each model's strengths and weaknesses in chart summarization.

The Llama-3 model shows strong consistency across both evaluation methods. It achieved the highest BLEURT score and lowest perplexity, indicating superior semantic fidelity and fluency, which aligns with its top Elo ratings in Coherence and Fluency during human evaluation. This suggests that Llama-3 consistently generates high-quality, coherent, and fluent summaries, as recognized by both automated metrics and human judgment.

Similarly, the Gemma-1.1 model performed well in both evaluations, with strong BLEU and BLEURT scores and a respectable perplexity score. Its high Elo rating in Coherence reinforces the idea that it produces accurate and coherent summaries, making it a reliable choice for text generation tasks.

However, the Starling-LM model reveals a discrepancy between its strong BLEU scores and lower Elo ratings in human evaluation, particularly in Coherence and Fluency. This suggests that, while Starling-LM excels in n-gram overlap with reference texts (as indicated by BLEU), it may lack the deeper coherence and fluency that human evaluators value. This highlights the limitations of relying solely on automated metrics like BLEU, which may not fully capture the quality of the generated text.

Overall, these findings emphasize the importance of combining automated benchmarks with human evaluation to obtain a comprehensive assessment of model performance, as each method captures different aspects of text quality.

4.6 Error Analysis and Challenges

Upon examining 100 generated summaries for the two cases presented in Table 4, several errors and challenges become apparent. These issues high-

Case (a)	Case (b)
Gold: The area chart illustrates the trend in the number of granted patents related to artificial intelligence by industry worldwide from 2010 to 2020. There has been a significant increase in patent grants across all industries over this period. The 'Personal devices and computing' field saw the highest growth, with patents rising from 727 in 2010 to 13,813 in 2020. Notably, every recorded field experienced growth, but the 'Life sciences' and 'Telecommunications' sectors also displayed prominent increases, with patents climbing to 2,290 and 4,848, respectively, in 2020. Businesses, transportation, and energy management sectors also had considerable numbers of patent grants, with 3,218, 1,679, and 1,318 patents granted in 2020. Among the least growth was seen in the 'Banking and finance' sector, with 751 patents in 2020, which still reflects substantial growth from the 31 patents in 2010. Overall, the chart shows an accelerating interest and investment in AI across various industries, with the sharpest overall increase occurring between 2016 and 2018. (Figure 3)	**Gold:** (...) For Belarus, there is a significant increase in the number of stored accessions over the years, starting from 1,144 in 1995 to a remarkable 27,494 in 2021. (...) (Figure 4)
Gemma-1.1: (...) The 'Business' sector also experienced significant growth, with numbers increasing from 95 in 2010 to 3,218 in 2020. (...) **Llama-3**: (...) Notably, the 'Physical sciences and engineering' industry had the lowest number of patents in 2020 with 965, down from 965 in 2018. (...) **Mistral**: (...) The field of 'Telecommunications' consistently received the highest number of patent grants, with a significant increase from 727 in 2010 to 13,813 in 2020. (...) The 'Energy management' and 'Physical sciences and engineering' fields saw a rise until 2018, followed by a slight decline in 2020. (...) The 'Personal devices and computing' field, which had the fewest grants, saw a gradual increase over the years. The data indicates a growing interest and investment in AI technologies across various industries, with a particularly sharp increase in the last two years. **Starling-LM**: (...) The 'Transportation' field, while smaller in comparison, grew from 189 patents in 2010 to 1,679 in 2020. The 'Life sciences' field saw an increase from 470 patents in 2010 to 2,290 in 2020. The 'Banking and finance' category, although the smallest in terms of number of patents, grew from 57 in 2010 to 751 in 2020. The 'Industry and manufacturing' field had 1,255 patents granted in 2020, up from 109 in 2010. The 'Energy management' field had 1,318 patents in 2020, an increase from 492 in 2010. The 'Physical sciences and engineering' field grew from 364 patents in 2010 to 965 in 2020. The 'Security' field saw an increase from 107 patents in 2010 to 1,527 in 2020. (...) **Comment:** Model struggles to produce a useful summary and accurately extract relevant statistics from the chart.	**Llama-2**: (...) For Belarus, there has been a significant increase from 1,144 accessions in 1995 to a peak of 35,210 in 2010, followed by a slight decline to 27,494 by 2021. (...) **Comment:** Model hallucinated about the number of unique plant genetic samples stored in Belarus reaching its peak of 35,210 in 2010.

Table 4: Sample error analyses of outputs from all five training models. **Red** indicates hallucination errors, **Blue** indicates tokens that are resulted in factual errors in the model output and **Violet** indicates misleading semantics

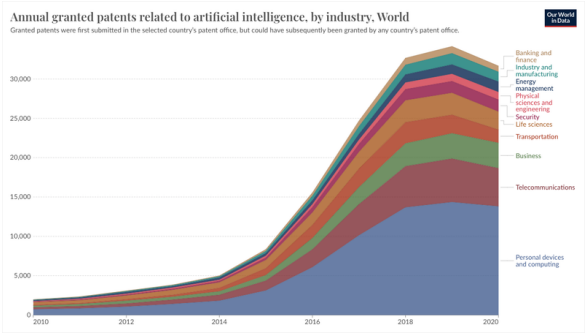

Figure 3: Case (a) - Artificial Intelligence Granted Patents By Industry

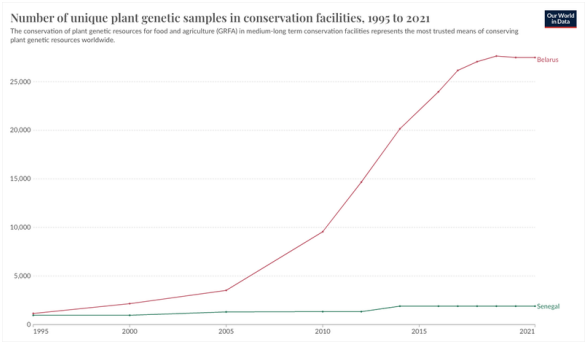

Figure 4: Case (b) - Number Of Accessions Of Plant Genetic Resources Secured In Conservation Facilities

light the difficulties faced by the language models in accurately understanding and summarizing the information conveyed in the charts.

In Case (a), the Gemma-1.1 model struggles to produce a useful summary and extract relevant statistics from the chart accurately. For instance, the model incorrectly states that the number of patent applications granted in the 'Business' sector increased from 95 in 2010 to 3,218 in 2020, whereas the correct values are 219 in 2010 and 3,218 in 2020. Similarly, the Llama-3 model makes an error in interpreting the data for the 'Physical sciences and engineering' industry, stating that the number of patents was down from 965 in 2018 to 965 in 2020, which is incorrect. The Mistral model also demonstrates several misinterpretations of the chart data, such as incorrectly claiming that the 'Telecommunications' field consistently received the highest number of patent grants and that the 'Personal devices and computing' field had the fewest grants.

In Case (b), the Llama-2 model hallucinates about the number of unique plant genetic samples stored in Belarus, stating that it reached a peak of 35,210 in 2010. However, this information is not supported by the chart data.

These errors and challenges in chart summarization can be attributed to several factors. Current models struggle with perceptual prowess, often missing subtle patterns and misinterpreting complex visual elements, as evidenced by the misinterpretation of trends and numbers in Case (a). Hallucinatory outputs occur when models generate false information not present in the input, leading to irrelevant or unsupported summaries, such as the hallucination in the Llama-2 model for Case (b). Data inconsistencies and training limitations result in models performing well on familiar data but faltering with less familiar formats, due to the broad variability in chart representations. Additionally, models excel at token-level predictions but struggle with maintaining semantically accurate summaries, leading to misleading information, such as Mistral's claims about the 'Telecommunications' field.

To address these challenges in chart summariza-

tion, several specific steps can be taken. First, curating larger and more diverse datasets covering various chart types and styles can help models generalize better. Second, developing more sophisticated model architectures that handle the nuances of visual data interpretation can reduce errors, possibly by integrating advanced vision-language models. Third, implementing grounding techniques to ensure outputs are closely tied to the input data can mitigate hallucinations by reinforcing the model's reliance on provided data. Continuously analyzing model outputs and feeding this information back into the training process can iteratively improve performance by identifying and refining common error patterns. Additionally, combining automated summarization with human oversight can enhance accuracy, as human reviewers can correct model outputs and provide additional training data (Kantharaj et al., 2022; Rahman et al., 2023; Moured et al., 2024).

By implementing these strategies, we can significantly improve the accuracy and reliability of language models in chart summarization.

5 Conclusion

In our study, we explored the effectiveness of state-of-the-art language models, including Llama-2, Llama-3, Starling-LM, Mistral, and Gemma-1.1, in summarizing chart data. Through a comprehensive fine-tuning process and tailored prompts, we evaluated their performance and identified the competitive results of the Llama-3 model, which achieved high BLEU scores, the highest BLEURT score, and the lowest perplexity value.

However, our analysis also revealed persistent challenges, such as perceptual limitations, hallucinatory outputs, and the need for improved data extraction methods. These challenges underscore the importance of continued research and development efforts to refine model architectures, diversify datasets, and explore novel approaches that integrate advances in natural language processing and computer vision.

The successful integration of summarization models with chart data holds immense potential for applications in data analysis, accessibility enhancement, and beyond. By addressing the identified challenges and building upon the strengths of the evaluated models, we can pave the way for more effective and efficient interactions between humans and machines in the realm of visual data comprehension.

Our study serves as a foundation for future research in this domain, providing valuable insights into the capabilities and limitations of state-of-the-art language models in summarization with chart data. We encourage further exploration and experimentation to push the boundaries of this field, ultimately contributing to the broader landscape of artificial intelligence and data science. By leveraging the strengths of these models and addressing the identified limitations, we can unlock new possibilities for data-driven decision-making and enhance the accessibility of visual information for a wider audience.

Acknowledgments

This research is supported by research funding from the Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

AI@Meta. 2024. Llama 3 model card.

David Caswell and Konstantin Dörr. 2018. Automated journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism Practice*, 12(4):477–496.

Chen Chen, Ran Zhang, Eunice Koh, Sangyoung Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. 2019. Figure captioning with reasoning and sequence-level training. *arXiv preprint arXiv:1906.02850*.

Zhiyuan Cui, Sunil K. Badam, Mehmet A. Yalçin, and Niklas Elmqvist. 2019. Datasite: Proactive visual data exploration with computation of insight-based recommendations. *Information Visualization*, 18(2):251–267.

Semir Demir, Sandra Carberry, and Kathleen F. McCoy. 2012. Summarizing information graphics textually. *Computational Linguistics*, 38(3):527–574.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Arpad Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York, NY, USA.

Leonel Ferres, Gitte Lindgaard, Linda Sumegi, and Becky Tsuji. 2013. Evaluating a tool for improving accessibility to charts and graphs. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(5):1–32.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Ehsan Hoque, Pooya Kavehzadeh, and Ahmed Masry. 2022. Chart question answering: State of the art and future directions. In *Computer Graphics Forum*, volume 41, pages 555–572. Wiley Online Library.

Ehsan Hoque, Vidya Setlur, Melanie Tory, and Ian Dykeman. 2017. Applying pragmatics principles for interaction with visual analytics. *IEEE transactions on visualization and computer graphics*, 24(1):309–318.

Tai-Yi Hsu, C. Lee Giles, and Tzu-Hao Huang. 2021. Scicap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264.

Ruopeng Huang. 2012. Enhancing survey quality through data quality assurance and quality control. In *International Conference on Social Science Methodology*, pages 257–270. Springer.

Karishma Ishwari, Aneeze Abdul Azeez, Sudheesan Sreedhar, Haritha Karunaratne, Arjuna Nugaliyadde, and Yassir Mallawarrachchi. 2019. Advances in natural language question answering: A review. *arXiv preprint arXiv:1904.05276*.

Antoine Q. Jiang, Alexandre Sablayrolles, Adam Mensch, Charles Bamford, Dhruv S. Chaplot, Diego de las Casas, Fabien Bressand, Gaél Lengyel, Guillaume Lample, and Lucas Saulnier et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Sharada Kantharaj, Ryan T. Leong, Xiaoran Lin, Ahmed Masry, Mihir Thakkar, Ehsan Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023.

Pooya Kavehzadeh. 2023. Chart question answering with an universal vision-language pretraining approach. Unpublished.

Do Hyun Kim, Ehsan Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13.

Gunwoo Kim, Taekyung Hong, Minki Yim, Jiyoung Nam, Jaewook Park, Junbeom Yim, Won Ik Hwang, Seunghyun Yun, Dongsoo Han, and Seungryong Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Zhaoyang Lai, Liang Yu, Shiqing Hu, and Xiaojun Chen. 2020. Automatic chart summarization. *arXiv preprint arXiv:2008.11223*.

Shujian Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.

Ahmed Masry, Xianglin Do, Jian Qiang Tan, Shafiq Joty, and Ehsan Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Vibhu Mittal, Johanna Moore, Giuseppe Carenini, and Steven F. Roth. 1998. Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24(3):431–477.

Omar Moured, Jinchao Zhang, Muhammad S. Sarfraz, and Rainer Stiefelhagen. 2024. Altchart: Enhancing vlm-based chart summarization through multi-pretext tasks. *arXiv preprint arXiv:2405.13580*.

Mina Namazifar, Alexandros Papangelis, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Language model is all you need: Natural language understanding as question answering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7803–7807.

Jocelyn Obeid and Ehsan Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, 2303.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever et al. 2019. Language models are unsupervised multitask learners.

Rezwan Rahman, Rezwana Hasan, Ashraful Farhad, Md Tahmid Rifat Laskar, Md Ashmafee, and Arjun Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. In *Proceedings of the Canadian Conference on Artificial Intelligence*.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 97–104.

Max Roser, Hannah Ritchie, and Esteban Ortiz-Ospina. 2015. Our world in data.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Antonio Spreafico and Giuseppe Carenini. 2020. Neural data-driven captioning of time-series line charts. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–6.

Arvind Srinivasan, Steven M. Drucker, Alex Endert, and John Stasko. 2018. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):672–681.

Gemma Team, Théo Mesnard, Chris Hardin, Raphaël Dadashi, Sriram Bhupatiraju, Sharada Pathak, Laurent Sifre, Matthieu Riviére, Megha S. Kale, and James Love et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikita Bashlykov, Shruti Batra, Pratik Bhargava, and Sandeep Bhosale et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Binyang Zhu, Eric Frick, Tong Wu, Haoyu Zhu, Kavitha Ganesan, Wei-Lin Chiang, Jing Zhang, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness & harmlessness with rlaif. Unpublished.

Jing Zhu, Jing Ran, Raymond K.-W. Lee, Zhenyu Li, and Kang Choo. 2021. Autochart: A dataset for chart-to-text generation task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1636–1644.

A Appendices

A.1 Pairwise Comparison Results

Table 5 presents the results of our human evaluation, which compares the quality of summaries generated by different models. The results are based on factual correctness, coherence, and fluency, highlighting which model performed better in each comparison. These comparisons provide insights into the strengths and weaknesses of the models in summarizing chart data.

A.2 Fine-Tuning Prompts

In this section, we provide the specific prompts used for fine-tuning the different language models in our study. These prompts were designed to guide each model in understanding the task of chart summarization and producing accurate summaries.

Llama-2, Mistral, and Starling-LM Prompts: The following prompt structure was used consistently across these three models.

```
<s>[INST] From the below input full content of a chart,
write a summary that reflects the meaning and trend of the
chart.
Chart content:
{sample['input']}[/INST]{sample['output']}</s>
```

Gemma-1.1 Prompt: For the Gemma-1.1 model, the prompt included user and model tags to structure the input and output more explicitly.

```
<bos><start_of_turn>user
From the below input full content of a chart, write a
summary that reflects the meaning and trend of the chart.
Chart content:
{sample['input']}<end_of_turn>
<start_of_turn>model
{sample['output']}<end_of_turn>
```

Llama-3 Prompt: The Llama-3 model used a prompt with specific header IDs and end-of-turn markers.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
From the below input full content of a chart, write a
summary that reflects the meaning and trend of the chart:
<|eot_id|> <|start_header_id|> user <|end_header_id|>
Chart content:
{sample['input']}<|eot_id|><|start_header_id|>assistant
<|end_header_id|>
{sample['output']} <|eot_id|> <|end_of_text|>
```

These prompts were tailored to each model's architecture to maximize their performance in chart summarization tasks.

	GEMMA-1.1 vs. LLAMA-2			GEMMA-1.1 vs. LLAMA-3			GEMMA-1.1 vs. MISTRAL		
Summary	Factual	Coherence	Fluency	Factual	Coherence	Fluency	Factual	Coherence	Fluency
First Wins	36.0%	29.5%	19.5%	17.0%	23.0%	12.5%	30.5%	31.0%	15.5%
Second Wins	43.0%	16.0%	11.0%	51.0%	26.0%	19.5%	42.0%	26.0%	19.5%
Tie	21.0%	54.5%	69.5%	32.0%	51.0%	68.0%	27.5%	43.0%	65.0%
	GEMMA-1.1 vs. STARLING-LM			GEMMA-1.1 vs. GOLD			LLAMA-2 vs. LLAMA-3		
Summary	Factual	Coherence	Fluency	Factual	Coherence	Fluency	Factual	Coherence	Fluency
First Wins	43.0%	35.5%	13.5%	23.5%	25.0%	12.0%	19.0%	24.0%	7.5%
Second Wins	32.5%	13.5%	10.0%	42.5%	21.5%	25.5%	57.0%	27.5%	19.0%
Tie	24.5%	51.0%	76.5%	34.0%	53.5%	62.5%	24.0%	48.5%	73.5%
	LLAMA-2 vs. MISTRAL			LLAMA-2 vs. STARLING-LM			LLAMA-2 vs. GOLD		
Summary	Factual	Coherence	Fluency	Factual	Coherence	Fluency	Factual	Coherence	Fluency
First Wins	24.0%	36.5%	12.0%	29.5%	26.0%	15.5%	19.5%	17.5%	5.0%
Second Wins	56.0%	18.0%	14.0%	46.5%	17.0%	12.0%	57.0%	41.0%	20.0%
Tie	20.0%	45.5%	74.0%	24.0%	57.0%	72.5%	23.5%	41.5%	75.0%
	LLAMA-3 vs. MISTRAL			LLAMA-3 vs. STARLING-LM			LLAMA-3 vs. GOLD		
Summary	Factual	Coherence	Fluency	Factual	Coherence	Fluency	Factual	Coherence	Fluency
First Wins	35.5%	44.5%	22.0%	37.5%	30.5%	21.5%	31.5%	25.0%	15.5%
Second Wins	38.5%	11.5%	6.0%	34.5%	12.0%	8.0%	36.5%	26.0%	18.0%
Tie	26.0%	44.0%	72.0%	28.0%	57.5%	70.5%	32.0%	49.0%	66.5%
	MISTRAL vs. STARLING-LM			MISTRAL vs. GOLD			STARLING-LM vs. GOLD		
Summary	Factual	Coherence	Fluency	Factual	Coherence	Fluency	Factual	Coherence	Fluency
First Wins	31.5%	38.0%	20.5%	26.5%	20.0%	13.5%	22.0%	23.0%	18.0%
Second Wins	39.0%	24.0%	11.0%	54.5%	33.5%	19.0%	52.5%	33.5%	23.5%
Tie	29.5%	38.0%	68.5%	19.0%	46.5%	67.5%	25.5%	43.5%	58.5%

Table 5: Human evaluation results for summary quality comparison among models.