StockGenChaR: A Study on the Evaluation of Large Vision-Language Models on Stock Chart Captioning

Le Qiu and Emmanuele Chersoni
Department of Language Science and Technology
The Hong Kong Polytechnic University
lani.qiu@connect.polyu.hk, emmanuele.chersoni@polyu.edu.hk

Abstract

Technical analysis in finance, which aims at forecasting price movements in the future by analyzing past market data, relies on the insights that can be gained from the interpretation of stock charts; therefore, non-expert investors could greatly benefit from AI tools that can assist with the captioning of such charts.

In our work, we introduce a new dataset *StockGenChaR* to evaluate large vision-language models in image captioning with stock charts. The purpose of the proposed task is to generate informative descriptions of the depicted charts and help to read the sentiment of the market regarding specific stocks, thus providing useful information for investors[1].

1 Introduction

In finance, technical analysis is the discipline that aims to predict future price trends based on historical data (Ponsi, 2016; Edwards et al., 2018). Analysts usually anticipate price directions by inspecting *stock charts*, where they are represented in the form of wave patterns (see Figure 1). Such patterns allow traders to identify potential entry or exit positions and make informed investment decisions.

Reading those charts, however, requires specific financial knowledge and expertise, and it can be challenging for non-expert traders to extract useful insights from merely visual cues. Therefore, we would like to propose framing this problem as an image captioning task where, given an image, an automatic system has to produce a description of its content (Vinyals et al., 2015). In our view, automating the manual interpretation process and producing descriptive texts for the charts would open up a shortcut to understanding the market dynamics for various groups of users, including experienced traders, novices, and individuals seeking support for their fast-paced trading activities.

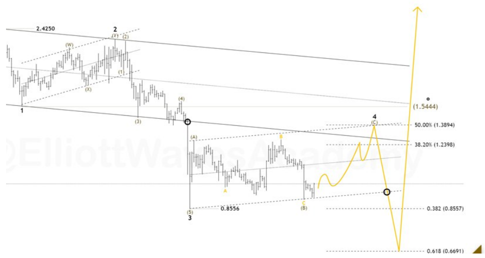


Figure 1: Future outlook for ARBUSD on the daily time frame based on Elliott-wave Theory (Elliott Waves Academy, 2024). The yellow line indicates the predictive movements of the stock.

Previous evaluation work on image captioning largely revolved around general-domain data, with relatively limited coverage of specialized domains [2]; and it typically aimed at summarizing the image content with just 1-2 concise sentences (Bernardi et al., 2016), which might not be ideal for the goals of technical analysis and financial decision-making. The image captioning setting needs to be adapted for stock chart reading: given an annotated chart image I, a system should generate a multi-sentence description C that provides a holistic narrative of the chart, covering the past movements and predictive trends, and ideally with trading advice. The generated text C is expected to be accurate and informative to lead to a well-grounded conclusion for the audience. Additionally, the text should remain as concise and comprehensible as possible, in order to be easily understandable even by less experienced traders.

To this purpose, we introduce *StockGenChaR*, a new dataset for the re-formulated stock-chart captioning task. To establish baseline performance

[1]The data and code will be made available on https://github.com/Laniqiu/GenChaR

[2]A summary of evaluation datasets for Image Captioning can be found in Table 4 in the Appendix.

Proceedings of The 10th Workshop on Financial Technology and Natural Language. EMNLP-2025, Suzhou, China

levels, we tested some representative LVLMs (i.e., LLMs with visual capabilities (Li et al., 2023b)) by using metrics that focus on different aspects of the generated texts, including n-gram overlap, semantic similarity, sentiment alignment, and accuracy of metadata information.

2 Related Work

Prior work around stock charts focused on utilizing the numerical data for goals such as financial return prediction and portfolio optimization (Hu et al., 2018; Kusuma et al., 2019; Ho and Huang, 2021; Norasaed and Siriborvornratanakul, 2024), while other studies made use of the graphical component in image or pattern recognition tasks (e.g., Velay and Daniel, 2018; Zheng et al., 2021). However, to our knowledge, the task of stock chart captioning has received limited attention so far.

The most recent approach to the image captioning problem consists of the vision-language pre-training approach (VLP). VLP models are pre-trained on a large amount of image-text pairs, and then fine-tuned for downstream tasks (Gan et al., 2022; Chen et al., 2023). Popular VLP models that can be applied to image captioning tasks include, for example, SimVLM (Li et al., 2019), OSCAR (Li et al., 2020b) and CLIP (Radford et al., 2021). Large Vision Language Models (LVLMs) can be considered as enhanced and ready-to-use versions of VLP models: in recent research work, models such as GPT-4 Vision (OpenAI, 2023), Gemini (Gemini Team Google, 2024), BLIP-2 (Li et al., 2023a) and LLaVa (Liu et al., 2024) proved their ability of successfully carrying out several multimodal tasks, including image captioning and visual reasoning (Li et al., 2023b; Zhang et al., 2024).

There have been examples of customized LVLMs that have exhibited some chart reasoning abilities. For example, Liu et al. (2023) developed MMCA, a MultiModal Chart Assistant achieving state-of-the-art performance on several chart question answering benchmarks; they also introduced a new and more challenging benchmark with nine different tasks evaluating reasoning capabilities over charts and plots, and reported that even the most sophisticated LVLMs have important limitations in interpreting charts. The works of Bhatia et al. (2024) and Xie et al. (2024) both introduced large instruction datasets for tuning LVLMs for the financial domain, together with two models, FinTral and FinLLaVA, that excel in solving tasks related to the interpretation of financial tables and charts.

Although such works challenge models in visual question answering on charts, we believe that framing stock chart understanding as an image captioning task would be closer to the needs of investors and practitioner in the financial industry, as image captioning could provide trend interpretations beyond the constrained setting of question-based benchmarks. To our knowledge, this type of task is not covered by any of the existing benchmarks for financial chart understanding.

In our study, we aim at filling this gap by building a new dataset for stock chart captioning, *StockGenChaR*. We will also present a systematic evaluation of the most commonly used LVLMs on the new benchmark.

3 Dataset Creation

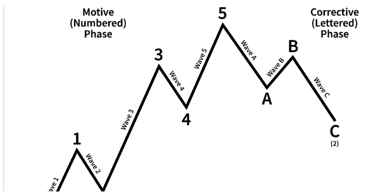


Figure 2: An EWP cycle.[3]

3.1 Sample Collection

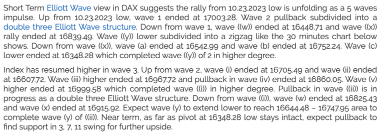
We identified an ideal source for this chart captioning task: the *ElliottWave-Forecast* website [4]. ElliotWave-Forecast is a worldwide top-notch technical analysis company, providing a wide range of coverage across about 80 markets, including Forex, Commodities, World Indices, and U.S. stocks & ETFs (ElliottWave-Forecast, 2024). The analyst team uses Elliott Wave Principle (EWP) as a major tool for chart analysis and offers forecasting and instructive guidance to its clients. EWP is a popular technical analysis approach: it is based on the belief that market prices have a tendency to move infinitely in a cycle (see Figure 2) in all time frames, exhibiting repetitive wave patterns (Poser, 2003). EWP provides the theoretical foundations for chart analysis and for the automatic completion of chart patterns within a specific timescale. The use of EWP also makes it easier to understand the charts and the captions, as it annotates the waves with

[3]Source: `https://www.investopedia.com/terms/e/elliottwavetheory.asp`

[4]`https://elliottwave-forecast.com`

the so-called *degrees* (the alphabetical or numeral indices along the wave patterns in Figure 2).

(a) Article



Short Term Elliott Wave view in DAX suggests the rally from 10.23.2023 low is unfolding as a 5 waves impulse. Up from 10.23.2023 low, wave 1 ended at 17003.28. Wave 2 pullback subdivided into a double three Elliott Wave structure. Down from wave 1, wave ((w)) ended at 16448.71 and wave ((x)) rally ended at 16839.49. Wave ((y)) lower subdivided into a zigzag like the 30 minutes chart below shows. Down from wave ((x)), wave (a) ended at 16542.99 and wave (b) ended at 16752.24. Wave (c) lower ended at 16348.28 which completed wave ((y)) of 2 in higher degree.

Index has resumed higher in wave 3. Up from wave 2, wave (i) ended at 16705.49 and wave (ii) ended at 16607.72. Wave (iii) higher ended at 16967.72 and pullback in wave (iv) ended at 16860.05. Wave (v) higher ended at 16999.58 which completed wave ((i)) in higher degree. Pullback in wave ((ii)) is in progress as a double three Elliott Wave structure. Down from wave ((i)), wave (w) ended at 16825.43 and wave (x) ended at 16915.92. Expect wave (y) to extend lower to reach 16644.48 – 16747.95 area to complete wave (y) of ((ii)). Near term, as far as pivot at 16348.28 low stays intact, expect pullback to find support in 3, 7, 11 swing for further upside.

(b) Textual content

(c) Chart image

Figure 3: A sample article from ElliottWave-Forecast in Fig. 3a, with the image and text displayed separately (EWFHendra, 2024). In the image captioning task, Figure 3c is the image input I, and Figure 3b the *gold* caption C.

The analysts publish their analysis (in English) together with chart images (see Figure 3a). For our dataset, we downloaded the articles released on the website by February, 2024. Each article was split into images and texts. We removed images that are not target charts, and texts and mark-up that are unnecessary, such as authorship, HTML tags, advertisements, and so on. So far, we only kept the samples containing one single stock chart (some articles may include two or more), to ensure a collection of one-to-one rather than one-to-many image-caption pairs. In addition, samples in which the text body is too long ($>$ 400 words) or too short ($<$100 words) have been excluded, according to the statistics on text length. The remaining stock chart-text pairs are our final dataset items, for a total of 1972 chart-caption pairs.

3.2 Chart Annotation

As shown in Figure 3c, stock charts map the price on the Y-axis against the time on the X-axis, and they typically come with several annotations. We categorized such annotations in four main types: Degree, Time, Price, and Add-on. *Degree* refers, roughly speaking, to a price movement; *Add-on* includes the information that is additionally applied to the charts, such as reminder messages and titles, while *Time* and *Price* are self-explanatory. We also annotated the endpoints of predictive patterns (categorized as *Point*) for further studies on automatic pattern completion[5]. The taxonomy and descriptions of the annotations are presented in Table 1.

[5]We did not annotate the endpoints for historical patterns, because they can be automatically generated with price data.

Table 1: A taxonomy of charting annotations. Here OHLC is used as a general term for *OHLC*, *Adj.* and *Volume* data. OHLC stands for Opening, Highest, Lowest, and Closing prices of a financial instrument during a timeframe, while *Adj.* is the adjusted closing price accounting for corporate actions, and *Volume* refers to the transaction amount.

Category	Description
Degree	EWP degrees
Time	X-axis ticks, timestamp, time markers
Price	Y-axis ticks, OHLC, price markers
Add-on	Supplementary indicators, annotations, and watermarks, etc.
Point	Endpoints of the predictive patterns

4 Evaluation with LVLMs

4.1 Model Choice

We ran evaluations with five recent general-purpose LVLMs that have showcased impressive capabilities in image captioning and visual question answering tasks (Li et al., 2023b; Zhang et al., 2024), including GPT-4V (OpenAI, 2023), mPLUG-Owl2 (Ye et al., 2023), LLaVA (Liu et al., 2024), InstructBLIP (Dai et al., 2024) and Gemini (Gemini Team Google, 2024). Each LVLM was prompted with the instruction below to produce candidate captions, and evaluated in a zero-shot setting. [6]

- Instruction: *Based on the chart image, generate a text around 100 to 400 words, describing the historical price movements and predictions and concluding the opinion of the chartist towards the stock trends.*

4.2 Evaluation Metrics

4.2.1 Text Similarity Metrics

Regarding evaluation metrics, we have considered the most popular ones, which are mainly based on n-gram overlapping, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and CIDEr (Vedantam et al., 2015). We also included two semantic-oriented metrics that are based on contextualized

[6]We had a pilot study with a few samples and some candidate instructions. The presented instruction achieved good performance overall. We observed that the few-shot learning setting is not applicable given the nature of the presented task, and thus we did not conduct the few-shot experiments

embeddings: BERTScore (Zhang et al., 2020), and paragraph-level cosine similarity, denoted as COS_F. After a pilot study, we selected the OpenAI embedding model as the best one to measure COS_F (see section A in the Appendix for more details).

4.2.2 Fine-grained Examination

Additionally, observing that the metrics above measure textual similarity in general, we also attempted to have a closer examination of the generated captions. To this purpose, we further analyzed the results in terms of *SA* and *IoU*.

IoU (Intersection over Union). We borrowed this metric from the field of object detection to measure how much important information has been included in the candidates relative to the references. In object detection, it measures the overlap over the union of predicted and ground-truth objects (Everingham et al., 2010). Different from its original formulation, an adaption has been made to fit into the settings of our work. In object detection, not all objects present in an image are annotated as ground truth. Also, a predicted object is only considered correct if it matches a labeled ground truth object.

In contrast, we assume that all metadata items present in the image are potentially valid and relevant for narration. Therefore, a candidate text should be rewarded whenever it correctly mentions an item from the image's full annotation — even if that item is not found in the reference text. Such *extra* information is still accurate and should neither be penalized nor neglected. Meanwhile, we attempt to bridge between the candidate and the reference. For this, we measure how the candidate text covers the objects compared to the reference text, while using the annotations as a *background* that has all possible metadata items. Let O_a be all metadata appeared in the image, and O_c and O_r be the metadata that is mentioned in the candidate and reference texts respectively. For each candidate-reference text pair (c, r), we first count how many objects they mention that also appear in O_a:

$$Hit_c = |O_c \cap O_a|, Hit_r = |O_r \cap O_a|$$

Then the relative coverage of c to r can be a ratio:

$$\frac{Hit_c}{|O_a|} / \frac{Hit_r}{|O_a|}$$

To further account for the impact of text length and large values, we redefine the IoU formula as: formulate IoU as:

$$IoU = \log\left(1 + \frac{\frac{|O_c \cap O_a|}{L_c}}{\frac{|O_r \cap O_a|}{L_r} + \epsilon}\right), \quad (1)$$

where L_c and Lr is the text length of c and r respectively, ϵ is a constant value to prevent division by zero and a logarithm could smooth the results.[7]

The dataset contains three categories of metadata i.e., the *metadata*, including *Degree*, *Time* and *Price*. For each sample, the IoU score is an aggregated result across these three categories. The IoU value would always be non-negative. A higher value may suggest that a more fine-grained description has been given, with more important information or meta information has been referred to. Conversely, lower scores may indicate underdescription, omission of key elements.

SA (Sentiment Alignment). The reference text states the historical movement of price, then outlines a prediction for future trends, indicating the existence of opinions, i.e., sentiment. An appropriate candidate is expected to express a similar sentiment in general. Accordingly, sentiment analysis can be performed on the reference and candidate texts to evaluate whether their sentiments were aligned. The assessment of sentiment alignment is presented as a text classification problem on three polarity categories, including *positive*, *negative* and *neutral*. The SA score of each reference-candidate pair is formulated as below:

$$\text{SA}(x, \hat{x}) = \begin{cases} 1, & \text{if } x = \hat{x} \\ 0, & \text{otherwise} \end{cases},$$

where x and $\hat{x}$ denote the reference and candidate text, and y and $\hat{y}$ denote their respective sentiments, which can be obtained using a BERT model fine-tuned for financial sentiment analysis [8], 1 indicates a correct sentiment alignment between x and $\hat{x}$ and 0 a mismatch.

Importantly, we don't perform sentiment analysis over the entire text. What is valued the most is the portion that contains opinions towards the future, which are typically located in the final part of the text. Therefore, our analysis only focus on the last a few sentences of a text. Given that this is an open-ended generation task and sentence-level alignment between reference and candidate

[7] ϵ takes the value of 1×10^{-6} during calculation.

[8] https://huggingface.co/ahmedrachid/FinancialBERT-Sentiment-Analysis

texts seems infeasible, we measure aggregated sentiments: The sentiment of each text segment is computed from the predicted labels of its individual sentences through a weighted voting approach, in which sentences closer to the end are assigned higher weights.

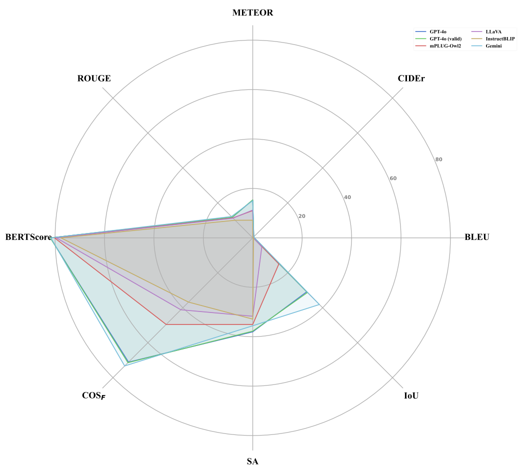


Figure 4: Radar chart of all evaluation scores. All values are presented on a 100% scale.

5 Results and Discussion

Table 2 presents the zero-shot evaluation results.

The instant observation is that all the LVLMs, despite their excellence in image understanding tasks, have achieved very low values for the evaluation metrics based on n-gram overlap (roughly the upper part of Table 2). The radar chart in Figure 4 presents the extremely unbalanced distribution between these common metrics and those introduced by us. Standard evaluation metrics for image captioning tasks — BLEU, ROUGE, CIDE$_r$ and METEOR — consistently report low scores for all models, even for the top ones like Gemini and GPT-4V. This is actually not unexpected: given the domain specificity and the length of the reference texts, it would have been surprising - and possibly an index of data contamination - to observe high values for such metrics.

For SA and IoU, GPT-4V and Gemini outperform the other models. It should be noticed that Gemini scores are much better on the IoU metric, meaning that its responses are more likely to contain more of the correct chart metadata, and has better coverage. In general, LVLMs achieve higher scores in BERTScore and COS$_F$, showing that even if they do not use the same words (i.e. low n-gram overlap), the texts generated by some of the models have high conceptual similarity with the reference. On the other hand, the SA scores, reflecting the alignment between model outputs and opinions of the analysts, have only around 40% in the best models. Higher values would be desirable, as this metric might be the most relevant in affecting the stock buying decisions of the investors and we would like LVLMs to be as much aligned as possible with human experts' insights, in order for them to be deployed in real-world applications.

Examples are provided in Table 6 (in the Appendix). From them, we can find that the reference text and model output appear to differ in narration. In this sense, the low scores are quite understandable, as these metrics in nature rely on lexical overlapping. While these methods may perform well on regular datasets, where texts are in short forms and show limited possibilities for paraphrasing, their effectiveness can be predictably compromised in this context which presents a open-ended long text generation task. The extremely low values of BLEU or CIDEr suggest they could be inappropriate for the evaluation of this task. Nonetheless, despite their variance in wording and constructions, some model outputs — such as those from Gemini and GPT-4V as listed — remain faithful to the given chart and provide meaningful interpretations for readers. Therefore, relying solely on traditional metrics would seem unfair. In this sense, the use of alternative metrics — BERTScore and COS$_F$ seem inevitable. BERTScore, according to its formulation (please refer to Zhang et al. (2020) for details), measures lexical similarity using contextualized embeddings from Transformers, which is more flexible and effective compared to its precedents. COS$_F$, on the other hand, encodes the entire text as vectors and then measures semantic similarity in a more blunt manner.

Additionally, we observe that the reference texts which are produced by professional analysts not only uncover the future price trends, but also try to disclose the full picture of the movement, i.e., to provide detailed context and justification for their description or analysis, making it more reliable. During the process, EWP markers (i.e., the numerical and alphabetical annotations around the waves) are frequently referred to, which in some way enhances the accessibility of the text. In contrast, the performance of LLMs varies significantly. Strong models like Gemini and GPT-4V appear to be aware of the EWP theory and able to recognize the visual elements such as markers and labels on

Table 2: A summary of the evaluation results on the dataset (1972 instances). Several images are rejected by GPT-4V, due to sensitivity concerns or other reasons such as image quality, according to the feedback from the model. The *Valid* column contains scores for valid captions only (1915 instances). Out of 1972 GPT-4V responses, 57 are found invalid. 4-gram scores are reported for BLEU and CIDEr, and F-score for BERTScore and for SA. The best scores are indicated in **bold** among all samples or <u>underlined</u> among valid samples. For all metrics, scores are reported in percentage (%) and rounded half up to two decimal precision and higher values indicate better performance.

	GPT-4V		mPLUG-Owl2	LLaVA	InstructBLIP	Gemini
	Overall	Valid				
BLEU	.65	.66	.44	.49	.25	**0.73**
CIDEr	.58	.60	.26	.24	.05	**0.62**
METEOR	**15.32**	<u>15.36</u>	11.05	11.24	7.14	15.05
ROUGE$_L$	11.74	11.78	11.39	11.28	10.07	**12.20**
BERTScore	**81.91**	<u>81.94</u>	80.22	79.51	77.63	81.71
COS$_F$	71.10	71.46	49.56	41.16	36.77	**73.32**
SA	**38.04**	37.74	35.04	31.73	32.89	35.58
IoU	30.88	31.31	14.94	5.19	.43	**38.15**

the chart. In contrast, weak models could even fail in recognition. For example, the red-highlighted sentence in Table 6 indicates that the LLAVA model could not recognize the time frame of the chart. From its generated textual description, nor can we find traces that the model understands the EWP theory or master the skills of giving suggestions for investment. Text-similarity metrics such as BERTScore, ROUGE, may fail to capture these subtle distinction in their measurement, and that is the reason that IoU is introduced in the evaluation.

Table 3: A breakdown of the IoU scores. The reported values (in %) represent the metadata coverage within different categories. For instance, suppose n Time markers are included in the full annotation and m found in the reference text, the *Time* coverage of the reference is then calculated as $\frac{m}{n}$.

	Degree	Time	Price
Reference	26.83	20.09	4.65
GPT-4V	3.72	7.07	5.83
GPT-4V (valid)	3.69	7.16	5.86
mPLUG-Owl2	.26	4.79	.22
LLaVA	.01	1.58	.05
InstructBLIP	.00[9]	.13	.01
Gemini	5.47	12.01	5.79

Besides from the overall scores presented in Table 2, Table 3 provides more statistics around IoU. These objective and quantitative results support our observation that human chartists tend to favor including metadata especially wave degrees, while LLMs are less attentive in their generation.

[9]The actual value is above zero, but it is displayed as zero due to standard rounding.

Regarding SA, we reported its F1-scores in Table 2. The confusion matrices of sentiment alignment are revealed in Figure 6 in the Appendix. A majority of errors occur between the *neutral* class and the others. Although the sentiment analysis model could be to blame, we noticed that subtle mismatches between the narrative tones could have contributed to the low scores. For instance, as shown in Table 5 (in Appendix), the reference text expresses a positive outlook towards the stock, while Gemini, although agrees, adopts a more cautious and restrained tone.

Based on these findings, it is evident that the some LLMs could have posed the capability of stock chart interpretation and demonstrate a certain level of financial domain knowledge, even in cases where their overall performance is not satisfying.

In the future, we believe that the performance can be further improved via a more systematic search for optimal prompt instructions, and possibly by introducing customized architectures for the task.

6 Conclusion

In this work, we have introduced StockGenChaR, a dataset for stock chart captioning, and we have reported the preliminary work on the captioning task, including the creation of the benchmark and a preliminary evaluation with some popular LVLMs. By first exploring these LVLMs, we hope to find out their capacity in this stock chart captioning without additional fine-tuning, and also to identify proper evaluation methods for the task.

The current findings suggest that these LVLMs could have possessed limited capabilities of stock chart captioning. However, for practical deployment in financial scenarios, task-specific fine-tuning is still required. Also, given the sensitive nature of financial data, directly using open-sourced models may pose risks related to confidentiality or information breach. Also, despite our efforts on evaluation metrics, automatic evaluation metrics alone may be insufficient to capture all aspects of the text quality, especially in our setting, where the task involves generating long-form texts intended for human readers, including amateurs to finance and stocking. In this sense, further explorations with customized approaches and the inclusion of human evaluation seem necessary. On the bright side, we have seen that these LVLMs, close-sourced or open-sourced, are equipped with the visual recognition and textual generation ability. It is therefore safe to assume that the performance on the presented task can be further improved via a more systematic search for optimal prompt instructions, and possibly by introducing customized architectures.

Limitations

Our work has some limitations that have to be acknowledged. First, the paper presents only preliminary evaluation results with general-purpose LVLMs. For future work, we plan to experiment with more customized LVLM architectures (e.g., those mentioned in Section 2) to further push the boundaries of model performance on the proposed chart captioning task.

Also, the present work only adopts automatic metrics for evaluation. From Table 2, we can find that metrics that are based on word overlapping such as BLEU and CIDEr, are insufficient in evaluating long texts; the two embedding-based methods, BERTScore and COS_F, despite capturing semantic similarity, demand further examination because they could fail to measure the user *accessibility* of the generated texts, as the task requires. In the future, we plan to explore more sophisticated prompting strategies to explicitly target accessibility and other *desiderata* aspects of the generated captions, and to include human annotators for evaluating the generated texts.

Acknowledgements

The authors acknowledge the support from the project "Analyzing the semantics of Transformers representations for financial natural language processing" (ZVYU), funded by the Faculty of Humanities of the Hong Kong Polytechnic University (PolyU-UGC). We would also like to thank our reviewers for the constructive feedback.

Ethical Considerations

The models and data used in this study are intended for research purposes only. No personally identifiable information or sensitive content is involved. Therefore, we believe the ethical risk of this work is minimal.

We utilized several models in this study, each governed by its individual license — LLaVA is released under the Apache License 2.0, mPLUG-Owl2 under the MIT License, and InstructBLIP under the CC BY-NC 4.0 License. For other models and the data we collected without publicly specified licenses, we used them in accordance with the terms of service or usage guidelines by their original provider, where available. Additionally, ChatGPT was employed as a writing assistant under API terms for translation and grammar checking purposes.

The annotators were recruited online from Mainland China. All had undergraduate or master's degrees in computer-related disciplines and were employed in relevant industries. Only individuals who passed a qualification test were selected to participate in the annotation task. All annotators provided informed consent, received compensation on a per-annotation basis in accordance with local labor standards, and retained the right to withdraw from the study at any time.

This study has been approved by the Institutional Review Board (IRB) from the Department of Language Science and Technology of the Hong Kong Polytechnic University.

Acknowledgement of Data Usage

The data used in this research was obtained from Elliott Wave Forecast with the necessary permission for usage. The provider has explicitly granted consent, ensuring compliance with relevant legal, ethical, and regulatory requirements. We affirm that the data will be handled responsibly and utilized strictly within the agreed scope.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel Object Captioning at Scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

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*.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models. *arXiv preprint arXiv:2402.10986*.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context Driven Entity-aware Captioning for News Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. VLP: A Survey on Vision-language Pre-training. *Machine Intelligence Research*, 20(1):38–56.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards General-purpose Vision-language Models with Instruction Tuning. *Advances in Neural Information Processing Systems*, 36.

Robert D Edwards, John Magee, and WH Charles Bassetti. 2018. *Technical Analysis of Stock Trends*. CRC Press.

Elliott Waves Academy. 2024. Future Outlook for the Digital Currency ARBUSD on the Daily Time Frame. Online; accessed on 13 June, 2024 from Twitter.

ElliottWave-Forecast. 2024. About Us. `https://elliottwave-forecast.com/about-us/`. Accessed: 04-Feb-2024.

Kawin Ethayarajh. 2019. How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of EMNLP*.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88:303–338.

EWFHendra. 2024. DAX Near Term Support Area. `https://elliottwave-forecast.com/news/dax-near-term-support-area/`. Accessed: 05-Feb-2024.

Zhaoxin Feng, Jianfei Ma, Emmanuele Chersoni, Xiaojing Zhao, and Xiaoyi Bao. 2025. Learning to Look at the Other Side: A Semantic Probing Study of Word Embeddings in LLMs with Enabled Bidirectional Attention. In *Proceedings of ACL*.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating Attractive Visual Captions with Styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, and 1 others. 2022. Vision-language Pre-training: Basics, Recent Advances, and Future Trends. *Foundations and Trends in Computer Graphics and Vision*, 14(3–4):163–352.

Gemini Team Google. 2024. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2023arXiv231211805G*.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *International Workshop OntoImage*, volume 2.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning Images Taken by People Who Are Blind. In *Proceeding of the European Conference of Computer Vision (ECCV)*.

Trang-Thi Ho and Yennun Huang. 2021. Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation. *Sensors*, 21(23):7957.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Guosheng Hu, Yuxin Hu, Kai Yang, Zehao Yu, Flood Sung, Zhihong Zhang, Fei Xie, Jianguo Liu, Neil Robertson, Timpathy Hospedales, and 1 others. 2018. Deep Stock Representation Learning: From Candlestick Charts to Investment Decisions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2706–2710. IEEE.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123:32–73.

Rosdyana Mangir Irawan Kusuma, Trang-Thi Ho, Wei-Chun Kao, Yu-Yen Ou, and Kai-Lung Hua. 2019. Using Deep Learning Neural Networks and Candlestick Chart Representation to Predict Stock Market. *arXiv preprint arXiv:1903.12258*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of EMNLP*.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled Transformer for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8928–8937.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping Language-image Pre-Training with Frozen Image Encoders and Large Language Models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742. PMLR.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, and 1 others. 2020b. Oscar: Object-semantics Aligned Pre-training for Vision-language Tasks. In *Computer Vision–ECCV*, pages 121–137. Springer.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating Object Hallucination in Large Vision-language Models. *arXiv preprint arXiv:2305.10355*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in COntext. In *Proceedings of Computer Vision (ECCV)*, pages 740–755. Springer.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023. MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. *arXiv preprint arXiv:2311.10774*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated Flower Classification over a Large Number of Classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE.

Witawat Norasaed and Thitirat Siriborvornratanakul. 2024. Market Movement Prediction Using Chart Patterns and Attention Mechanism. *Discover Analytics*, 2(1):1.

OpenAI. 2023. GPT-4V(ision) System Card. *Published on: openai.com*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing Images Using 1 Million Captioned Photographs. *Advances in Meural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*.

Ed Ponsi. 2016. *Technical Analysis and Chart Interpretations: A Comprehensive Guide to Understanding Established Trading Tactics for Ultimate Profit*. John Wiley & Sons.

Steven W Poser. 2003. *Applying Elliot Wave Theory Profitably*, volume 169. John Wiley & Sons.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR.

Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. 2017. Breakingnews: Article Annotation by Image and Text Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1072–1085.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning. In *Proceedings of ACL*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A Dataset for Image Captioning with Reading Comprehension. In *Proceedings of the European Conference of Computer Vision (ECCV)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE cConference on Computer Vision and Pattern Recognition*.

Marc Velay and Fabrice Daniel. 2018. Stock Chart Pattern Recognition with Deep Learning. *arXiv preprint arXiv:1808.00418*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-UCSD Birds 200. *California Institute of Technology, Technical Report*.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, and 1 others. 2024. Open-FinLLMs: Open Multimodal Large Language Models for Financial Applications. *arXiv preprint arXiv:2408.11878*.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. In *Proceedings of the European Conference of Computer Vision (ECCV)*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. *arXiv preprint arXiv:2311.04257*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in Multimodal Large Language Models. *arXiv preprint arXiv:2401.13601*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of ICLR*.

Yuechu Zheng, Yain-Whar Si, and Raymond Wong. 2021. Feature Extraction for Chart Pattern Classification in Financial Time Series. *Knowledge and Information Systems*, 63(7):1807–1848.

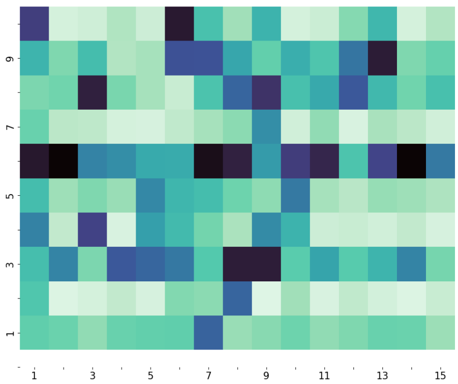
A Embedding Model Selection

Following Li et al. (2020a), we designed a similarity experiment to find the ideal embedding model:

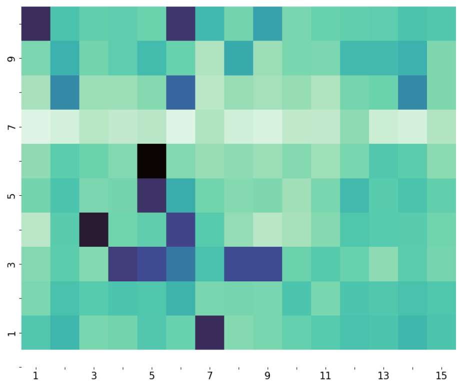
- First, we collected generated texts from GPT-4V using randomly-picked samples and designed instructions, and then created gold label similarities between pairs of gold truth captions and generated texts (The gold label similarity is actually a pseudo similarity. Currently, we haven't conducted a standardized human evaluation of the generated texts. But GPT-4V sometimes produces invalid responses due to sensitivity concerns or other reasons, so we take the extreme values by assigning a similarity score of 0 between the caption and invalid response pair, and a score of 1 for other cases.);
- We obtained text embeddings from each candidate embedding model;
- We computed COS_F scores between each paired embeddings as the model-wise predicted similarities, and then calculate the Spearman's correlation coefficients between them. We experimented with BERT, RoBERTA and the OpenAI embedding models, and reported the COS_F and Spearman scores in Figure 5. An ideal model is supposed to exhibit strong correlation with the gold labels. Additionally, it should be able to differentiate between different instructions, meaning that the predicted similarities should vary upon the provided instructions. Based on these criteria, the OpenAI embedding model outperformed the others. Besides, considering its larger token window (1536 tokens), we decided to select the OpenAI embedding model as the embedding source for the COS_F metric.

B Supplementary Tables and Figures

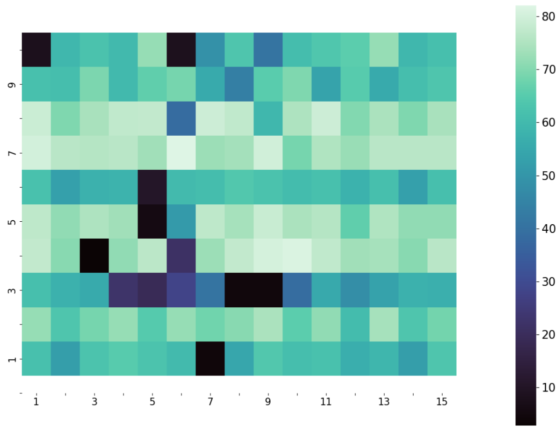
[10]The number of captions per image.



(a) BERT
r_s. = 27.48
Avg. = 90.97



(b) RoBERTa
r_s. = 44.82
Avg. = 99.74



(c) OpenAI model
r_s. = 47.23
Avg. = 61.79

Figure 5: COS_F scores are presented in heatmaps, where the X-labels indicate different instructions and Y-labels different samples. The Spearman's correlation coefficients are denoted as r_s., with p-value $< 0.1\%$ in all cases, while Avg. indicates the average COS_F score (notice that the OpenAI model is the one with the highest correlation while having at the same time a lower average cosine score. The other two models seem to have a high level of anisotropy of the vector space, cf. Ethayarajh (2019); Feng et al. (2025), assigning similarities close to 1 to most text pairs). Regarding model versions, we chose *bert-base-uncased* for BERT, *rflike berta-base* for RoBERTa, and *text-embedding-3-small* for the OpenAI model. All scores are reported in percentage.

Dataset	Domain	Total	Caps.[10]	Source
Conceptual Caption (Sharma et al., 2018)	Generic	3.3M	5	Web
MS COCO (Lin et al., 2014)	Generic	328K	5	Web
Flickr30k (Young et al., 2014)	Generic (people, animals)	31K	5	Flickr.com
Flick 8K (Hodosh et al., 2013)	Generic	8K	1-5	Flickr.com
FlickrStyle10K (Gan et al., 2017)	Generic	10K	2	Flickr.com
SBU Captions (Ordonez et al., 2011)	Generic	1M	5	Web
Visual Genome (Krishna et al., 2017)	Generic	108K		Web
VizWiz Captions (Gurari et al., 2020)	Assistive	39K	5	VizWiz APP
CUB-200 (Welinder et al., 2010)	Birds	12K	10	Web
Oxford-102 (Nilsback and Zisserman, 2008)	Flowers	8K	10	Web
Fashion Captions (Yang et al., 2020)	Fashion	52K	5	Web
BreakingNews (Ramisa et al., 2017)	News(sports, arts, etc.)	100K	5	Web
GoodNews (Biten et al., 2019)	News	466K	1	New York Times
SentiCap (Mathew et al., 2022)	Generic	3.2K	6	MS COCO
TextCaps (Sidorov et al., 2020)	OCR	28.4K	5-6	Web
nocaps (Agrawal et al., 2019)	Generic	15.1K	11	Web
IAPR TC-12 (Grubinger et al., 2006)	Generic	20K	1-5	Viventura
PASCAL 1K (Rashtchian et al., 2010)	Generic (people, animals)	1K	5	PASCAL VOC

Table 4: A general summary of the benchmark datasets for image captioning

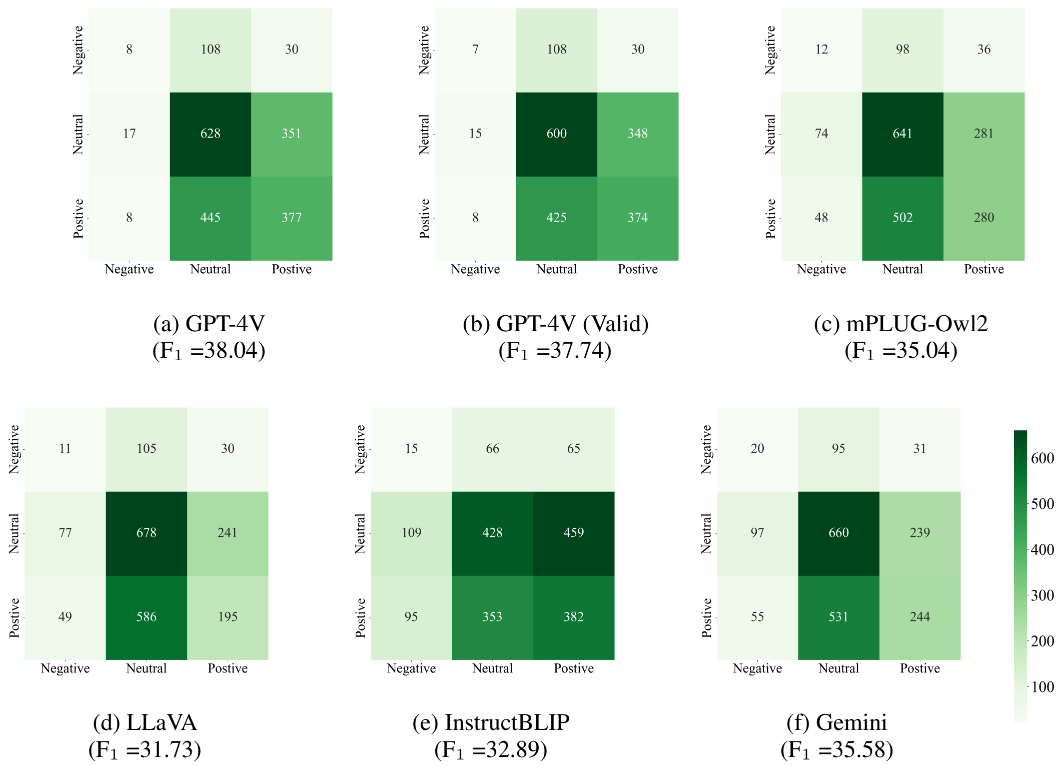
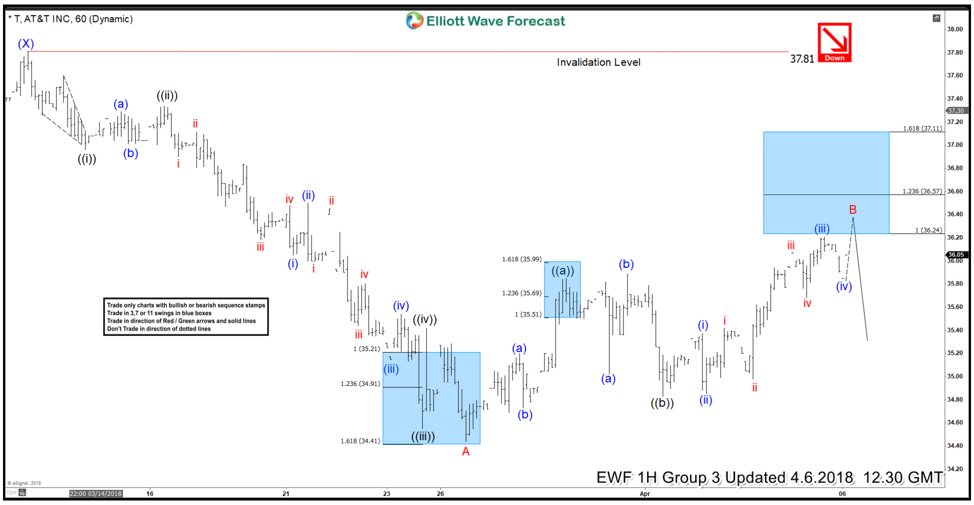


Figure 6: Confusion matrices of SA. Macro F_1 scores of the models are presented in the brackets, reported on a 100% scale.

Table 5: An example of the false neutral case from Gemini (i.e., a positive sample is classified as a neutral). Only the last a few sentences are presented.

Reference: "... The internal structure of wave Z is unfolding as a double three structure where wave (w) of Z ended at \$53.79 and wave (x) of Z has ended at \$52.08. Near term, the rally from wave (x) low ended at \$54.37 as 5 waves, and CL_F is expected to do a wave b pullback to correct the rally from \$52.08 low in 3, 7, or 11 swing before the rally resumes towards \$55.93 – \$56.83. We don't like selling the proposed pullback and expect buyers to appear again once wave b pullback is complete in 3, 7, or 11 swing as far as pivot at \$52.07 remains intact in the first degree. If pivot at \$52.07 fails, then CL_F has ended cycle from \$49.95 low and will do a larger correction and still expected to turn higher while second degree pivot at \$49.95 stays intact."

Gemini: ... The identified wave structure combined with critical support levels provides a framework for possible upward momentum. Traders and investors are advised to watch for a rebound from the defined support levels, which could signify a continuation of the bullish trend established earlier in the cycle. This analysis reinforces a strategic approach, encouraging trades in alignment with the identified Elliott wave patterns and Fibonacci levels for optimal risk management.



Reference: AT&T ticker symbol: ($T) short-term AT&T Elliott Wave view suggests that the bounce to 313 high (37.81) ended Intermediate wave (X) bounce. Below from there, intermediate Y leg lower remains in progress as a Zigzag correction. Where initial decline unfolded in 5 waves Elliott Wave Impulse Sequence with extension in Minor wave A. Which can be seen in the Chart below, with the internal distribution of each leg consisted of 5 waves structure thus it favored being impulse. Now let's take a quick look at the short-term price action in the stock from 313 peak (37.81). The decline from there can be seen in 5 waves impulse sequence, where internals of each leg is consist of 5 waves structure. Where Minute wave ((i)) ended at 36.97, Minute wave ((ii)) ended at 37.34 and Minute wave ((iii)) ended at 34.55. Then from there, Minute wave ((iv)) ended at 35.29 and Minute wave ((v)) of A ended at 34.44 low. Up from there, the stock is correcting the 313 cycle (37.81) in Minor wave B bounce. And expected to find seller's in 3, 7 or 11 swings as far as the pivot from 313 peak (37.81) stays intact. So far the bounce looks to be in 3 swings as a Flat structure, where Minute wave ((a)) ended at 35.89, Minute wave ((b)) ended at 34.84 and Minute wave ((c)) of B remains in progress towards 36.24-36.57 100%-123.6% Fibonacci extension area of ((a))-((b)). Before finding seller's again for another extension lower ideally provided the pivot from 37.81 high remains intact. We don't like buying it and intraday traders can look to sell it from an above-mentioned 100%-123.6% ext area for 3 wave reaction lower at least. Keep in mind that the market is dynamic and the FTSE Elliott Wave view could change in the meantime...

Gemini: This Elliott Wave Forecast chart for AT&T INC (60-minute timeframe) depicts historical price movements and projects future trends using Elliott Wave principles. The chart shows a complex pattern, marked by distinct impulsive and corrective waves labeled with Roman and lowercase letters respectively. The analysis identifies two completed corrective patterns labeled 'A' and 'B'. 'A' shows a corrective three-wave structure, followed by an impulsive five-wave structure labeled 'B'. The 'B' wave appears to have reached a potential termination point, suggesting a significant price reversal is imminent. A dashed line indicates a potential invalidation level at approximately 37.81. If the price breaks above this level, the bearish projection would be invalidated. The chart's creator has drawn blue boxes around specific wave structures, suggesting these areas represent significant price swings suitable for trading. Red and green arrows indicate the predicted direction of price movement; dotted lines, however, should be avoided for trading signals. The analyst recommends trading only charts displaying clear bullish or bearish sequence stamps and focusing on 3, 7, or 11 swings within the blue box areas. The chartist's overall opinion is bearish. The structure of wave 'B' and its potential termination point suggest that a downward trend is likely to follow. The invalidation level at 37.81 provides a clear benchmark for assessing the validity of this bearish prediction. A break above this level would signal a significant shift in the underlying price trend and invalidate the current bearish outlook. The analyst emphasizes the importance of adhering to the specified trading rules, particularly avoiding trades counter to the direction of the solid and directional arrows.

GPT-4V: The chart displaying the historical price movements of AT&T Inc. (Ticker: T) utilizes the Elliott Wave theory to forecast future trends based on observed patterns in the price movements. The analysis reveals the stock has experienced significant volatility, characterized by both upward and downward movements delineated by specific wave formations. Initially, the stock faced a bearish trend, denoted by the downward movements labeled as waves (i), (ii), and (iii). This initial decline appears to have established a low point around 34.41, suggesting potential support levels. Subsequently, the price exhibited a corrective wave structure wherein it managed to regain some value, as indicated by the formations tagged as ((iii)), ((iv)), and subsequent retracements represented by (a) and (b). The chart depicts key Fibonacci retracement levels, illustrating potential areas of resistance and support. For instance, the (1) level at 35.51 indicates a notable area where the price experienced fluctuations, while the 1.236 and 1.618 levels offer insights into possible price targets for the next wave movements. The areas shaded in blue emphasize critical zones for potential trading decisions, guiding market participants on optimal entry and exit points. Moving forward, the chartist's...

LLAVA: The chart image displays a historical price movement of a stock, with a red line indicating a downward trend. The stock has been experiencing a decline since the beginning of the year, with a notable drop in January. The chartist's opinion on the stock trends is that the stock is likely to continue its downward trajectory, as the current trend is not showing any signs of reversal. The chart also features a forecast for the stock's future performance, with a red line indicating a potential decline. The chartist believes that the stock's current downward trend is likely to continue, and the forecast suggests that the stock may reach a lower low in the near future. The chartist's conclusion is that the stock is not a good investment at the moment, as the current trend is not favorable for long-term investors. The stock's historical price movements and the forecast indicate that it may continue to decline, making it a risky investment option.

Table 6: We selected a chart image and presented it with its reference description and several model output for comparison. The chart presents a 1-hour short term movement of AT&T ticker.