Weak-to-Strong Generalization on Financial News Summarization

Yuyang Jiang
Department of Statistics
University of Chicago

An Hui Chang
Department of Statistics
University of Chicago

Abstract

In this project, we investigate the applicability of the weak-to-strong generalization framework in the context of financial news summarization. We constructed a dataset of 2,000 financial news records and assessed the performance of OpenAI's framework from lexical and pragmatic viewpoints. Our findings indicate that: (1) the weak-to-strong generalization does not hold in the context of financial news summarization for the model classes tested; (2) while the summarization models like BART series more closely align with the ground truth compared to the text generation model LLaMA-1, they are less effective in producing financially informative summaries.

1 Introduction

The initial motivation for this project is that weak-to-strong supervision is one of promising solutions to resolve the failure of human supervision. Human supervision will not only fail or become highly time- or money-consuming for complex tasks and superhuman models [1] but also risk the potential bias. [5] And this failure will cause significant impact in data-intensive fields like finance domain.

For an investor, "information" from tons of financial technical reports (like news, statements, announcements, etc.) is one of the most important factors in helping them decide how to allocate investment resources to different underlying assets. [2] Considering this, we believe a good point to start with is a finance-specific summarization task in the setting of "weak-to-strong" first to test the potential of applying this framework to financial domains.

Based on our research question, we first collect and clean raw financial news to build up our own dataset. Then following OpenAI's framework, we test weak-to-strong frame on different model pairs. Finally, we evaluate the generalization from both lexical and pragmatic performance.

Our key findings include:

1. Weak-to-strong fails to generalize in the setting of financial news summarization given our chosen model class. Weak-to-strong's performance is strongly affected by the ability of weak supervisors.
2. Summarization model (BART series) is better than text generation model (LLaMA-1) on generating summarization aligned with groundtruth.
3. Text generation model (LLaMA-1) can generate more financially informative summarization than BART series.

2 Related Work

Textual Analysis Among the many applications of Large Language Models (LLMs) in the financial domain, textual analysis has been widely utilized and applied to financial market analysis. Kao, Ma,

Preprint. Under review.

Ng, and Chua (2024) [4] introduced the Summarize-Explain-Predict (SEP) framework, which enables an LLM to teach itself how to generate explainable stock predictions.

Weak-to-Strong Generalization As it is a mature field in finance, it is ideal to apply the novel weak-to-strong generalization framework proposed by OpenAI's Burns et al. (2023) [1]. This framework allows us to determine whether a weak supervisor can effectively supervise the strong student's performance.

Evaluation Metrics To evaluate the model's performance, we referred to both computer science and economics empirical works to get the best of both worlds. From a purely technical perspective, the evaluation design should measure how precisely the model summarizes the news content, that is, how much similarity the summarization shares with the original news. We utilized ROUGE scores to measure this similarity. However, as ROUGE scores have their limitations, as pointed out by Zhou and Tan (2023) [7], we considered additional metrics.

The TF-IDF score and NLTK scores are feature extraction methods that transform textual information into numerical representations, measuring how well the models capture market sentiment. In Ke, Kelly, and Xiu's (2019) [3] study, NLTK enabled key text preprocessing and tokenization steps to convert the raw news articles into analyzable word count vectors, and TF-IDF provided a way to weight word importance when aggregating sentiment word counts into article-level scores. In our design, we calculated the Pearson correlation coefficients of NLTK and TF-IDF scores with the stock price movement minus the general market movement, which is the alpha of the stock as suggested by the classic Capital Asset Pricing Model (CAPM) theory [6].

3 Data and Methodology

3.1 Data

This study uses Benzinga news website and its cooperater Alpaca API, Our input data consists of news content, and our ground truths (GTs) are news headlines, specifically focusing on Tesla (TSLA). We prepared train and test datasets, each containing 2000 entries. We have formalized the data preprocessing process to ensure the quality and relevance of the data. First, we eliminated any news containing tickers of companies other than Tesla. Second, we cleaned the data by normalizing the text and removing news with no content, redundant words, and punctuation. Third, we calculated the 20 most frequent words, which were more sensible this time, including terms like "production" and "delivery." Lastly, we filtered out news articles that lacked substantial content or did not contain any of the 20 most frequent words.

3.2 Experiment Setup

In our dataset setting, we used a training set of 2000 entries, which we split into two equal parts: A and B. Our test set also consists of 2000 entries. This setup is similar to the model setting used by OpenAI. We trained a weak model on set A and used the test set to obtain weak baseline results. Next, we trained a strong model, termed weak-to-strong, on the weak model's generation of set B and used the test set to obtain weak-to-strong results. Finally, we trained a strong model on set B and used the test set to get strong ceiling results.

For model choice, we have designed different pairs of weak-to-strong models to test the generalization ability of the models and their corresponding weak-to-strong effects. We used GPT-2-XL as the weak model and BART-Large, BART-CNN-Large, or LLAMA as the strong models. Additionally, we tested BART-Large and BART-CNN-Large as weak models with LLAMA-7B as the strong model. This setup allows us to evaluate the effectiveness of the weak-to-strong framework across various model combinations.

3.3 Evaluation Framework

Language Metrics ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a widely used set of metrics for evaluating the quality of text summaries generated by automatic summarization systems. Here are the ROUGE metrics we implemented in our experiments:

Table 1: An Example of Automatic Summarization on Financial News

Headline (GT)	**elon musk claims many improvements for tesla cybertruck before release**	**Relevant[Y/N]**
GPT-2-xl	if you like the content on this site please share it with your friends on social media and leave us feedback.	N
BART	elon musk says cybertruck may be better than what we showed	Y
GPT-2-xl → BART	this article was generated by benzinga automated content engine and reviewed by an editor	N
BART-cnn	elon musk says cybertruck is going to be better than what we showed it may be in time for deliveries by end of year but could be pushed back if gigafactory texas isn't fully operational by next year	Y
GPT-2-xl → BART-cnn	this is an instruction that describes a task, paired with an input that provides further context. write a response that appropriately completes the request	N
LLaMA-1-7b	elon musk confirms tesla cybertruck production at gigafactory texas will be better than what we showed	Y
GPT-2-xl → LLaMA-1-7b	this is an instruction that describes a task, paired with an input that provides further context. write a response that appropriately completes the request	N
BART-cnn → LLaMA-1-7b	tesla ceo elon musk says cybertruck may be better than what we showed so far here what he says about the upcoming model and why he says it will be produced at the new gigafactory texas by next may if there are any unforeseen problems in gigafactory construction or vehicle design and assembly	Y
BART → LLaMA-1-7b	elon musk says cybertruck will be better than what we showed	Y

1. **ROUGE-1** measures the overlap of unigrams (single words) between the generated summary and a set of reference summaries. It captures the basic content similarity at the word level. High ROUGE-1 scores indicate that the generated summary includes many of the same words as the reference summaries.

2. **ROUGE-2** evaluates the overlap of bigrams (pairs of consecutive words) between the generated summary and the reference summaries. It provides a deeper level of comparison by considering word pairs, which helps in assessing the fluency and coherence of the summary. High ROUGE-2 scores suggest that the summary preserves some of the original text's context and phrasing.

3. **ROUGE-L** measures the longest common subsequence (LCS) between the generated summary and the reference summaries. Unlike ROUGE-1 and ROUGE-2, which focus on exact n-gram matches, ROUGE-L captures sentence-level structure similarity. It is useful for evaluating summaries where maintaining the order of information is important.

4. **ROUGE-LSum** is a variant of ROUGE-L specifically designed for summarization tasks. It computes the LCS based on sentence-level matching rather than word-level matching, making it particularly suitable for comparing the overall structure and coherence of long summaries.

Pragmatic Metrics To assess the sentiment and relevance of text data, we utilize both TF-IDF scoring and NLTK sentiment analysis, combined with a sentiment-weighted sum approach.

1. NLTK: a library in Python for natural language processing (NLP). It provides easy-to-use interfaces to over 50 corpora and lexical resources, such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. NLTK can do text analysis, sentiment analysis, and machine learning in NLP.

Table 2: Rouge2 Performance of Weak-to-Strong on Financial News Summarization. Model Performance: BART > LLaMA-1-7b > BART-cnn » GPT-2-xl.

Experiment	Weak Baseline	Weak-to-Strong	Strong Ceiling
GPT-2-xl → BART-cnn	0.001	0.001	0.075
GPT-2-xl → BART	0.001	0.000	0.167
GPT-2-xl → LLaMA-1-7b	0.001	0.000	0.114
BART-cnn → LLaMA-1-7b	0.075	0.059	0.114
BART → LLaMA-1-7b	0.167	0.135	0.114

2. TF-IDF: It stands for Term Frequency-Inverse Document Frequency, is a widely used statistical measure in information retrieval and text mining to evaluate the importance of a word in a document relative to a collection of documents (or corpus). The TF component measures how frequently a term appears in a document, while the IDF component assesses how unique or rare the term is across the entire corpus. By combining these two measures, TF-IDF highlights words that are significant to a particular document while downplaying common words that appear frequently across many documents.

- First, we define our vocabulary and sentiment dictionaries: a positive word dictionary with scores such as 'a': 0.1, 'b': 0.2 and a negative word dictionary with scores such as 'd': 0.1, 'e': 0.2.
- Secondly, we calculate the weighted sum of scores, we assign weights based on sentiment, with positive words receiving a weight of 1 and negative words a weight of -1. For example, using the formula: $I = 1 \times 0.1 + 1 \times 0.2 + (-1) \times 0.1$, we compute the weighted sum of the score.

Performance Gap Recovered (PGR) It measures the fraction of the performance gap between the weak supervisor model and the strong student model ceiling performance that the student model recovers when trained on the weak labels. Define the following:

1. Weak performance: Accuracy of the weak supervisor model.

2. Weak-to-strong performance: Accuracy of the strong student model trained on weak labels.

3. Strong ceiling performance: Accuracy of the strong student model trained on ground truth labels.

Then PGR is defined as: PGR = (Weak-to-strong performance - Weak performance) / (Strong ceiling performance - Weak performance)

4 Results and Discussion

Lexical Performance According to Table 1 and Table 2, bart-large can generate summarisation most aligned with groundtruth summarisation while gpt-2-xl tends to generate noisy and ineffective summarisation. Considering zero values of weak-to-strong results, we can find out that all strong models are significantly distracted by noisy patterns learned from gpt-2-xl. Among them, only bart-large-cnn struggles to generate a small proportion of meaningful summarization.

Pragmatic Performance Text generation model can generate more financially informative summarization than BART series (See Table 3).

1. LLaMA can generate more informative summarization than BART series. On both NLTK and TF-IDF scores, we can see that LLaMA outperforms compared with BART series. This suggests that although BART can generate summarization more aligned with groundtruth, it doesn't necessarily contain more financial information to reflect market behavior.

2. Compared with LLaMA, weak-to-strong generalization of BART is easier to overfit especially on noise patterns in weak generation. Weak-to-strong results becomes even worse than weak baselines consistently across three strong models, which means that weak-to-strong results are strongly distracted by noisy patterns learned by weak supervisors. It is noted that negative magnitude

Table 3: Pragmatic Performance of Weak-to-Strong on Financial News Summarization. Groundtruth summarisation generated by human editors achieves NLTK of 0.03 and TF-IDF of 0.02.

Experiment	**Weak Baseline**		**Weak-to-Strong**		**Strong Ceiling**	
[Weak → Strong]	NLTK	TF-IDF	NLTK	TF-IDF	NLTK	TF-IDF
GPT-2-xl → BART-cnn	0.05	0.03	0.04	-0.04	0.06	0.04
GPT-2-xl → BART	0.05	0.03	0.03	-0.05	0.04	0.01
GPT-2-xl → LLaMA-1-7b	0.05	0.03	-0.03	-0.03	0.39	0.09
BART-cnn → LLaMA-1-7b	0.06	0.04	0.04	0.036	0.39	0.09
BART → LLaMA-1-7b	0.04	0.01	0.00	0.01	0.39	0.09

Table 4: Lexical PGR of Weak-to-Strong on Financial News Summarization

Experiment	**Rouge1**	**Rouge2**	**RougeL**	**RougeLsum**
GPT-2-xl → BART-cnn	-0.034	0.000	-0.043	-0.036
GPT-2-xl → BART	-0.056	-0.006	-0.050	-0.053
GPT-2-xl → LLaMA-1-7b	-0.057	-0.009	-0.056	-0.034
BART-cnn → LLaMA-1-7b	-0.239	-0.410	-0.191	-0.191

of pragmatic metrics is caused by the reason that most information metrics, i.e., sentiment scores, are zeros.

3. There exists inconsistency between language metrics and pragmatic metrics, especially supported by the inverted performance of BART. Its summarization might be more aligned with human editors, who generate our groundtruth summarization, but it is doubtful whether its summarization contains enough related information to reflect market behavior. This can also suggest that human editors might also possibly fail to generate

PGR Performance According to PGR of language metrics, we can find out that weak-to-strong fails to generalize in the setting of financial news summarization in our chosen model class (see Table 4).

1. Overall, the negative magnitude of PGR across different rouge scores is a direct signal showing that weak-to-strong generalization fails to work in our task based on our model choices.

2. When fixing gpt-2-xl as the weak supervisor and testing different strong models, we can find out that summarization model (BART series) is better to achieve weak-to-strong generalization than text generation model (LLaMA-1). Specifically, bart-large-cnn performs even better than bart-large.

3. When fixing LLaMA-1-7b as the strong model and testing different weak supervisors, we can find out that gpt-2-xl seems to outperform. However, this doesn't mean that using gpt-2-xl can lead to a more meaningful weak-to-strong framework and the higher PGR is probably because the result magnitude of gpt-2 is already very low.

We didn't mention PGR of pragmatic metrics here because it's very hard to define "strong" and "weak" in the pragmatic context. Aligned with our previous discussion about inconsistency, semantically strong model like BART isn't for sure to generate meaningful summarization to reflect financial information. And this will lead to the result that sometimes strong models perform even worse than weak supervisors, causing the difficulty to interpret PGR of pragmatic metrics. However, the overall trend from Table 3 still supports the conclusion that weak-to-strong generalization fails in our chosen model class.

5 Conclusions and Future Work

Our research concludes that the weak-to-strong approach fails to generalize effectively in the context of financial news summarization with the chosen model class. The performance of the weak-to-strong method is highly dependent on the capabilities of the weak supervisors. We found that the BART

series summarization model performs better in generating summaries that align with the ground truth compared to the LLaMA-1 text generation model. However, the LLaMA-1 model excels in producing more financially informative summaries than the BART series.

In our future study, we aim to make several improvements to enhance the accuracy and effectiveness of our model. First, we plan to find a better data source with less noise to improve the quality of our dataset. For model training, we should identify a more reliable ground truth (GT), ideally without human judgment, as current use of "headlines" as GTs may not provide the best reference for summarization due to potential biases from human editors. Additionally, we intend to explore different weak and strong model pairs, such as using LLAMA-7B as the weak model and LLAMA-13B as the strong model, to better evaluate the weak-to-strong effect. Lastly, we recognize that current evaluation metrics, such as NLTK and TF-IDF, have proven inadequate for our task. Therefore, we will seek alternative metrics or methods to more accurately assess model performance.

References

[1] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

[2] Itay Goldstein. Information in financial markets and its real effects. *Review of Finance*, 27(1):1–32, 2023.

[3] Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. Predicting returns with text data. Technical report, National Bureau of Economic Research, 2019.

[4] Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. Learning to generate explainable stock predictions using self-reflective large language models. *arXiv preprint arXiv:2402.03659*, 2024.

[5] Hao Liu, Matei Zaharia, and Pieter Abbeel. Self guided exploration for automatic and diverse ai supervision. 2023.

[6] William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

[7] Karen Zhou and Chenhao Tan. Entity-based evaluation of political bias in automatic summarization. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.