Demonstration Selection Strategies for Numerical Time Series Data-to-Text

Masayuki Kawarada, Tatsuya Ishigaki, Goran Topić and **Hiroya Takamura**
Artificial Intelligence Research Center, AIST
{kawarada.masayuki, ishigaki.tatsuya
goran.topic, takamura.hiroya}@aist.go.jp

Abstract

Demonstration selection, the process of selecting examples used in prompts, plays a critical role in in-context learning. This paper explores demonstration selection methods for data-to-text tasks that involve numerical time series data as inputs. Previously developed demonstration selection methods primarily focus on textual inputs, often relying on embedding similarities of textual tokens to select similar instances from an example bank. However, this approach may not be suitable for numerical time series data. To address this issue, we propose two novel selection methods: (1) sequence similarity-based selection using various similarity measures, and (2) task-specific knowledge-based selection. From our experiments on two benchmark datasets, we found that our proposed models significantly outperform baseline selections and often surpass fine-tuned models. We also found that scale-invariant similarity measures such as Pearson's correlation work better than scale-variant measures such as Euclidean distance. Manual evaluation by human judges also confirms that our proposed methods outperform conventional methods.

1 Introduction

This paper explores demonstration selection approaches in data-to-text tasks. In particular, we focus on tasks involving numerical time series input. Two examples are depicted in Figure 1. The first example is drawn from *market comment generation* (Murakami et al., 2017), while the second example is drawn from *line graph-to-text generation* in the chart-to-text dataset (Kantharaj et al., 2022). For both tasks, graphs are represented as time series, i.e., as sequences of numbers, not images. In market comment generation, we generate concise textual explanations from fixed-length numerical time series, which represent price movements in the Japanese stock market Nikkei. In line graph-to-text generation, we generate explanations for varying-length numerical time series in various domains.

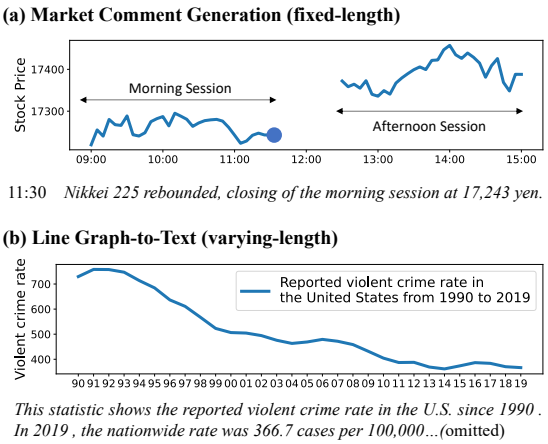

Figure 1: Examples of our target tasks: (a) market comment generation (fixed-length numerical time series), (b) line graph-to-text (varying-length numerical time series).

For both tasks, such time series are often hard to understand for non-experts, so language generation techniques are used to aid readers' comprehension.

In-context learning, where a prompt is fed into large language models, has recently gained prominence across various language processing tasks. A pivotal step therein that significantly influences task performance is demonstration selection, the process of selecting instances to include in the prompt (Liu et al., 2022). Given its significance, our focus is specifically on tasks with numerical time-series data as input—tasks that have not been extensively explored in existing studies. Existing methods of demonstration selections for various language processing tasks are based on either supervised scoring of examples with a classifier/regressor or an unsupervised similarity measure. Specifically, we focus on the latter, i.e., methods based on unsupervised similarity measures, because we aim to devise demonstration selection methods that are applicable to various domains without the additional cost of annotating training data. A prevalent approach for unsupervised selec-

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7378–7392

tion methods involves extracting the top-k examples that are most similar to the target test instance. Such methods often use similarity metrics based on embedded tokens or surface similarity (Agrawal et al., 2023; Liu et al., 2022). However, numerical data-to-text proves challenging due to the absence of token embeddings[1].

To overcome the aforementioned challenge, we propose (1) using the similarity of numerical time series and (2) leveraging auxiliary data aligned with the input, such as timestamps or graph legends (e.g., "Reported violent crime rate in the United States from 1990 to 2019," as shown in Figure 1(b)). First, we treat the input series as a vector or sequence and explore various vector similarity measures and different correlation functions. Second, we suggest incorporating auxiliary information with the series, such as timestamps or graph legends. For example, a comment with a timestamp of 9:00AM would often mention the opening of the market as in "*Nikkei opens at the price of 17,000 yen*". Thus, such task-specific knowledge—i.e., the fact that timestamps can influence market comment generation—is useful for extracting similar examples.

Our experiments use two datasets: market comment generation[2] and single-line graphs in chart-to-text dataset (Murakami et al., 2017; Kantharaj et al., 2022). From the results, we found that (1) models using our proposed methods outperform the baselines, (2) scale-invariant similarity measures, such as Pearson's correlation, perform better, and (3) combining task-specific knowledge-aware methods with sequence similarity-based methods further enhances performance, often rivaling or surpassing strong fine-tuning-based models. Manual evaluation by human judges corroborates these results in terms of correctness, i.e., whether generated comments mention correct price movements or numbers in graphs, against the input series.

In summary, our contributions are threefold: (1) we present the first exploration of demonstration selection for data-to-text tasks involving numerical time series, (2) we introduce two distinct selection methods, and (3) we verify their effectiveness on two datasets through both automatic evaluation using various common metrics and manual evaluation by human judges.

[1]The primary input is time series, but graph legends can be used in the line graph-to-text task, which are represented as tokens as an auxiliary.

[2]The data can be accessed through a contractual agreement, and the preprocessing code will be made publicly available for reproducibility.

2 Related Work

This paper connects two research fields: data-to-text and demonstration selection in in-context learning.

Numerous studies have been conducted on data-to-text. Traditionally, studies have focused on various input types, including tables (Puduppully et al., 2019; Lebret et al., 2016), graphs (Bai et al., 2022; Konstas et al., 2017), sets of tuples (Gardent et al., 2017) and numerical time series data (Gardent et al., 2017). While initial studies utilized rule-based approaches (Goldberg et al., 1994; Reiter et al., 2005), but more recently, there has been a shift towards neural network-based models (Sutskever et al., 2014; Murakami et al., 2017). Among neural network-based approaches, in-context learning using large pretrained language models has been recognized as a promising direction (Liu et al., 2022). Regarding the combination of data-to-text and demonstration selection, Liu et al. (2022) explored demonstration selection for tabular data, where the input is represented as a sequence of tokens. This paper presents a novel attempt to address demonstration selection in the context of numerical time series data.

The existing studies on demonstration selection were conducted on tasks other than data-to-text (Zhang et al., 2022; Agrawal et al., 2023; Chang and Jia, 2023; Nguyen and Wong, 2023; Yang et al., 2023; Peng et al., 2024). As an exception, Liu et al. (2022) explored k-nearest neighbour-based approach for a data-to-text setting, i.e., the Wikipedia table-to-text on the ToTTo dataset (Parikh et al., 2020). In contrast to this work, we focus on numerical data-to-text. Existing studies take one of three approaches: token similarity-based (Liu et al., 2022), surface similarity-based (Agrawal et al., 2023), or learning-based approachs (Chang and Jia, 2023; Nguyen and Wong, 2023; Zhang et al., 2022; Yang et al., 2023). These studies all use texts, while our focus is on numerical time series input.

3 Tasks

We describe two different numerical data-to-text tasks: market comment generation and line graph-to-text generation.

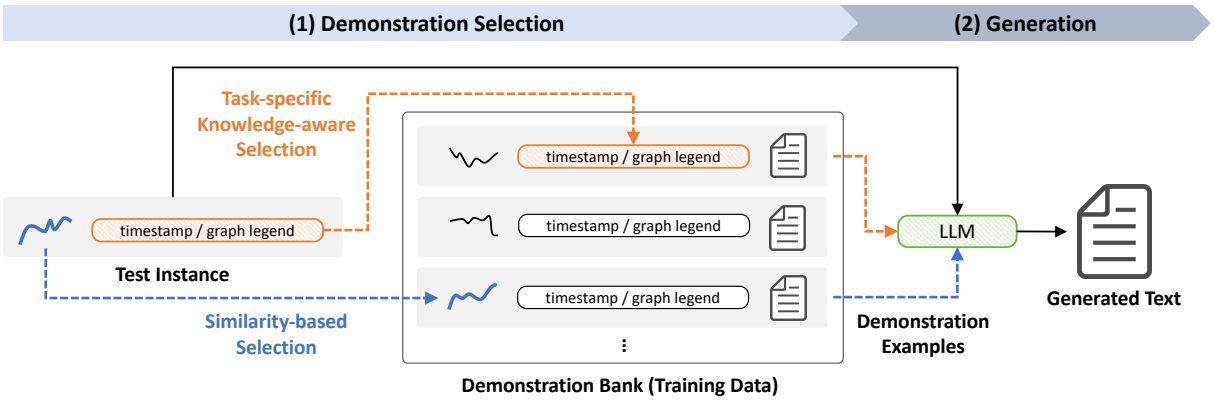

Figure 2: Overview of our proposed methods: (1) searching for similar cases from the training data based on a time series of prices or values in a graph, possibly with additional auxiliary data (e.g., timestamps or graph legends), and then (2) using the extracted examples for text generation.

3.1 Task Definitions

These two tasks involve taking a set of numerical sequences as input, with the model generating an explanation for the input series.

Market Comment Generation: The goal of this task is to generate concise market comments for a target timestamp based on two fixed length time series i.e., short-term series S_N and long-term series L_M. S_N includes prices recorded every five minutes, while L_M consists of daily closing prices, i.e., the last prices at every trading day, for the last M days. Each value in S_N corresponds to a specific timestamp, and similarly, closing prices in L_M align with specific dates. Notably, instances originate exclusively from the Nikkei market domain, enabling the use of specific market knowledge, such as the tendency to mention the market's opening at 9:00. Note that comments in this task are in Japanese.

Line Graph-to-Text: This task aims to generate concise paragraphs explaining input time series C_l across various domains, such as fluctuations in world population or variations in crime rates in the United States. Unlike the market comment generation, this task employs varying input lengths (l) for each C_l, allowing for flexibility in the input time series length. In addition, each line graph represented as time series C_l comes with an auxiliary graph legend represented as tokens indicating its domain, such as "*alterations in world population*" or "*shifts in criminal numbers in the U.S.*". Note that 1) market comment generation employs fixed values for N and M across all instances, while line graph-to-text uses varying input lengths l for each C_l, and 2) each time series in line graph-to-text comes with an auxiliary graph legend indicating its domain, while such textual data is not available in market comment generation.

3.2 Prompts

Our prompt used for experiments is divided into three segments, i.e., task description, demonstration examples, and target input, separated by "###":

In market comment generation, we write "*Generate a market comment for the current time...* (omitted)." as the task description. The second part presents selected demonstration examples (shots) in a table format, representing short-term series S_N (*Time* and *Nikkei Average Price*) and long-term series L_M (*Date* and *Nikkei Closing Price*). For example, the left column of S_N displays timestamps such as "*10:10, 10:05, 10:00,*" while the right column shows corresponding stock prices such as "*37,540, 37,569, and 37,552.*". We adopted this table-like format by following Kawarada et al. (2024). Finally, in the third part, we describe the target input in a table format similar to the shot inputs. The LLMs generate the continuation following the expression "Output: ".

For the line graph-to-text task, a prompt similar to the one for market comment generation is employed. We provide specific examples of the prompts used for the market comment generation task and the line graph-to-text task in Appendix A.

4 Demonstration Selection Approaches

Demonstration selection involves selecting shots from an example bank based on specific criteria. We present two distinct approaches for demonstration selection: 1) time-series similarity-based and

2) task-specific knowledge-based. Additionally, we explore the combination of similarity-based methods with task-specific knowledge-aware approach.

4.1 Similarity-based Selection

In contrast to existing demonstration selection methods employing the similarity between token embeddings (Liu et al., 2022; Agrawal et al., 2023), we use the similarity between time-series numerical sequences. In particular, we comprehensively compare six similarity measures: Manhattan distance, Euclidean distance, dynamic time warping (DTW) (Mueller, 2007), cosine similarity, and Spearman's and Pearson's correlation coefficients. Note that, for the line graph-to-text task, the varying lengths of input time series in each test instance restrict us from using the above measures except for DTW.

We posit that gold texts describing analogous numerical time series data are likely to resemble each other, e.g., these comments mention similar price movements. Thus, it may be possible to improve correctness by including such examples in prompts.

The first three measures are scale-variant, meaning they are affected by the exact size of the values. In contrast, the latter three measures are scale-invariant, meaning they are less influenced by the absolute magnitude of the values. Market commentary often mentions price movements such as increases and decreases. Thus, we believe that the use of scale-invariant measures may facilitate the retrieval of similar price movements. Consequently, we expect a decrease in critical errors regarding price fluctuations, leading to an improvement in the correctness of output comments.

4.2 Task-specific Knowledge-aware Selection

The Use of Timestamps: To enhance our approach, we incorporate task-specific knowledge, specifically timestamps, in the market comment generation task. For example, a comment with a timestamp of 9:00AM would often mention the opening of the market as in "*Nikkei opens at the price of 17,000 yen*". When dealing with a target instance, it is more likely to cover a similar topic as other comments issued at the same timestamp. To integrate this task-specific knowledge, our approach involves randomly selecting examples with comments issued at the same time as the target comment. It is crucial to note that this task-specific knowledge integration technique is exclusive to the market comment generation task. The variability in line graphs over the same period does not guarantee similarity, making this approach not feasible for the line graph-to-text task.

The Use of Graph Legends: In the task of line graph-to-text task, where time series are associated with a graph legend (e.g., tokens of "*the population of the U.S.*"), we propose using graph legends to select similar examples. To implement this approach, we use BERTScore to measure the embedding similarity between graph legends and then select the top-k examples with the highest similarity from our example bank. While the effectiveness of this strategy, involving the use of tokens, is well-established in existing studies (Liu et al., 2022), it is specifically designed for tasks involving graph legends and cannot be applied to domain-dependent scenarios that lack graph legends.

4.3 Combination of Two Methods

task-specific knowledge-aware selections can be combined with similarity–based methods. In market comment generation, we first sample examples with comments issued at the target timestamp and then rank them based on the similarity measures explained in Section 4.1. Similarly, for line graph-to-text generation, we extract the top-50 examples using BERTScore and include the top-k using DTW as similarity metrics.

5 Experiments

5.1 Datasets

We use two datasets: the market comment generation dataset (Murakami et al., 2017; Aoki et al., 2018; Hamazono et al., 2021) and a subset of the chart-to-text dataset (Kantharaj et al., 2022).

The market comment generation dataset (Hamazono et al., 2021) comprises 18,489 single-sentence market comments, each paired with a numerical time-series. This dataset contains 15,035 examples for training, 1,759 examples for validation, and 1,695 examples for testing. We sampled 500 instances from the dataset for testing, taking into account the financial constraints associated with using the OpenAI API. We used the training instances as the example bank for demonstration selection. The time-series dataset, spanning from December 2010 to September 2016, is sourced from IBI-Square[3], while the comments are gathered from Nikkei Quick News. Aligning with prior

[3] http://www.ibi-square.jp/index.html

studies, the short-term series captures Nikkei stock prices recorded every 5 minutes during a trading day, with the series length N set at 62. For the long-term series, representing closing prices, we define the series length M as 7.

The line graph-to-text dataset is derived by extracting 2,360 single-line graphs from the chart-to-text dataset (Kantharaj et al., 2022). These graphs were originally collected from Statista[4]. Further refinement is done by extracting 1,912 graphs with the x-axis labeled as "year.". We adhere to the official split of the original dataset to create training, validation, and test subsets, and the split details are publicly available for reproducibility[5]. The resulting subsets consist of 1,188, 334, and 390 instances for training, validation, and testing, respectively. In both datasets, the training data segments are utilized as example banks.

5.2 Compared Methods

To evaluate our proposed selections, we compare them with baseline methods in two categories: 1) prompting methods and 2) fine-tuned encoder-decoder.

5.2.1 Prompting Methods

In addition to the proposed methods introduced in Section 4, we compare two baseline selection methods. **Random Selection** involves randomly selecting examples from the example bank. it represents the most prevalent approach in numerous studies. For this baseline, preliminary experiments revealed significant variations in model performance based on the selected examples. To account for this, we perform five random selections using different seeds and report the average scores. **Embedding-based Selection** involves selecting the top-k similar examples based on the embeddings calculated by SentenceBERT (Reimers and Gurevych, 2019). To obtain embeddings, we feed the linearized tokens of this tabular representation into SentenceBERT.

We employed OpenAI's GPT-3.5-turbo[6] and open-sourced LLM for the prompting methods. For reproducibility, we set the temperature to 0 during inference. Since the chart-to-text dataset is available on GitHub, there is a possibility that it was included in OpenAI's training data. To address this, we also conducted experiments using GPT-NeoX-20B (Black et al., 2022) for open-sourced LLMs, which is trained with the publicly available Pile dataset (Gao et al., 2020)[7]. We manually verified that our evaluation datasets are not contaminated by the training data of this LLM. In experiments with the open-source LLM, we only conducted 3-shot and 5-shot experiments because the context length becomes too long with 10-shot settings. Note that we used the same prompts for both models. Implementation details are provided in the Appendix B.

5.2.2 Fine-tuned Encoder-Decoder

EncDec-MLP is an existing encoder-decoder-based approach. This architecture has become standard, as demonstrated by previous work Murakami et al. (2017); Aoki et al. (2018); Hamazono et al. (2021), which employed LSTM-based encoder-decoder models. Our implementation, however, is based on BART (Lewis et al., 2020), a more commonly used model in recent data-to-text studies (Tang et al., 2022; Ishigaki et al., 2023). In our configuration, Multi-Layer Perceptrons (MLPs) convert long- and short-term vectors (S_N and L_M) into embeddings of size 768, aligning with the embedding layer size in the pretrained BART. During fine-tuning, we initialize MLP parameters randomly, while BART parameters inherit pretrained weights. The parameters are described in detail in Appendix B. **EncDec-token** is a common approach for data-to-text studies where the table—or, in this context, the time series—is linearized into a sequence of tokens, which are then fed into an encoder-decoder model. We use BART as the encoder-decoder.

5.3 Evaluation Metrics

We conducted evaluations using common automatic metrics and human judges.

5.3.1 Automatic Metrics

We employ BLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020), following existing studies (Murakami et al., 2017; Kantharaj et al., 2022). The F1 score is utilized to compute the BERTScore. All metric calculations used the HuggingFace library[8]. These

[4] https://www.statista.com/

[5] https://github.com/our_repo/

[6] We used `gpt-3.5-turbo-0125` for our experiments.

[7] EleutherAI released the data included in the Pile dataset, which can be accessed at https://github.com/EleutherAI/the-pile. We confirmed that the Chart-to-text GitHub page is not included in this dataset.

[8] The BERTScore uses https://huggingface.co/tohoku-nlp/bert-base-japanese.

Method	3-shot BLEU	3-shot MET.	3-shot BScore	5-shot BLEU	5-shot MET.	5-shot BScore	10-shot BLEU	10-shot MET.	10-shot BScore
Baseline Selections									
Random	8.21	22.66	72.47	8.76	23.90	72.91	9.65	25.68	73.59
Embedding-based similarity	7.98	22.95	72.58	8.49	23.42	72.61	9.15	24.70	73.28
Proposed: Selection using Scale-variant Measures									
Manhattan distance	8.72	25.45	72.95	9.91	25.62	73.81	10.75	27.77	74.42
Euclidean distance	9.19	25.97	73.43	10.07	26.43	74.13	10.85	28.29	74.61
DTW	8.91	25.84	73.01	9.46	25.62	73.58	10.55	27.87	74.33
Proposed: Selection using Scale-invariant Measures									
Cosine similarity	11.11	28.10	74.42	11.93	29.69	74.95	12.43	31.12	75.83
Spearman's correlation	10.65	27.80	74.47	11.22	28.74	74.76	13.11	31.24	75.69
Pearson's correlation	**12.41**	**31.74**	**75.63**	**12.62**	**31.72**	**75.59**	**14.53**	**35.21**	**76.75**
Proposed: Task-knowledge									
Same timestamp	11.62	29.56	74.69	12.67	31.00	75.18	14.07	33.24	75.75
Proposed: Combination using Scale-variant Measures									
Manhattan distance	14.03	34.65	76.67	14.07	33.85	76.22	16.99	37.69	77.34
Euclidean distance	**14.74**	35.73	**76.98**	14.68	34.44	76.13	16.50	37.54	77.25
DTW	14.12	34.93	76.66	**15.44**	**34.92**	76.64	**17.07**	38.17	77.64
Proposed: Combination using Scale-invariant Measures									
Cosine similarity	14.01	34.55	76.69	13.92	34.45	76.60	14.40	35.15	77.08
Spearman's correlation	12.85	33.28	76.00	14.08	34.43	76.22	16.55	37.96	77.55
Pearson's correlation	14.21	**36.01**	76.80	14.71	36.57	**77.13**	16.66	**38.97**	**78.20**

Table 1: Results for the market comment generation task using GPT-3.5-turbo as LLM, evaluated with BLEU, METEOR (MET.), and BERTScore (BScore).

Selection Method	3-shot BLEU	3-shot MET.	3-shot BScore	5-shot BLEU	5-shot MET.	5-shot BScore	10-shot BLEU	10-shot MET.	10-shot BScore
Baseline Selections									
Random	22.90	45.33	89.35	25.22	47.71	89.84	29.00	51.00	90.48
Embedding-based similarity	24.29	45.63	89.52	27.63	49.10	90.16	32.01	53.25	90.86
Proposed									
Time-series similarity.	33.57	55.14	91.00	36.89	58.68	91.60	41.82	62.35	92.16
Task-knowledge	**47.01**	**66.17**	**92.71**	**49.85**	**68.47**	**93.09**	**52.10**	**70.44**	**93.47**
Combination	44.93	64.28	92.51	48.67	67.40	92.92	50.85	69.98	93.34

Table 2: Evaluation results for line graph-to-text task using GPT-3.5-turbo as LLM. MET. and BScore stand for METEOR and BERTScore, respectively. Note that similarity-based approaches are not applicable to this task due to the varying lengths of the inputs.

metrics cannot directly capture the correctness of generated comments.

5.3.2 Human Evaluation

Automatic metrics cannot evaluate factual correctness, such as identifying hallucinations. To further assess the quality of generated texts, we conducted a human evaluation on 30 randomly sampled instances from both datasets. For market comment generation, two native Japanese speakers evaluated the market comment generation in terms of two criteria: **correctness** (which comment is more factually correct, i.e., without contradictions against the reference facts) and **fluency** (which comment is more fluent, i.e., free of grammatical errors). We compare the outputs of the combination model, EncDec-MLP, and the random baseline. For chart-to-text, three fluent English speakers assessed the line graph-to-text instances. We use three criteria: **correctness**, **fluency**, and **coherence** (which comment is more coherent, i.e., sentences are semantically well connected). We compared the outputs of the knowledge-based selection model, EncDec-token, and the random baseline. We did not evaluate market comments for coherence, as each comment consists of a single sentence. We report the percentage of comments that were judged to be better than those generated by another method.

6 Results and Discussion

6.1 Main Results

Table 1 presents the results of market comment generation using GPT-3.5-turbo as the LLM. The

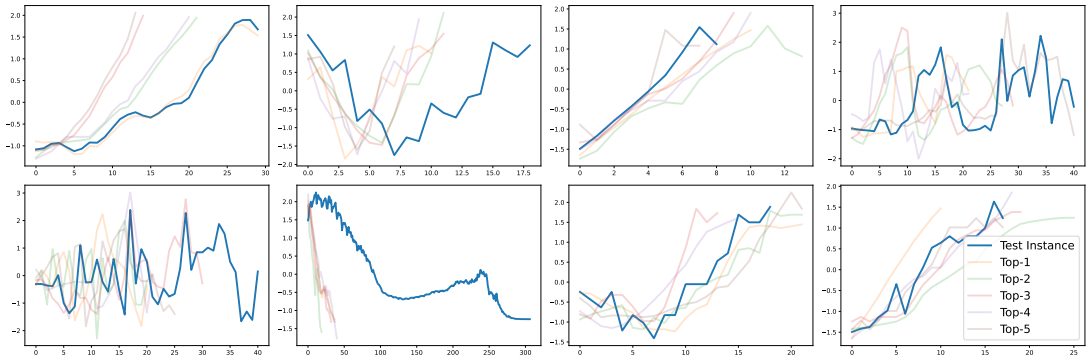

Figure 3: The line graphs of selected examples in the chart-to-text dataset. The blue line shows the test instance and lines colored in other than blue are the lines of top-5 selected examples.

	Method		BLEU	MET.	BScore
Market Comment	3-shot	Rand.	5.91	17.79	63.34
		Embed.	6.06	17.80	61.79
		Knowl.	7.93	21.73	61.57
		Sim.	8.13	22.16	**63.54**
		Comb.	**9.54**	**23.91**	62.30
	5-shot	Rand.	7.68	22.48	71.68
		Embed.	7.24	22.82	71.88
		Knowl.	9.68	26.98	70.60
		Sim.	10.43	28.56	**72.73**
		Comb.	**11.47**	**31.21**	72.53
Line Graph	3-shot	Rand.	24.25	42.93	89.28
		Embed.	31.96	50.21	89.58
		Knowl.	**50.99**	**65.52**	92.93
		Sim.	38.83	56.55	92.01
		Comb.	48.57	63.99	**93.01**
	5-shot	Rand.	25.96	43.66	88.75
		Embed.	33.23	51.89	89.66
		Knowl.	**51.21**	**65.51**	92.15
		Sim.	41.37	58.56	92.39
		Comb.	49.97	64.70	**92.50**

Table 3: Comparison of BLEU, METEOR (MET.), and BERTScore (BScore) using GPT-NeoX-20B as the LLM. "Rand." refers to random selections, "Embed." refers to embedding-based selections,"Knowl." to knowledge-based selections, "Sim." to similarity-based selections, and "Comb." to combined selections. For the market comment generation task, we use Pearson's correlation as the similarity metric.

	Method		BLEU	MET.	BScore
Market Comment	10-shot	Rand.	9.65	25.68	73.59
		Embed.	9.15	24.70	73.28
		Knowl.	14.07	33.24	75.75
		Sim.	14.53	35.21	76.75
		Comb.	**17.07**	**38.17**	**77.64**
	EncDec-MLP		12.74	35.43	76.35
	EncDec-token		12.08	33.40	75.58
Line Graph	10-shot	Rand.	29.00	51.00	90.48
		Embed.	32.01	53.25	90.86
		Knowl.	**52.10**	**70.44**	**93.47**
		Sim.	41.82	62.35	92.16
		Comb.	50.85	69.98	93.34
	EncDec-MLP		-	-	-
	EncDec-token		51.03	66.79	93.47

Table 4: Comparison of few-shot learning using GPT-3.5-turbo and fine-tuned encoder-decoder based on BLEU, METEOR (MET.), and BERTScore (BScore). EncDec-MLP cannot be used for line graph-to-text conversion because the input lengths vary for each test instance.

scores are reported for different numbers of demonstration examples selected for the prompt. In the proposed time-series similarity-based selections, performance improvements are observed across all settings compared to the baseline selections. Notably, the BLEU score exhibits enhancements of 4.2 (8.21 up to 12.41), 3.86 (8.76 up to 12.62), and 4.88 (9.65 up to 14.53) for 3-, 5-, and 10-shot scenarios of Pearson's correlation-based method, respectively. Other metrics also reflect improvements, with the Pearson's correlation outperforming other proposed similarity measures. Combined methods exhibit further performance improvements. For instance, the combination of task knowledge and Euclidean distance achieves a BLEU score of 14.74, surpassing individual models (9.19 for Euclidean distance alone and 11.62 for task knowledge alone).

Table 2 shows the results on line graph-to-text using GPT-3.5-turbo as the LLM. Both similarity- and knowledge-based methods outperform the random selection-based and embedding-based baselines. The knowledge-based approach significantly enhances the BLEU score from 22.90 to 47.01, 49.85, 52.10 in 3-, 5-, and 10-shot settings, respectively. This underscores the utility of token similarity in line graph-to-text, aligning with the findings from existing studies (Liu et al., 2022; Agrawal et al., 2023). For the combined model,

	Comb. (1) vs. EncDec-MLP (2)		Comb. (1) vs. Rand. (2)		EncDec-MLP (1) vs. Rand. (2)	
Summary	**Correctness**	**Fluency**	**Correctness**	**Fluency**	**Correctness**	**Fluency**
Summary 1 Win	36.66%	0%	43.30%	0%	36.66%	0
Summary 2 Win	10.00%	0%	10.00 %	0%	13.33%	0
Tie	53.33%	100%	46.70%	100%	50.00%	100%
p-value (sign test)	0.0286†	-	0.0106†	-	0.0592	-

Table 5: Human evaluation results for market comment generation. † indicates that the difference between two methods is significant ($p < 0.05$).

	Knowl. (1) vs. EncDec-token (2)			Knowl. (1) vs. Rand. (2)			EncDec-token (1) vs. Rand. (2)		
Summary	**Correctness**	**Coherence**	**Fluency**	**Correctness**	**Coherence**	**Fluency**	**Correctness**	**Coherence**	**Fluency**
Summary 1 Win	40.00%	3.33%	10.00%	43.30%	10.00%	13.30%	36.70%	10.00%	0%
Summary 2 Win	13.33%	0%	0%	16.70 %	6.80%	6.80%	20.00%	6.70%	6.70 %
Tie	46.66%	96.66%	90.67%	40.00%	83.33%	80.00%	43.33%	83.33%	93.33%
p-value (sign test)	0.0380†	0.500	0.125	0.0481†	0.5000	0.3430	0.1666	0.5000	0.2500

Table 6: Human evaluation results for line graph-to-text generation. † indicates that the difference between two methods is significant ($p < 0.05$).

unlike in market comment generation, we did not observe performance gains from the task knowledge method alone. The possible reason for this is that graphs in the same domain can already be extracted by using the token sequence of graph legends, so including graphs in different domains by using the method for extracting similar series may result in noise.

We also present the experimental results of using GPT-NeoX-20B as the LLM in Table 3. These results show that our proposed method outperforms the baseline random selection in both tasks. This result is consistent with the results obtained using GPT-3.5-turbo.

6.2 Scale-variant vs. Scale-invariant measures

Scale-invariant measures such as Pearson's correlation coefficient are more effective than scale-variant measures such as Euclidean distance. Methods using scale-invariant measures can find instances where the absolute values of two compared numerical time series are different, but the changes in the values are similar. Since comments often mention number movements, e.g., *rises* and *falls*, this may reduce errors in these expressions.

6.3 The Comparison with Fine-tuned Encoder-Decoder

In market comment generation, our proposed models outperform existing fine-tuned models (EncDec-MLP), as shown in Table 4. For instance, the combined model with 10-shot examples achieves scores of 17.07 in BLEU, 38.17 in METEOR, and 77.64 in BERTScore, outperforming the optimal existing method (EncDec-MLP). In the line graph-to-text task, our proposed knowledge-based approach with 10-shot examples achieves scores of 52.10 in BLEU, 70.44 in METEOR, and 93.77 in BERTScore, surpassing the existing fine-tuned model (EncDec-token). A comparison of the output texts generated by our proposed method, random selection, and the fine-tuned encoder-decoder model is presented in Appendix D.

6.4 Examples of Selected Time Series

Figure 3 shows the line graphs that were actually selected in the line graph-to-text task. The blue line in each graph shows the test instance and we select the top-5 similar examples by using our similarity-based approach with DTW. The lines colored in other than blue represent the lines of the top-5 selected examples. These graphs suggest that DTW can capture similar movements of line graphs even when the series lengths are different. For example, among the lines in the left-top graph, the most similar selected line is colored in orange, which is almost identical to the blue line of the test instance. The other four lines in this graph represent shorter time series, but the movements are similar to that of the test instance. A similar tendency pattern is observed in the other graphs.

6.5 Evaluations by Human Judges

The results of the human judges' evaluations, as shown in Table 5 and 6, indicate a higher rate of factual correctness for both tasks. The Cohen's kappa (Cohen, 1960) for the correctness of market comment generation is 0.472. Meanwhile, Fleiss's

Reference		*The Nikkei 225* ***continues to rise****, having recovered to the 16,000 yen mark.*
Random	Examples	*The Tokyo Stock Exchange dips, but stabilizes at 2:00 p.m., supported by rising Chinese stocks.* *The Tokyo Stock Exchange closed,* ***rebounding to above 9,600 yen****, supported by a rise in the U.S. Dow.* *The Nikkei 225 opens with a decline, falling to 14,100 yen.*
	Output	*The Tokyo Stock Exchange closed* ***down for the third consecutive day****, falling below 16,000 yen.*
Proposed	Shots	*The Nikkei* ***continues to rise substantially****, up 276 yen to close at 11,662 yen.* *The Nikkei 225 closed* ***sharply higher*** *due to reports on Bank of Japan personnel and expectations of participation in TPP negotiations.* *The Tokyo Stock Exchange closed at its* ***highest level*** *in 4 years and 5 months, due to news about the Bank of Japan's personnel changes.*
	Output	*The Nikkei 225* ***continues to rise****, up 184 yen to close at 16,015 yen.*

Table 7: Comparison of the demonstration examples (shots) and output for the market comment generation task. Words describing movements for stock prices in both the reference and output texts are highlighted in **bold**, and terms similar to the reference in the shots are marked in **red**. Note that the actual examples are in Japanese, and these are translations into English.

kappa (Fleiss et al., 1971) for the correctness, fluency, and coherence of line graph-to-text are 0.558, 0.425, and 0.369, respectively. In terms of fluency on market comment generation, all comments were judged fluent, thus, the value of tie is 100%. For both tasks, our proposed method demonstrates a significant difference in correctness compared to other methods ($p < 0.05$). These results reveal that our proposed method is capable of improving correctness while maintaining fluency and coherence.

6.6 The Impact of Selected Demonstration on Generated Texts

Table 7 presents a comparison of the generated text and selected demonstrations for both the proposed method and the baseline in the market comment generation task. "Random" refers to the random selection baseline, while "Proposed" denotes our method combining similarity-based and knowledge-aware selection, using Pearson's correlation as the similarity metric. The reference text contains the phrase "*continues to rise*" indicating an increase in stock price. However, the text generated by the random selection method includes "*down for the third consecutive day*", implying a decrease in stock price. Conversely, our proposed method correctly generates the phrase "*continues to rise*", which describes increasing stock prices. This is likely because our demonstration selection method effectively retrieves instances similar to the reference. As highlighted in red in the table, the examples selected by our method frequently include expressions such as "*continues to rise substantially*", "*sharply higher*" and "*highest level*":all of which suggest rising stock prices. This enables the LLMs to better understand the relationship between the time series of the given test instance and the desired text.

7 Conclusion

This paper explores demonstration selection in numerical time series data-to-text tasks. We explore various similarity measures for demonstration selection and also propose the task knowledge-aware approaches. We verified that our approaches outperform the random and embedding similarity-based baselines, which have been used in many existing studies. Among the similarity measures, our experiments suggest that scale-invariant measures work better. We also found that the proposed methods with demonstration selection perform better than the method with fine-tuning, which had previously shown good performance as the de facto standard. As future work, we plan to enhance our demonstration selections for multiple time series and multi-modal settings.

Acknowledgements

This paper is based on results obtained from project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), and a project Programs for Bridging the gap between R&D and the IDeal society (society 5.0) and Generating Economic and social value

(BRIDGE)/Practical Global Research in the AI × Robotics Services, implemented by the Cabinet Office, Government of Japan.

Limitations

Input Type: In this paper, we assume that the input is a single type of time series data, i.e., price movements in the Japanese stock market or a line graph in chart-to-text dataset. However, various types of data exist in real-world such as pie and bar charts (Kantharaj et al., 2022). Among studies that treat time-series as input, there is multiple time-series, for example, time series used in Chang et al. (2022); Ishigaki et al. (2021) contain boolean values in addition to numerical numbers. Our current work is limited to use numerical values. Also, our current proposed methods are limited to be applied to a single sequence for demonstration selection.

The Use of OpenAI API: The use of OpenAI API in our experiments might raise a reproducibility issue when the API is discontinued. Also, because the Chart-to-Text dataset is publicly available on GitHub, there is no guarantee that it has not been used for training GPT-3.5-turbo[9]. Therefore, we are also conducting experiments using GPT-NeoX-20B, which is an open-source LLM.

Constraints on Cost and Computation Resources: We use a single A100 GPU (80GB) for experiments that use EncDec-based methods. We need to spend several hours to conduct experiments in this paper. There are also costs associated with OpenAI: it takes over USD1,000 to reproduce our results.

Ethical Considerations

It is difficult to control the tokens output when using large language models. Therefore, there is a risk of offensive words being output.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Tatsuya Aoki, Akira Miyazawa, Tatsuya Ishigaki, Keiichi Goshima, Kasumi Aoki, Ichiro Kobayashi, Hiroya Takamura, and Yusuke Miyao. 2018. Generating market comments referring to external resources. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 135–139, Tilburg University, The Netherlands. Association for Computational Linguistics.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

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.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Ernie Chang, Alisa Kovtunova, Stefan Borgwardt, Vera Demberg, Kathryn Chapman, and Hui-Syuan Yeh. 2022. Logic-guided message generation from raw real-time sensor data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6899–6908, Marseille, France. European Language Resources Association.

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144, Toronto, Canada. Association for Computational Linguistics.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on*

[9]Regarding market comment generation, the dataset is not publicly available and can only be used under a contract. Therefore, the likelihood of it being included in the training data for LLMs is extremely low.

Natural Language Generation, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Eli Goldberg, Norbert Driedger, and Richard I. Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert: Intelligent Systems and Their Applications*, 9(2):45–53.

Yumi Hamazono, Tatsuya Ishigaki, Yusuke Miyao, Hiroya Takamura, and Ichiro Kobayashi. 2021. Unpredictable attributes in market comment generation. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 217–226, Shanghai, China. Association for Computational Lingustics.

Tatsuya Ishigaki, Goran Topić, Yumi Hamazono, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2023. Audio commentary system for real-time racing game play. In *Proceedings of the 16th International Natural Language Generation Conference: System Demonstrations*, pages 9–10, Prague, Czechia. Association for Computational Linguistics.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Masayuki Kawarada, Tatsuya Ishigaki, and Hiroya Takamura. 2024. Prompting for numerical sequences: A case study on market comment generation. In *Proceedings of The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING2024)*, pages xxxx–xxxx, Turin, Italy.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Meinard Mueller. 2007. *Dynamic Time Warping*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg.

Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. 2017. Learning to generate market comments from stock prices. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1384, Vancouver, Canada. Association for Computational Linguistics.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences.

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

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Tianyi Tang, Junyi Li, Zhipeng Chen, Yiwen Hu, Zhuohao Yu, Wenxun Dai, Wayne Xin Zhao, Jian-yun Nie, and Ji-rong Wen. 2022. TextBox 2.0: A text generation library with pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 435–444, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. 2023. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5443–5456, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Generate a market comment for the current time and enclose the output comment by tags <comment>output comment</comment>.

###
Input:
Time Nikkei Average Price
10 : 10 37, 540
10 : 05 37, 569
⋮
Day Nikkei Closing Price
1_*DayAgo* 37, 320
2_*DayAgo* 37, 701
⋮
Output:
<comment>*Nikkei increased over 200 yen from yesterday... </comment>*

###
Input:
Time Nikkei Average Price
10 : 10 37, 240
10 : 05 37, 163
⋮
Day Nikkei Closing Price
1_*DayAgo* 37, 221
2_*DayAgo* 37, 701
⋮
Output:

Figure 4: The translated prompt from the original Japanese one. The prompt can be divided by the token "###" and it is composed of three parts: 1) task description, 2) demonstration examples and 3) the target input series.

A Full Prompts

We show examples of the prompts used for the market comment generation task in Figure 4 and for the line graph-to-text task in Figure 5.

B Implementation Details

Table 8 shows the hyperparameters used during inference with GPT-NeoX-20B. We performed inference on both market comment generation and line graph-to-text tasks using eight A100 (40GB) GPUs. After performing 4-bit quantization, we set the batch size to 1 and the new max tokens to 256. We also show the hyperparameters used for training using fine-tuned encoder-decoder in Table 9. We conducted experiments using BART[10] as

[10]We used https://huggingface.co/stockmark/bart-base-japanese-news for market comment generation task and employed https://huggingface.co/facebook/bart-base for line graph-to-text generation task.

Generate a caption that describes the table below, and format it by placing the caption between <caption>output</caption> tags.

###
Input:
Year Consumer Price Index
2019 103.83
2018 104.63
⋮

###
Input:
Year The population of U.S.
2006 299,753,098
2007 302,743,399
⋮
Annual population in U.S. from 2006 to 2010
Output:

Figure 5: The prompt used for the line graph-to-text generation. The prompt can be divided by the token "###" and it is composed of three parts: 1) task description, 2) demonstration examples and 3) the target input series.

the encoder-decoder model. The experiments were performed on a single A100 (80GB) GPU.

	Hyperparameter
batch_size	1
new_max_tokens	256
beam_size	1
load_in_4bit	True

Table 8: Hyperparameters for inference using GPT-neoX-20B.

C Quality of Selected Examples

We assume that the comments in the selected examples are likely to similar to the gold comments. Table 10 shows BLEU, METEOR, and BERTScore calculated for selected examples compared to the gold comments in the market comment generation task under the 10-shot setting. We verify that the achieved scores are higher, e.g., 13.02 in BLEU, than the scores obtained by the random baseline, e.g., 6.82 in BLEU. This suggests that our approach selects examples that more closely resemble the gold-standard comments.

D Examples of Generated Texts

Table 11 shows examples of generated texts and gold texts for market comment generation, while Table 12 presents examples of generated texts

	Hyperparameter
batch_size	8
max_source_length	512
max_target_length	256
epoch	200
dropout	0.1
adam_beta1	0.9
adam_beta2	0.998
bf16	True

Table 9: Hyperparameters for EncDec-token and EncDec-emb.

Method	BLEU	MET.	BScore
Baseline			
Random	6.82	19.93	71.67
Embedding-based similarity	6.23	18.99	71.52
Proposed: Selection using Scale-variant Measures			
Manhattan distance	8.09	22.80	72.67
Euclidean distance	8.46	23.43	72.83
DTW	7.91	22.67	72.52
Proposed: Selection using Scale-invariant Measures			
Cosine Similarity	7.96	23.67	72.99
Spearman's correlation	**10.65**	24.42	73.21
Pearson's correlation	10.63	**28.24**	**74.30**
Proposed: Combination using Scale-variant Measures			
Manhattan distance	12.36	31.17	75.29
Euclidean distance	12.65	31.53	75.35
DTW	12.40	31.25	75.25
Proposed: Combination using Scale-invariant Measures			
Cosine similarity	10.36	29.50	74.76
Spearman's correlation	11.83	30.87	74.97
Pearson's correlation	**13.02**	**33.17**	**75.77**

Table 10: The selected examples for the market comment generation task were evaluated against the gold comments using BLEU, METEOR, and BERTScores.

and gold texts for line graph-to-text. We observe hallucinations, shown in bold in the table, produced more by the EncDec-token and random models than the combination models. For example, EncDec-token wrongly generated "crude oil" where "trade goods" would be expected. We could find many other instances of similar errors, e.g., country names and product names. We find many instances where such critical errors are alleviated in the outputs by combination models. In the random baseline, a currency unit is mistakenly produced as "U.N. dollar" instead of the correct unit "U.S. dollar". By selecting examples, we found several cases where such errors are fixed.

Reference	Nikkei 225 down 624 yen as selling accelerates due to the Bank of Japan easing off.
EncDec-tok.	*The Nikkei 225* ***rises sharply*** *due to strong U.S. stocks.*
GPT-3.5-turbo (Random)	*The Nikkei 225* ***rebounded sharply****, up 330 yen to close at 17,500 yen.*
GPT-3.5-turbo (Proposed)	*The Nikkei 225 closes sharply lower, at a one-week low.*
GPT-NeoX-20B (Random)	*The Nikkei 225 showed minor movements,* ***fluctuating*** *in the 17,000 yen range.*
GPT-NeoX-20B (Proposed)	*The Nikkei 225 fell by more than 500 yen as major stocks decline.*
Reference	*At 10:00, the Tokyo Stock Exchange is up, supported by favorable yen depreciation and the BOJ's monetary policy.*
EncDec-tok.	*At 10:00, the Tokyo Stock Exchange* ***slight price fluctuations*** *around the low 16,500s.*
GPT-3.5-turbo (Random)	*The Nikkei 225 is* ***down*** *by more than 100 yen due to export-related sell-off amidst the yen's appreciation.*
GPT-3.5-turbo (Proposed)	*At 10:00 AM, the Tokyo Stock Exchange was up more than 200 yen at one point.*
GPT-NeoX-20B (Random)	*At 10:00 AM, the Tokyo Stock Exchange* ***fells*** *further, with the decline exceeding 200 yen at one point.*
GPT-NeoX-20B (Proposed)	*At 10:00 AM, the Tokyo Stock Exchange is up more than 100 yen.*

Table 11: Examples of outputs obtained from the market comment generation task. Hallucinations are shown in **bold**.

Reference	*This graph shows the growth in the U.S. import volume of trade goods from Saudi Arabia from 1985 to 2019 . In 2019 , U.S. imports from Saudi Arabia amounted to approximately 13.44 billion U.S. dollars .*
EncDec-tok.	*This graph shows the import volume of* **crude oil** *in the U.S. from 1985 to 2019. In 2019, the U.S. imported approximately 13.44 billion U.S. dollars worth of* **crude oil** *.*
GPT-3.5-turbo (Random)	*This graph shows the growth in the U.S' import volume of trade goods from Saudi Arabia from 1985 to 2019. In 2019, U.S. imports from Arabia amounted to approximately 13.44 billion* **U.N. dollars**.
GPT-3.5-turbo (Proposed)	*This graph shows the growth in the U.S. import volume of trade goods from Saudi Arabia from 1985 to 2019. In 2019, U.S. imports from Saudi Arabia amounted to approximately 13.44 billion U.S. dollars.*
GPT-NeoX-20B (Random)	*This graph shows the volume of U.S. imports of trade goods from Saudi Arabia from 1985 to 2019 . In 1985 , U.S. imported 13.44 billion U.S. dollars worth of goods .*
GPT-NeoX-20B (Proposed)	*This graph shows the volume of U.S. imports of trade goods from Saudi Arabia from 1985 to 2019 . In 2019 , U.S. imports from Saudi Arabia were about 13.44 billion U.S. dollars .*
Reference	*There were just over 616 thousand deaths in the United Kingdom in 2018 , compared with 607 thousand in 2017 . Between 2003 and 2011 , the annual number of deaths in the UK fell from 612 thousand to just over 552 thousand . Since 2011 however , the annual number of annual deaths in the United Kingdom has steadily grown , with the number recorded in 2018 a high for the provided time-period .*
EncDec-tok.	*Between 2000 and 2018, each year roughly between* ***610.58 and 602****. thousand people died in the UK. From 2014 onwards, the number of deaths increased annually, peaking in 2018 at* ***608.01*** *thousand.*
GPT-3.5-turbo (Random)	*This statistic shows the number of deaths in the United Kingdom from 2000 to 2018 . In* ***2011*** *, there were approximately 616,010 deaths recorded in the United Kingdom .*
GPT-3.5-turbo (Proposed)	*This statistic displays the number of deaths in the United Kingdom (UK) from 2000 to 2018. In 2018, there were approximately 616,010 deaths recorded. The number of deaths has generally increased over the years, with a slight decrease in some years.*
GPT-NeoX-20B (Random)	*This data indicate the overall number of deaths in* ***the United States*** *from 2000 to 2018. In 2018, approximately 616,010 deaths were documented in the* ***US****.*
GPT-NeoX-20B (Proposed)	*This data represents the total number of deaths in the United Kingdom between 2000 and 2018. In 2018, around 616,010 deaths were reported in the UK.*

Table 12: Examples of outputs obtained from the line graph-to-text task. Hallucinations are shown in **bold**.