문서 설명 기반 차트 생성 및 차트 추론 질의응답을 위한 한국어 데이터셋

파티마 페사란 자데[1], 임준영[1], 이신행[2], 이미연[2], 김채린[3], 김건희[1]

1 서울대, 2 더바이럴, 3 한샘가온

fatemehpesaran@vision.snu.ac.kr, junyoung.lim@vision.snu.ac.kr, data@the-viral.co.kr, lee@the-viral.co.kr, hanssemgaon.ai@gmail.com, gunhee@snu.ac.kr

Korean Dataset for Document-Description-Based Chart Generation and Chart Reasoning Question Answering

Fatemeh Pesaran Zadeh[1], Junyoung Lim[1], Shinhaeng Lee[2], Miyeon Lee[2], Chaerin Kim[3], Gunhee Kim[1]

fatemehpesaran@vision.snu.ac.kr, junyoung.lim@vision.snu.ac.kr, data@the-viral.co.kr, lee@the-viral.co.kr, hanssemgaon.ai@gmail.com, gunhee@snu.ac.kr

요 약

Recent progress in large language models and multimodal models has increased interest in chart generation and chart understanding, but Korean resources remain limited. We present a Korean two-task dataset for document-to-chart generation and chart reasoning question answering. Each instance aligns a Korean document description with a structured table, visualization code, a rendered chart image, and three reasoning QA pairs. The dataset contains 10,000 chart construction instances and 30,000 reasoning QA pairs across multiple chart types and reasoning categories. Experiments show that fine-tuning improves chart reasoning accuracy from 79.7% to 88.1% and chart-generation CodeBLEU from 0.08 to 0.94. These results suggest that the dataset is a practical benchmark and training resource for Korean chart-centered document understanding.

1. Introduction

Charts are a common way to present numerical information in reports, news articles, and public documents. By converting raw numbers into visual form, charts make it easier to compare values, identify trends, and understand relationships across categories. Accordingly, prior work has introduced chart understanding benchmarks such as DVQA, PlotQA, and ChartQA, which study question answering over charts with visual and logical reasoning [1]–[3]. More recent work has also explored explanation generation for chart reasoning [4]. However, existing chart benchmarks are largely built on English-language data [1]–[4].

This creates an important gap for Korean document understanding. In real Korean reports and public documents, numerical information is often expressed through language-specific formatting, units, category names, and writing styles that differ from English sources. As a result, models or benchmarks developed mainly on English chart data may not transfer well to Korean documents. Existing Korean work has mainly focused on table question answering, rather than chart generation and chart reasoning [5]. To the best of our knowledge, there is no publicly released Korean dataset that jointly supports document-to-chart generation and chart reasoning QA with aligned

intermediate representations such as document descriptions, structured tables, chart code, chart images, and multi-step reasoning annotations.

To address this gap, we present a Korean two-fold dataset for (1) document-to-chart generation and (2) chart reasoning question answering. Each instance is organized as a unified pipeline of document description → structured table → visualization code → chart image → reasoning QA. This design supports both generation and reasoning within a single benchmark and enables systematic evaluation of Korean language and multimodal models on chart-related document understanding tasks.

2. Dataset Construction

2.1 Task Definition

We build the dataset for two tasks. The first is **document-to-chart generation**, where the input is a Korean document description containing numerical information and the output is a chart. Depending on the evaluation setting, the target can be a structured table, chart code, or a rendered chart image. The second is **chart reasoning QA**, where the input is a chart and a question, and the output is an answer with intermediate reasoning steps. As shown in Figure 1, the two tasks are naturally connected: the chart generated from the document serves as the basis for downstream reasoning and question answering.

2.2 Data Collection and Filtering

We collected source documents from public reports, news articles, and statistical materials, and retained only documents with usable numerical descriptions. Duplicate and low-quality samples were removed during filtering. Each selected description was then converted into a structured table containing categories, values, and units. Based on the table, we generated chart visualization code and rendered the final chart image. Finally, we wrote three QA pairs with three-step reasoning for each chart. As a result, each example forms a single aligned pipeline: description → structured table → chart code → chart image → QA/reasoning.

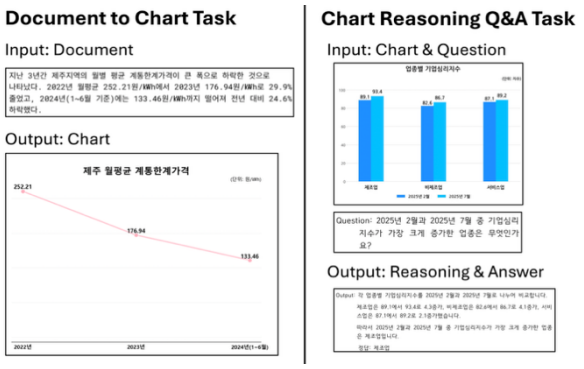

Figure 1. Qualitative example data points for each task that we propose in the dataset.

2.3 Data Processing Pipeline

Each example contains source metadata, chart annotations, and QA annotations. The chart annotation includes fields such as chart type, title, labels, unit, data values, and visualization code, while the QA annotation includes the question, reasoning type, three reasoning steps, and final answer.

To ensure quality, the dataset was constructed with expert-guided design and multi-stage review. According to the project report, external advisors supported the dataset and CoT design, and the data were inspected for labeling errors, numerical consistency, chart appropriateness, and the coherence of the question–reasoning–answer chain. Samples with errors were revised through repeated review. Since the dataset forms a single aligned pipeline from description to reasoning, quality assurance also emphasized cross-stage consistency among the description, table, chart, and QA annotations.

3. Results

Our dataset contains **10,000 document descriptions**, with aligned **10,000 structured tables**, **10,000 chart code instances**, **10,000 chart images**, and **30,000 QA-reasoning pairs** (Table 1). It covers five major chart types and two reasoning categories: 9,309 arithmetic and 20,691 logical reasoning instances. For experiments, we split the dataset into 80% train, 10% validation, and 10% test. The training split is used for fine-tuning, the validation split for model selection, and the test split for final evaluation.

We evaluate both data quality and downstream model performance. In Table 2, description–chart consistency measures whether the chart matches the source description in values, items, and units; reasoning appropriateness measures whether the reasoning steps correctly support the answer; and reasoning consistency measures whether the three-step CoT remains logically coherent throughout, where human experts go through the data and report the portion of consistent, appropriate data points. The resulting scores are 97.6%, 99.47%, and 99.5%, respectively, showing strong alignment across the full pipeline.

We evaluate chart reasoning QA using Qwen2.5-VL-7B-Instruct [6] and LLM-as-a-Judge Accuracy. To validate this automatic evaluation, we manually labeled 60 ChartQA samples and found 86.67% raw agreement with the LLM judge, supporting its use as a reasonable evaluation proxy. For document-to-chart generation, we evaluate Meta-Llama-3-8B-Instruct [7] using CodeBLEU, following prior chart-generation work such as Text2Chart31 [8], where it was used as a reliable automatic metric for generated visualization code. As shown in Table 3, fine-tuning improves chart reasoning accuracy from 79.7% to 88.1% and chart-generation CodeBLEU from 0.08 to 0.94.

Table 1. Dataset components and the size.

Component	Size
Document Descriptions	10,000
Structured Tables	10,000
Visualization Code Instances	10,000
Chart images	10,000
QA-Reasoning Pairs	30,000

Table 2. Data quality evaluation results.

Metric for Data Quality	Result (%)
Description–chart consistency	97.6
Reasoning appropriateness	99.5
Reasoning consistency	99.5

Table 3. Results of before and after fine-tuning with the proposed dataset.

Chart Q&A Reasoning Model	Accuracy (%)
Qwen2.5-VL-7B [6]	79.7%
+ Training with proposed dataset	**88.1%**
Chart Generation Model	**CodeBLEU**
Llama3-8B-Instruct [7]	0.08
+ Training with proposed dataset	**0.94**

4. Conclusion

In this paper, we presented a Korean dataset for two related tasks: document-to-chart generation and chart reasoning question answering. The dataset is designed as a unified pipeline linking document descriptions, structured tables, visualization code, chart images, and three-step reasoning QA. With 10,000 chart construction instances and 30,000 reasoning QA pairs, the dataset provides a practical benchmark for Korean document understanding, chart generation, and reasoning research. In future work, we plan to add baseline model comparisons and expand the dataset to broader document domains, chart types, and reasoning patterns.

References

[1] K. Kafle, B. Price, S. Cohen, and C. Kanan, “DVQA: Understanding Data Visualizations via Question Answering,” in CVPR, 2018.

[2] N. Methani, P. Ganguly, M. Khapra, and P. Kumar, “PlotQA: Reasoning over Scientific Plots,” in WACV, 2020.

[3] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, “ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning,” in Findings of ACL 2022.

[4] S. Hegde, P. Fazli, and H. Seifi, “ChartQA-X: Generating Explanations for Visual Chart Reasoning,” arXiv preprint arXiv:2504.13275, 2025.

[5] C. Jun, J. Choi, M. Sim, H. Kim, H. Jang, and K. Min, “Korean-Specific Dataset for Table Question Answering,” in LREC 2022.

[6] Qwen, “Qwen2.5-VL Technical Report,” arXiv preprint arXiv:2502.13923, 2025.

[7] Meta, “The Llama 3 Herd of Models,” arXiv preprint arXiv:2407.21783, 2024.

[8] Zadeh, Fatemeh Pesaran, Juyeon Kim, Jin-Hwa Kim, and Gunhee Kim. "Text2Chart31: Instruction tuning for chart generation with automatic feedback." In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 11459-11480. 2024.