electronics

MDPI

Article

Causal Inference and Prefix Prompt Engineering Based on Text Generation Models for Financial Argument Analysis

Fei Ding [1], Xin Kang [1,*], Linhuang Wang [1], Yunong Wu [2], Satoshi Nakagawa [3] and Fuji Ren [4]

[1] Faculty of Engineering, Tokushima University, Tokushima 770-8506, Japan; c502147003@tokushima-u.ac.jp (F.D.); c502147006@tokushima-u.ac.jp (L.W.)
[2] Dataa Robotics, Chengdu 610000, China
[3] Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-0033, Japan
[4] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; renfuji@uestc.edu.cn
* Correspondence: kang-xin@is.tokushima-u.ac.jp

Abstract: The field of argument analysis has become a crucial component in the advancement of natural language processing, which holds the potential to reveal unprecedented insights from complex data and enable more efficient, cost-effective solutions for enhancing human initiatives. Despite its importance, current technologies face significant challenges, including (1) low interpretability, (2) lack of precision and robustness, particularly in specialized fields like finance, and (3) the inability to deploy effectively on lightweight devices. To address these challenges, we introduce a framework uniquely designed to process and analyze massive volumes of argument data efficiently and accurately. This framework employs a text-to-text Transformer generation model as its backbone, utilizing multiple prompt engineering methods to fine-tune the model. These methods include Causal Inference from ChatGPT, which addresses the interpretability problem, and Prefix Instruction Fine-tuning as well as in-domain further pre-training, which tackle the issues of low robustness and accuracy. Ultimately, the proposed framework generates conditional outputs for specific tasks using different decoders, enabling deployment on consumer-grade devices. After conducting extensive experiments, our method achieves high accuracy, robustness, and interpretability across various tasks, including the highest F1 scores in the NTCIR-17 FinArg-1 tasks.

Keywords: generative learning; financial argument analysis; prompt engineering; causal inference

Citation: Ding, F.; Kang, X.; Wang, L.; Wu, Y.; Nakagawa, S.; Ren, F. Causal Inference and Prefix Prompt Engineering Based on Text Generation Models for Financial Argument Analysis. *Electronics* **2024**, *13*, 1746. https://doi.org/10.3390/electronics13091746

Received: 5 March 2024
Revised: 11 April 2024
Accepted: 30 April 2024
Published: 1 May 2024

1. Introduction

As deep learning (DL) technology undergoes continuous iterative updates, it has found extensive applications in various financial fields. These include financial forecasting and trading [1], risk management and fraud detection [2], and asset portfolio optimization [3], among others. One notable area of application is financial argument (FinArg) analysis [4]. This involves a critical examination and evaluation of various financial statements, data, and information presented in reports, presentations, or documents. A robust FinArg analysis forms the foundation for constructing logical and well-supported arguments, enabling individuals to make and present informed judgments regarding a company's financial health, performance trends, investment opportunities, and strategic decisions. The use of AI to automatically extract key information from large-scale financial data on the internet can profoundly impact consumer electronic behavior. However, due to specific challenges inherent to financial analysis—such as data quality and scarcity, easy model overfitting and generalization issues, and lack of result interpretability—the integration of deep learning into FinArg analysis remains fraught with difficulties [5].

After experiencing the development of rule-based (e.g., dictionary), machine learning-based (e.g., SVM [6] and LDA [7]), and deep learning-based (e.g., LSTM [8] and GRU [9]) paradigms, natural language processing (NLP) technology has become increasingly mature.

In recent years, large language models (LLMs) such as GPT-4 [10] and LLaMa-2 [11] have garnered significant attention from researchers in the NLP area. This surge in interest is primarily due to their performance in text generation tasks, which is often comparable to human-level output, and their potential applicability in various NLP tasks [12]. Concurrently, the causal capabilities of LLMs have ignited rigorous debate [13], especially given their profound implications for pivotal sectors such as finance, science, law, and policy. Yet, these models are not devoid of challenges [14]. They grapple with issues like prohibitive training costs, inconsistent performance across varied tasks, unpredictable failure scenarios, and the intricacies involved in prompt engineering and ensuring a coherent chain-of-thought reasoning.

In this paper, we introduce the Prefix Prompt Engineering Framework (PPEF) for fine-grained financial argument analysis, grounded in text generation models. Central to this framework is a text-to-text generative model that serves as its backbone. To ensure that this model generates responses in a specified format, we employ several prompt engineering strategies tailored to distinct datasets. Specifically, our approach encompasses (1) fine-tuning the model using task-specific prefix instructions; (2) additional pre-training of the backbone network on alternative financial datasets; and (3) fine-tuning the backbone network using Causal Inference produced by ChatGPT (https://openai.com/blog/chatgpt, accessed on 20 January 2024) as the explainable target. These methods can function independently or synergistically, depending on the task at hand. Section 3 delves into the intricacies of various combinations and discusses the nuanced adjustments required for handling diverse task combinations. Upon training a consistent backbone network, we employ distinct decoding methodologies tailored to various financial subtasks. These encompass both Argument Unit Identification (AUI) and Argument Relation Identification (ARI) [15]. Notably, the text-to-text generation model, when fine-tuned using prefix prompt engineering, offers several advantages over both LLMs and conventional deep learning techniques:

1. We have crafted a comprehensive set of prompts that are effective for both LLMs and text generation models. With Prefix Instruction Fine-tuning and in-domain further pre-training, our proposed framework boasts superior accuracy for specific financial tasks and demonstrates exceptional performance in the face of imbalanced datasets. It achieves the highest F1 scores in the NTCIR-17 FinArg-1 tasks [15].
2. In contrast to LLMs with hundreds of billions of parameters, our framework can be trained and perform inference on consumer-grade GPUs, even surpassing LLMs in certain tasks. Moreover, it can manage various downstream tasks using task-specific decoders without changing the backbone of the model.
3. By integrating diverse prompt engineering strategies and employing Causal Inference from ChatGPT, the text generation model not only gains proficiency in text classification and relational identification but also acquires the ability to interpret its output. To the best of our knowledge, no existing work has simultaneously achieved such high performance, robustness, and interpretability.

Finally, the proposed framework achieved good results on multiple financial argument analysis subtasks. The remainder of the paper is structured as follows: Section 2 briefly introduces related works. Section 3 details our proposed framework. Section 4 presents in-depth experiments and discussions. Finally, Section 5 concludes our work and discusses future plans.

2. Related Works

Financial information is inherently dynamic, which is one of the great challenges faced when using DL technology to process financial data. To process dynamic financial data, BloombergGPT [16] retrains an LLMs using a mixed dataset of finance and general data sources. This endeavor consumed approximately 1.3M GPU hours, translating to a staggering cost of around JPY 5M.Given the prohibitive expenses associated with retraining LLMs on a monthly or even weekly basis, there is a pronounced preference for more lightweight

adaptations within the finance sector. In response, [17] unveiled an interpretable neural network framework tailored for financial analysis. This solution adopts a hierarchical strategy, complemented by a query-driven attention mechanism, to discern sentiments in financial news texts. In a similar vein, Dogu Tan Araci [18] introduced FinBERT, a language model rooted in BERT, designed to address nuanced tasks specific to the financial landscape. Keane Ong et al. proposed FinXABSA [19], a novel approach for enhancing explainability in financial analysis. This technique employs the Pearson correlation coefficient to draw connections between Aspect-Based Sentiment Analysis and stock price fluctuations. In a similar vein, Hongyang Yang et al. presented an open-source large language model named FinGPT [20], tailored for the finance domain. Setting it apart from proprietary counterparts, FinGPT champions a data-centric ethos, offering both researchers and industry professionals a transparent and readily available resource to evolve their financial LLMs. Although these financial models manage dynamic data effectively, they lack interpretability, and their outputs fail to convince the public. Furthermore, they show less solicitude for opinion mining, and only mention a few fine-grained financial opinion-mining tasks, such as argument unit classification and Argument Relation Identification.

Argument analysis is one of the emerging research areas in NLP tasks. Some research [21,22] reviewed existing argumentation systems and applications, and discussed challenges and perspectives of this exciting new research area. Schaefer and Stede [23] focused on argumentation in social media, especially Twitter. They explored methods of modeling the structure of arguments in the tweet context corpus annotation, and reviewed current progress in the task of detecting components and their relations in tweets. Argument analysis has been widely adopted in various domains like medical [24], education [25], legal [26], and finance [27–30]. Chen et al. [27] proposed structures that link opinions with financial instruments. They explored the use of opinions from various sources to extract opinion components and identify the relationships between these opinions. Additionally, they also [28] investigated the feasibility of applying opinion mining within the financial domain. In recent years, tasks related to financial argumentation have also been addressed [15]. Many researchers have made efforts in fine-grained financial argument analysis. Lin et al. [31] used a voting strategy to determine the optimal output from several language models. Tang and Li [32] added a multi-layer convolution mechanism based on the text features extracted by BERT to improve the robustness of argument analysis. Chakraborty et al. [33] employed pre-trained language models like BERT-SEC [34] and FinBERT [18], along with a cross-encoder architecture, to handle deep semantics and relationships. Although there have been significant advances in the field of financial argument analysis, issues of robustness and interpretability remain worthy of discussion. Furthermore, few studies have combined FinArg tasks with LLMs. The performance of mainstream methods or models is summarized in Table 1.

Table 1. Performance of mainstream methods or models in financial argument analysis task. It is primarily evaluated based on three aspects: accuracy, training cost, and interpretability.

Model	Accuracy	Training Cost	Interpretability
Machine learning based models (e.g., SVM and LDA)	low	low	unavailable
BERT based models (e.g., FinBERT and RoBERTa)	medium	medium	unavailable
LLMs based models (e.g., FinGPT and BloombergGPT)	high	high	available
proposed PPEF	high	medium	available

As one of the pioneering LLMs, T5 [35] is an encoder–decoder text generation model pre-trained on a multi-task mixture of unsupervised and supervised tasks converted into a text-to-text format. It offers a unified framework for the realm of NLP pre-trained models by unifying diverse tasks into a single format. Subsequent research such as [36,37] explored

the limits of text-to-text generative models applying to Aspect-Based Sentiment Analysis (ABSA) tasks. Jordan et al. [38] introduced a Chatbot Interaction with the AI framework in which T5 plays a pivotal role in data augmentation. Similarly, a study by [39] explored a framework for fact verification. This framework harnesses pre-trained sequence-to-sequence transformer models and employs T5 in a listwise method, paired with data augmentation techniques. To the best of our knowledge, no work has been performed to date on financial argument analysis with text generation LLMs. This paper mainly focuses on providing an in-depth look at the recent trend—fine-grained financial argument analysis integrated with large language models—and focuses on the robustness and interpretability of the model.

3. Proposed Prefix Prompt Engineering Framework

In the era of the information age, we are witnessing an unprecedented explosion of text data, generated across various digital platforms. This vast expanse of data holds a wealth of information, particularly in terms of argument analysis. The task of fine-grained argument analysis often encompasses various subtasks, each with distinct input and output in terms of both content and structure. To address this, we employ the T5 [35] model as the central backbone of our framework. T5 standardizes multiple NLP tasks into a unified text-to-text format, ensuring that both the input and output are consistently represented as text strings. A comprehensive visualization of the PPEF framework is provided in Figure 1. The dataset in the figure takes the FinArg-1 task [15] as an example. A detailed explanation of each module is presented in this section.

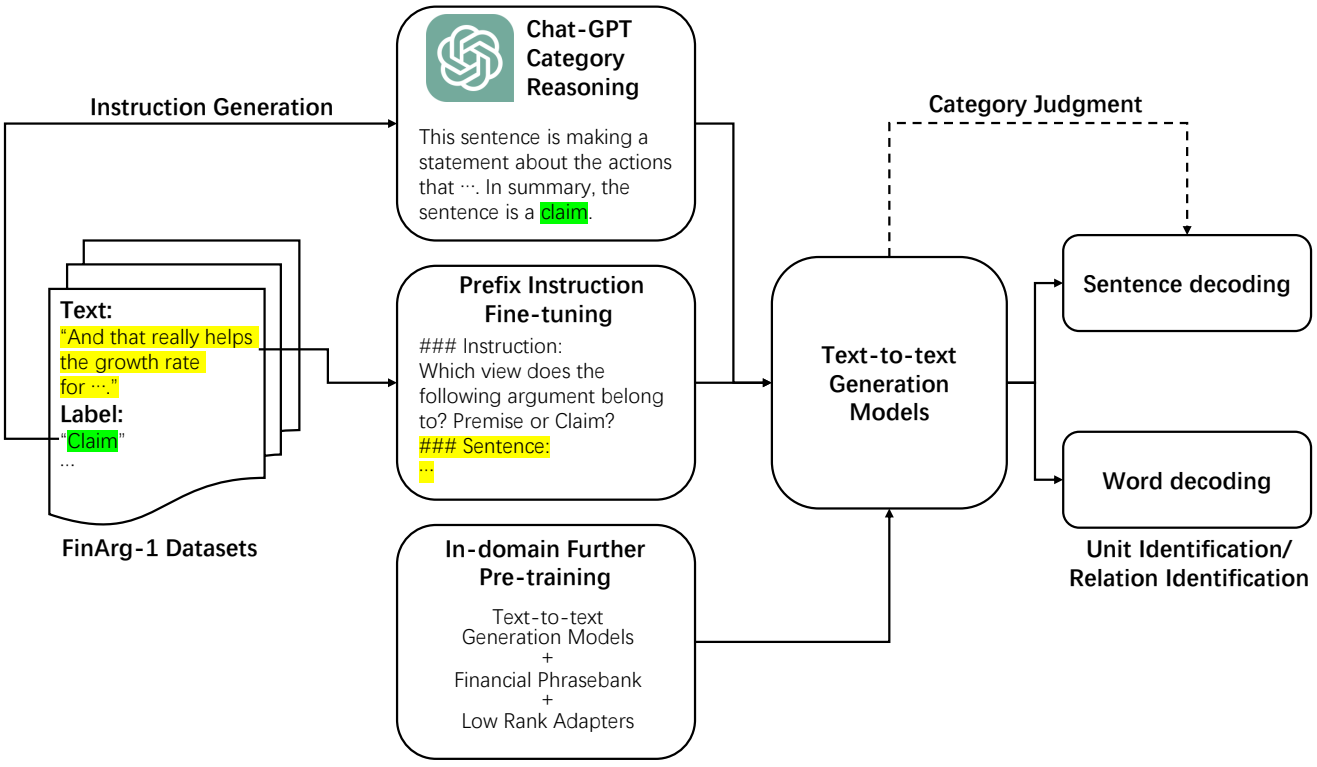

Figure 1. The proposed PPEF overview. Yellow text represents the original text of the dataset, and green text represents the labels.

3.1. Prefix Instruction Fine-Tuning

Both prompt-based learning [40] and instruction fine-tuning [41] have been demonstrated to effectively enhance the performance of various LLMs. In the paradigm of prompt-based learning, the description of the task is embedded in the input. For example, instead of giving certain parameters to the model implicitly, they are input in the form of

questions directly. To address different FinArg subtasks, we extensively tested a myriad of potential prompts and instructions. From this, we curated a generic prefix. The specifics of these prompts can be found in Table 2.

Table 2. Prompts and instructions for different subtasks. ARI stands for Argument Relation Identification and AUI stands for Argument Unit Identification. {text} and {label} present the original dataset inputs and outputs, respectively. The bold red text highlights the difference between ChatGPT Causal Inference prompt and Long Instruction for AUI (train with ChatGPT Causal Inference).

Sub Tasks	Prompt/Instruction
Short Instruction for ARI	Judge the relationship between the two sentences. Attack/Support/None: {text_1}{text_2}
Long Instruction for ARI	Below are two sentences that contain opinions. Please judge the logical relationship between sentence 1 and sentence 2. The relationship can only be Attack, Support, or no-relation. ### Sentence 1: {text_1} ### Sentence 2: {text_2}
Short Instruction for AUI	Premise or Claim:
Long Instruction for AUI	### Instruction: Which view does the following argument belong to? Premise or Claim? ### Sentence: {text} ### Argument:
Long Instruction for AUI (train with ChatGPT Causal Inference)	Below is a sentence belonging to an argumentation, contained with its component category of 'Premise' or 'Claim'. Write an explanation that appropriately explains which category the sentence belongs to and why the sentence falls into this category. Your explanation must end with 'In summary, the sentence is a premise' or 'In summary, the sentence is a claim'. ### Sentence: {text} ### Explanation:
ChatGPT Causal Inference prompt	Below is a sentence belonging to an argumentation, **Paired** with its component category of 'Premise' or 'Claim'. Write an explanation that appropriately explains which category the sentence belongs to and why the sentence falls into this category. Your explanation must end with 'In summary, the sentence is a premise' or 'In summary, the sentence is a claim'. ### Sentence: {text} **### Category:** **{Label}** ### Explanation:

For each subtask, we design one long and one short instruction, respectively. We embed the input text from the original FinArg-1 dataset into the designed prompts and instructions, with the embedding position indicated as {text} in Table 2. Additionally, for long prompts, we incorporate the special string '###' and newline characters to emphasize the structural information. Concurrently, in the long instruction for AUI (trained with ChatGPT Causal Inference), we aimed to align the instructions utilized for fine-tuning the T5 model with the prompts adopted by ChatGPT for compatibility. Subsequently, the refined dataset is employed to fine-tune the T5 model as depicted in Figure 1.

Prefix Instruction Fine-tuning significantly improves the accuracy and robustness of the model. For an in-depth examination of how varying instructions influence the outcomes across different subtasks, please refer to Section 4.

3.2. In-Domain Further Pre-Training

Pre-training language models on specific in-domain data, known as domain-adaptive pre-training, or on data relevant to particular tasks, termed task-adaptive pre-training, has been demonstrated to enhance performance in downstream tasks [42]. The nature of financial data is particularly apt for this approach. It is highly dynamic, deeply specialized, and has clear data demarcations. This makes it a prime candidate for further pre-training. Thanks to the commendable efforts of the huggingface team (https://huggingface.co/, accessed on 22 January 2024), fine-tuning and further pre-training are very convenient today.

In this paper, in addition to the original T5 model, we also fine-tune the Flan-t5 (https://huggingface.co/google/flan-t5-base, accessed on 22 January 2024) model through the financial phrasebank dataset (https://huggingface.co/datasets/financial_phrasebank, accessed on 22 January 2024) with Low-Rank Adapters (LoRA) [43]. The experimental results indicate that in some cases, further pre-training can indeed enhance the model's effectiveness, especially with highly imbalanced datasets. However, the instructions used in downstream tasks need to be adjusted accordingly. Further pre-training allows the model to acquire domain-specific knowledge, which enables it to obtain ideal output without using complex prompts and instructions when fine-tuning. But at the same time, long prompts may lead to a decrease in accuracy. For a detailed discussion and parameter settings, see Section 4.

3.3. Causal Inference from ChatGPT

Causal Inference in the context of financial arguments involves analyzing and determining the cause-and-effect relationships between financial variables or events. This is crucial for understanding how specific actions or occurrences can impact financial markets, investment outcomes, or economic indicators. Inspired by Chain-of-Thought Prompting [44] and causal reasoning [13], we let ChatGPT automatically generate material for the Causal Inference of financial data, and try to equip the T5 model with the capability to not just classify the category of AUI but also furnish explanations for its decisions. Initially, we employed a set of specific prompts to let the ChatGPT-3.5 model (model number 'text-davinci-003') generate an "Explanation" that rationalizes the classification determinations made. To ensure that the "Explanation" yielded by ChatGPT aligns with the label of the original dataset, we embedded a strong correlation constraint within the prompt (the red bold part in Table 2). Under the restriction of strong constraints, all "Explanation" generated by ChatGPT conform to the ground truth label of the data. The generated text always ends up with "In summary, the sentence is a premise." or "In summary, the sentence is a premise." Table 3 contains some examples of explanations generated by ChatGPT.

Table 3. "Explanation" examples generated by ChatGPT.

Label	ChatGPT "Explanation"
Premise	"This sentence provides evidence to support a larger argument by providing an example of how success can be achieved. It does not make a statement that can be proven or disproven, but rather provides an example of how success can be achieved. In summary, the sentence is a redpremise."
Claim	"This sentence is making a statement about the actions that Amazon is taking, which is a claim. It is not providing any evidence or reasoning to support a conclusion, which is what a premise would do. In summary, the sentence is a claim."

Subsequently, the "Explanation" generated by ChatGPT is harnessed as the target text for the fine-tuning process of the T5 model as shown in Figure 1. As mentioned in Section 3.1, when the reasoning of ChatGPT is used as the label, we aim to align the instruction during the T5 fine-tuning with the prompt that was initially utilized by ChatGPT to gain the same knowledge and logical reasoning. Specifically, the prompts are shown in "Long Instruction for AUI (train with ChatGPT Causal Inference)" and "ChatGPT Causal Inference prompt" in Table 2. In this process, ChatGPT is regarded as a "teacher" model, guiding the T5 model to output in the form of causal reasoning, making the model's decisions somewhat interpretable. An example of the difference between using Causal Inference from ChatGPT or not is shown in Figure 2. Causal Inference from ChatGPT enables the T5 model to simultaneously output text labels and inferences. When evaluating model performance, we use the last word of the T5 model inference text as the output label.

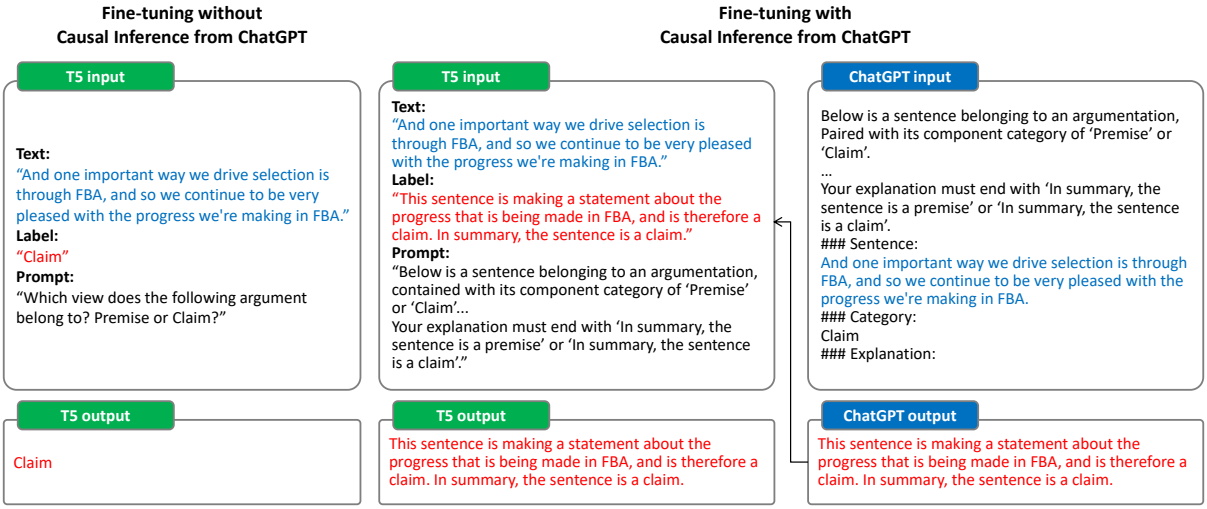

Figure 2. Causal Inference from ChatGPT enables the T5 model to simultaneously output text labels and inferences. The original FinArg-1 text and the labels used by the different methods are highlighted.

3.4. Task-Specific Decoding

Given that the T5 model always utilizes text strings for both input and output, we transform the labels from the FinArg-1 task dataset into corresponding textual representations. Notably, when employing the "Explanation" generated by ChatGPT as the ground truth label, we consider the penultimate token of the output string as the model's label output word. This approach is adopted primarily because, in most instances, the last token of the string tends to be a period.

In the AUI subtask, we use the mapping $f_u : 0 \rightarrow$ 'premise', $1 \rightarrow$ 'claim' to map the corresponding labels. It is worth noting that words with a capitalized first letter cannot be

used because in the tokenizer of T5, words with capital letters are sometimes split into two tokens. In the ARI subtask, we use the mapping $f_r : 0 \rightarrow$ 'none', $1 \rightarrow$ 'support', $2 \rightarrow$ 'attack' to map the corresponding labels. When using "no relation" or "unrelated" as the mapping word, the tokenizer will also split them into two tokens. Finally, we choose "none" as the mapping word for label "0". Experimental results show that after at least one epoch of training, the output of the T5 model is always included in the mapping vocabulary. Therefore, the final output label can be obtained without any other decoding.

When using the "Explanation" generated by ChatGPT as the ground truth label, due to the existence of instruction, almost all the text sequences output by the T5 model will end with words in the mapping vocabulary, just like the examples in Table 3. In some extreme cases, when the output of the T5 model exceeds the set max length parameter, the output will be truncated. This situation can usually be solved by limiting the output length of ChatGPT or increasing the max length of the T5 model. But for the sake of saving computing resources, we usually prefer a smaller max length. Therefore, when the output string does not end with the words in the mapping vocabulary, there are two ways to deal with it: (1) simply considering the output to be wrong, or (2) re-inputting the output sentence into the fine-tuned model and judging what the category label contained in the sentence is, which we call Category Judgment. The results of the experiments are elaborated in Section 4.

Through the use of INT8 quantization techniques and the method of Category Judgment, our framework can be deployed on consumer-grade graphics cards. Specifically, the entire framework can be trained using only 12 GB of GPU memory and can perform inference with a minimum requirement of 6 GB of GPU memory.

4. Experiments and Discussions

4.1. Experiment Setup

4.1.1. Dataset

In the current study, our proposed method is evaluated on two datasets with two subtasks for fine-grained argument understanding in financial analysis. The aim of these tasks is to comprehend the arguments present in investor-generated text, which includes both professional and amateur textual data. Table 4 show the statistics of the datasets. The datasets details are as follows:

- **Argument Unit Identification** This subtask requires models to distinguish whether a given argumentation sentence is a claim or a premise. The dataset contains a total of 9691 sentences, of which 5078 are premises and 4613 are claims. Data examples are as follows:
  ["First of all, I want to remind you that Q3 is typically a lower operating income quarter as we're preparing for the Q4 holiday peak.", Claim],
  ["On the international, on an FX neutral basis, the growth was 15% in Q3 and 19% in Q4.", Premise];
- **Argument Relation Identification** This subtask necessitates the identification of the relationship between arguments, specifically discerning whether it is one of the support, attack, or other categories. The textual portion of the dataset comprises two separate sentences. It is worth mentioning that this dataset is significantly imbalanced, with the "attack" label accounting for only 1.1% of the total labels. Data examples are as follows:
  ["Some have a 24-month clock, and there are even some that have a 30-month clock.", "They come back in and they pay less for the service but they pay more for their smartphone.", There is no detected relation between the two sentences.],
  ["Japan as a geography for us is a high transactional market.", "The improvement in that in Q3 is obviously very high margin and also the bottom.", There is a "Support" relation from sentence 1 to sentence 2.],

["And that in fact in Q1 caused the market to expand.", "So, at least in the intermediate timeframe, we do not see cannibalization.", There is an "Attack" relation from sentence 1 to sentence 2.].

Table 4. Data statistics of Argument Unit Identification and Argument Relation Identification.

Argument Unit Identification				
	Train	Dev	Test	Whole
Preminse	4062	508	508	5078 (52.4%)
Claim	3691	461	461	4613 (47.6%)
Total	7753	969	969	9691
Argument Relation Identification				
	Train	Dev	Test	Whole
Support	3859	482	482	4823 (69.9%)
Attack	62	8	8	78 (1.1%)
Other	1600	200	200	2000 (29.0%)
Total	5521	690	690	6901

4.1.2. Computer Configuration

All experiments were run on the following servers. OS: CentOS Linux release 7.6.1810. Linux Core: 3.10.0-957.el7.x86 64. CPU: Intel Core i7 6700k. GPU: NVIDIA GeForce RTX 3090Ti 24 GB. RAM: TeamGroup 32 GB. Python version: 3.7.3.

4.1.3. Evaluation Metrics

We report both Micro-F1 and Macro-F1 scores of all tasks. Due to the presence of imbalanced datasets, we use Macro-F1 to rank the results. The formula for the F1 score is:

$$\text{F1} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}. \quad (1)$$

Specifically, Micro-F1 calculates metrics globally by counting the total true positives, false negatives, and false positives. Macro-F1 calculates metrics for each label, and finds their unweighted mean.

4.2. Experiment Results

To prove the effectiveness of our PPEF framework in fine-grained argument understanding in financial analysis, we compare our proposed model with some strong baselines. We also compare the results from the state-of-the-art LLMs, such as GPT-4. The details of the baselines are as follows:

- **BERT:** ref. [45] is an NLP model developed by Google's AI that stands for Bidirectional Encoder Representations from Transformers. We use "bert-base-uncased" (https://huggingface.co/bert-base-uncased, accessed on 5 September 2023) as the baseline of pre-traind BERT in this experiment. The hidden representation of the [CLS] token is extracted, and a single-layer MLP is added for label classification.
- **RoBERTa:** ref. [46] is an NLP model builds upon the BERT architecture, utilizing dynamic masking and larger batch size. We use "xlm-roberta-base" (https://huggingface.co/xlm-roberta-base, accessed on 5 September 2023) as the baseline of pre-traind RoBERTa in this experiment. We add the same MLP layer as BERT.
- **FinBert:** ref. [18] is a pre-trained NLP model designed to analyze the sentiment of financial texts. It is developed by further training the BERT language model in the finance domain, using a large corpus of financial documents.
- **T5:** ref. [35] is an encoder–decoder model that has been pre-trained on a multi-task mixture of unsupervised and supervised tasks, with each task converted into a text-

to-text format. We use "t5-large" (https://huggingface.co/t5-large, accessed on 8 September 2023) with 770 million parameters as baseline. The text and labels of the dataset are directly used as input and output of T5, without any prompts.

- **ChatGLM:** ref. [47] is an open bilingual language model based on General Language Model (GLM) framework. We choose ChatGLM-6B (https://github.com/THUDM/ChatGLM-6B, accessed on 11 September 2023), with 6.2 billion parameters, as the baseline.
- **GPT-4:** ref. [10] is a large multimodal model accepting image and text inputs and emitting text outputs, which was recently released by OpenAI. Since the parameters of GPT-4 is not public, we use the few-shot method to complete the experiment.
- **TMUNLP:** ref. [31] uses a voting strategy to determine the optimal output from several language models.
- **IDEA:** ref. [32] adds a multi-layer convolution mechanism based on the text features extracted by BERT to improve the robustness of argument analysis.
- **LIPI:** ref. [33] employs multiple pre-trained language models and a cross-encoder architecture to handle deep semantics and relationships.

For comparative experiments, we also use "t5-large" as the backbone of the proposed framework. BERT-like models use a learning rate of 5×10^{-5} while other models use a learning rate of 3×10^{-4}. All experiments are trained for 10 epochs, and the final result is the average among five runs with different random seeds.

In the AUI subtask, we used all framework modules mentioned in Sections 3.2–3.4 simultaneously. However, in the ARI subtask, Causal Inference from ChatGPT is not used. That is because, even with strong constraints added, when inferring the argument relationship, a considerable part of the reasoning strings generated by the ChatGPT and T5 models still exceed the mapping vocabulary. The experimental results are shown in Table 5. Our proposed framework exceeds the comparative baselines on both tasks. Since the GPT-4 model cannot be fine-tuned, it does not perform well on specific fine-grained financial analysis tasks, even worse than BERT-like models. Our proposed framework also ranks first in the ARI subtask and third in the AUI subtask of NTCIR-17 FinArg-1, respectively. For the task details, refer to [15].

Table 5. Experiment results for AUI and ARI subtasks. 'CIC' presents Causal Inference from ChatGPT and 'FFP' presents Financial Further Pre-training. Results are reported as average and standard deviation among 5 runs. Bold fonts represent the best results. The results of TMUNLP, IDEA and LIPI are the best results reported in FinArg-1 tasks.

	Argument Unit Iden.		**Argument Relation Iden.**	
	Micro-F1	**Macro-F1**	**Micro-F1**	**Macro-F1**
BERT	75.14 ± 0.12	75.27 ± 0.03	82.67 ± 0.07	52.66 ± 0.12
RoBERTa	74.69 ± 0.15	74.7 ± 0.13	81.74 ± 0.11	55.52 ± 0.09
FinBert	75.67 ± 0.08	75.36 ± 0.10	82.53 ± 0.10	51.81 ± 0.03
T5	73.75 ± 0.09	73.66 ± 0.10	82.01 ± 0.11	53.74 ± 0.18
ChatGLM	76.17 ± 0.19	75.97 ± 0.31	79.58 ± 0.03	60.11 ± 0.04
GPT-4 (few-shot)	62.41 ± 0.06	62.39 ± 0.06	69.82 ± 0.04	48.72 ± 0.17
TMUNLP	76.57	76.55	82.07	57.90
IDEA	76.47	76.46	81.74	51.85
LIPI	73.89	73.86	79.42	60.22
Ours PPEF	**77.23 ± 0.10**	**77.27 ± 0.03**	85.61 ± 0.03	**61.44 ± 0.04**
PPEF w/o CIC&FFP	74.49 ± 0.07	74.51 ± 0.03	-	-
PPEF w/o FFP	76.36 ± 0.07	76.32 ± 0.06	**85.83 ± 0.07**	55.17 ± 0.12
PPEF w/o CIC	76.31 ± 0.04	76.19 ± 0.10	-	-

4.3. Discussions

4.3.1. Ablation Experiments

This section provides the ablation study for the proposed framework as shown in the lower part of Table 5. In the AUI task, if only Prefix Instruction Fine-tuning (PIF) is used instead of Financial Further Pre-training (FFP) and Causal Inference from ChatGPT (CIC), the obtained Macro-F1 score is 74.56, which is only slightly higher than the original T5 model, not as good as other SOTA LLMs model. If FFP or CIC is not used, the effect of the model will slightly decrease.

In the ARI task, we observe analogous results. Notably, in the absence of FFP, the model's Micro-F1 score for the ARI task exhibits an increase. However, this is accompanied by a significant decline in the Macro-F1 score. This underscores that the FFP module notably enhances the model's stability, especially when handling imbalanced datasets. For all ablation studies, we employ a substantial max length to mitigate the potential influence of the mapping vocabulary issue. A more in-depth discussion on max length can be found in Section 4.3.3.

4.3.2. Long or Short Instructions

When not using CIC, for each subtask, we use one long instruction and one short instruction for each subtask respectively as mentioned in Table 2. At this point, we observe that whether or not FFP is used will have a significant impact on the inference results as shown in Figure 3. In general, shorter prompts give better results if FFP is used; on the other hand, longer prompts are better if FFP is not used. This could be attributed to the model acquiring relevant background knowledge during further pre-training. Without this knowledge, a more specific instruction-guided approach might be necessary during fine-tuning to produce the appropriate response.

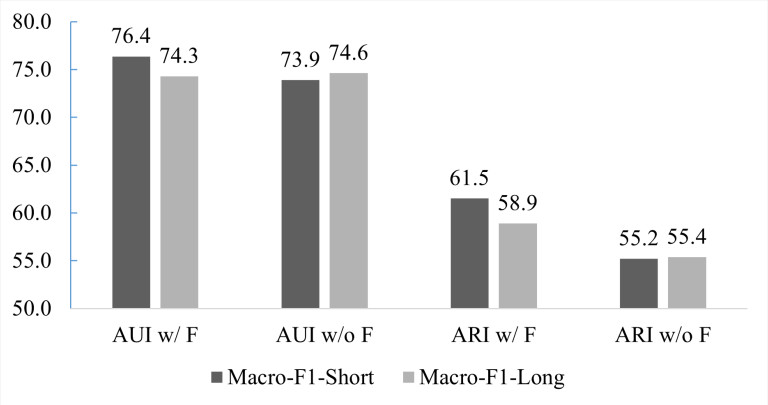

Figure 3. Long/short instructions act on inference results.

4.3.3. Case Study and Category Judgment

In the actual inference process, when the output of the T5 model exceeds the max length parameter, it will be truncated. This situation can usually be avoided by setting the max length to be greater than 256. But reducing the max length of the T5 model can exponentially speed up the inference; therefore, we always prefer a smaller max length. In order to solve this problem, we propose the Category Judgment method, which is to re-enter the outputs outside the mapping vocabulary into the fine-tuned model for further reasoning.

Some examples of case study and Category Judgment are listed in Table 6. Overall, the shorter the input sentence and the fewer compound arguments it contains, the more accurate the output of the model will be. The wrong predictions are concentrated in the case when the input sentence contains some complex ideas. When the max length is set to be 40, with the Category Judgment method, most of the output exceeding the max length can be accurately identified. We also observe that some incorrectly predicted sentences

lead to correct results after Re-Output as shown in the third example in Table 6. Based on these results, we find that after Category Judgment, the prediction accuracy of the model can reach a comparable level to that of using a large max length, thus greatly saving the computing resources of inference.

Table 6. Case study and Category Judgment examples. Words in curly brackets indicate truncated parts when the max length was set to be 40. Re-Output is the result of re-entering the output into the model with a larger max length. Green ticks represent correct predictions and red crosses represent incorrect predictions.

Predictions	Sentences
Output: Premise✓ **Re-Output:** Premise✓	**Input:** "And in that context, they lifting and shifting some of the older workloads, but they're modernizing the entire business process flow." **Output:** "This sentence provides evidence to support the argument by providing an example of how the argument is being implemented. It explains the action being taken to achieve the goal, which is to lift and shift older workloads, {and to modernize the entire business process flow. In summary, the sentence is a premise.}" **Re-Output:** "This sentence provides evidence to support the argument by providing an example of how the argument is being implemented. It does not make a statement that is meant to be accepted as true, but rather provides evidence to support the argument. In summary, the sentence is a premise."
Output: Claim✓ **Re-Output:** Claim✓	**Input:** "In this particular case, back to commenting on this one, we're hopeful that over time the restrictions we're seeing are loosened. Because innovation really requires freedom to collaborate and communicate, and I know that is a major focus there." **Output:** "This sentence is making a statement about the potential outcome of the restrictions that are being seen. It is expressing a hope that the restrictions will be loosened, and that this is a major focus of the argument{. In summary, the sentence is a claim.}" **Re-Output:** "This sentence is making a statement about the potential outcome of the restrictions being seen, and is expressing a hope that the restrictions will be loosened. This is a statement of opinion, and is therefore a claim. In summary, the sentence is a claim."
Output: Claim× **Re-Output:** Premise✓	**Input:** "I also really like what I've seen with the iPhone SE and the fact that it's opening the door to customers. We weren't reaching before and likely convincing some people to upgrade that wanted a smaller form factor but wanted to stay with iPhone, and so they were waiting for the iPhone SE." **Output:** "This sentence is making a claim about the success of the iPhone SE and the potential for it to convince customers to upgrade to a smaller form factor. It is not providing any evidence or facts to support this {claim, but is instead making a statement about the potential of the iPhone SE. In summary, the sentence is a claim.}" **Re-Output:** "This sentence provides evidence to support the argument by providing an example of a success of the iPhone SE and the potential for it to convince customers to upgrade to a smaller form factor. It does not make a statement about the argument itself, but rather provides evidence to support the argument. In summary, the sentence is a premise."
Claim✓	**Input:** "And one important way we drive selection is through FBA, and so we continue to be very pleased with the progress we're making in FBA." **Output:** "This sentence is making a statement about the progress that is being made inFBA, and is therefore a claim. In summary, the sentence is a claim."

5. Conclusions

In this paper, we propose a novel financial argument analysis framework PPEF. Through the combination of different modules, PPEF can handle various subtasks of fine-grained argument analysis. Through the PIF and FFP methods, PPEF obviously improves the accuracy and robustness of the AUI and ARI tasks. Among models at the same parameter level, PPEF achieves a state-of-the-art F1 score. Moreover, with CIC approach, the framework can conduct explainable causal reasoning on consumer-grade graphics cards like LLMs such as GPT-4.

In the future, we plan to investigate the effectiveness of our proposed method for other financial analysis tasks, such as aspect-based financial classification and financial text generation. Another direction worth exploring is whether more precise instructions can enable the application of Causal Inference from the "teacher" model to more complex tasks.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization: F.D., X.K. and F.R.; Formal analysis: F.D., Y.W. and S.N.; Methodology: F.D. and L.W.; Resources: Y.W., S.N. and F.R.; Writing (original draft): F.D.; Writing (review and editing): X.K., L.W. and F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Project of Discretionary Budget of the Dean, Graduate School of Technology, Industrial and Social Sciences, Tokushima University.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, Xin Kang, upon reasonable request.

Acknowledgments: Authors acknowledge the support of their respective institutes.

Conflicts of Interest: The authors whose names are listed immediately below report the following conflicts of interest related to the work under consideration: Yunong Wu is employed by Dataa Robotics. The authors whose names are listed immediately below certify that they have NO conflicts of interest to declare: Fei Ding, Xin Kang, Linhuang Wang, Satoshi Nakagawa, Fuji Ren.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep Learning
FinArg	Financial Argument
LLMs	Large Language Models
NLP	Natural Language Processing
PPEF	Prefix Prompt Engineering Framework
AUI	Argument Unit Identification
ARI	Argument Relation Identification
ABSA	Aspect-Based Sentiment Analysis
CIC	Causal Inference from ChatGPT
FFP	Financial Further Pre-training
PIF	Prefix Instruction Fine-tuning
GLM	General Language Model

References

1. Barra, S.; Carta, S.M.; Corriga, A.; Podda, A.S.; Recupero, D.R. Deep learning and time series-to-image encoding for financial forecasting. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 683–692. [CrossRef]
2. Leo, M.; Sharma, S.; Maddulety, K. Machine learning in banking risk management: A literature review. *Risks* **2019**, *7*, 29. [CrossRef]
3. Soleymani, F.; Paquet, E. Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder—DeepBreath. *Expert Syst. Appl.* **2020**, *156*, 113456. [CrossRef]
4. Van Eemeren, F.H.; Henkemans, A.F.S.; Grootendorst, R. *Argumentation: Analysis, Evaluation, Presentation*; Routledge: New York, NY, USA, 2002.
5. Huang, J.; Chai, J.; Cho, S. Deep learning in finance and banking: A literature review and classification. *Front. Bus. Res. China* **2020**, *14*, 1–24. [CrossRef]
6. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

7. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
8. Graves, A.; Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
9. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
10. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
11. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
12. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
13. Kıcıman, E.; Ness, R.; Sharma, A.; Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv* **2023**, arXiv:2305.00050.
14. Shen, Y.; Heacock, L.; Elias, J.; Hentel, K.D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and other large language models are double-edged swords. *Radiology* **2023**, *307*, e230163. [CrossRef] [PubMed]
15. Chen, C.C.; Lin, C.Y.; Chiu, C.J.; Huang, H.H.; Alhamzeh, A.; Huang, Y.L.; Takamura, H.; Chen, H.H. Overview of the NTCIR-17 FinArg-1 Task: Fine-Grained Argument Understanding in Financial Analysis. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan, 12–15 December 2023.
16. Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; Mann, G. BloombergGPT: A Large Language Model for Finance. *arXiv* **2023**, arXiv:2303.17564.
17. Luo, L.; Ao, X.; Pan, F.; Wang, J.; Zhao, T.; Yu, N.; He, Q. Beyond Polarity: Interpretable Financial Sentiment Analysis with Hierarchical Query-driven Attention. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; pp. 4244–4250.
18. Araci, D. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv* **2019**, arXiv:1908.10063.
19. Ong, K.; van der Heever, W.; Satapathy, R.; Mengaldo, G.; Cambria, E. FinXABSA: Explainable Finance through Aspect-Based Sentiment Analysis. *arXiv* **2023**, arXiv:2303.02563.
20. Yang, H.; Liu, X.Y.; Wang, C.D. FinGPT: Open-Source Financial Large Language Models. *arXiv* **2023**, arXiv:2306.06031.
21. Lippi, M.; Torroni, P. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol. (TOIT)* **2016**, *16*, 1–25.
22. Lawrence, J.; Reed, C. Argument mining: A survey. *Comput. Linguist.* **2020**, *45*, 765–818. [CrossRef]
23. Schaefer, R.; Stede, M. Argument mining on Twitter: A survey. *IT-Inf. Technol.* **2021**, *63*, 45–58. [CrossRef]
24. Dou, R.; Kang, X. TAM-SenticNet: A Neuro-Symbolic AI approach for early depression detection via social media analysis. *Comput. Electr. Eng.* **2024**, *114*, 109071. [CrossRef]
25. Martins, M. Analysis of High School Students' Argumentative Dialogues in Different Modelling Situations. *Sci. Educ.* **2024**, *33*, 175–212. [CrossRef]
26. Xu, H.; Ashley, K. Multi-granularity Argument Mining in Legal Texts. *arXiv* **2022**, arXiv:2210.09472.
27. Chen, C.C.; Huang, H.H.; Chen, H.H. *From Opinion Mining to Financial Argument Mining*; Springer Nature: Berlin/Heidelberg, Germany, 2021.
28. Chen, C.C.; Huang, H.H.; Chen, H.H. A research agenda for financial opinion mining. In Proceedings of the International AAAI Conference on Web and Social Media, virtually, 7–10 June 2021; Volume 15, pp. 1059–1063.
29. Ma, X.; Zheng, F.; Tang, D. Identifying the Head-and-Shoulders Pattern Using Financial Key Points and Its Application in Consumer Electronic Stocks. *IEEE Trans. Consum. Electron.* 2023, *in press*. [CrossRef]
30. Roy, R.; Ghosh, S.; Naskar, S.K. Financial Argument Analysis in Bengali. In Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, Panjim, India, 15–18 December 2023, pp. 88–92.
31. Lin, H.Y.; Sy, E.; Peng, T.C.; Huang, S.H.; Chang, Y.C. TMUNLP at the NTCIR-17 FinArg-1 Task. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan, 12–15 December 2023. [CrossRef]
32. Tang, S.; Li, L. IDEA at the NTCIR-17 FinArg-1 Task: Argument-based Sentiment Analysis. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan, 12–15 December 2023. [CrossRef]
33. Chakraborty, S.; Sarkar, A.; Suman, D.; Ghosh, S.; Naskar, S.K. LIPI at the NTCIR-17 FinArg-1 Task: Using Pre-trained Language Models for Comprehending Financial Arguments. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan, 12–15 December 2023. [CrossRef]
34. Loukas, L.; Fergadiotis, M.; Chalkidis, I.; Spyropoulou, E.; Malakasiotis, P.; Androutsopoulos, I.; Paliouras, G. FiNER: Financial numeric entity recognition for XBRL tagging. *arXiv* **2022**, arXiv:2203.06482.
35. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
36. Zhang, W.; Deng, Y.; Li, X.; Yuan, Y.; Bing, L.; Lam, W. Aspect sentiment quad prediction as paraphrase generation. *arXiv* **2021**, arXiv:2110.00796.
37. Gao, T.; Fang, J.; Liu, H.; Liu, Z.; Liu, C.; Liu, P.; Bao, Y.; Yan, W. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 7002–7012.

38. Bird, J.J.; Ekárt, A.; Faria, D.R. Chatbot Interaction with Artificial Intelligence: Human data augmentation with T5 and language transformer ensemble for text classification. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 3129–3144. [CrossRef]
39. Jiang, K.; Pradeep, R.; Lin, J. Exploring listwise evidence reasoning with t5 for fact verification. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Online, 1–6 August 2021, pp. 402–410.
40. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
41. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Stanford Alpaca: An Instruction-following LLaMA Model. 2023. Available online: https://github.com/tatsu-lab/stanford_alpaca (accessed on 13 March 2023).
42. Zhu, Q.; Gu, Y.; Luo, L.; Li, B.; Li, C.; Peng, W.; Huang, M.; Zhu, X. When does further pre-training MLM help? An empirical study on task-oriented dialog pre-training. In Proceedings of the Second Workshop on Insights from Negative Results in NLP, Online, 10 November 2021; pp. 54–61.
43. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
44. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
45. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, p. 2.
46. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
47. Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; Tang, J. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 320–335.