Commodity Market Forecasting using NLP

Dana Golden, Dikshya Mohanty, Khushboo Singh

I. Introduction

Commodities markets depend heavily on movements in physical supply and demand. The government releases reports on commodities to provide the market with information on these movements [1]–[3]. We have developed a model that generated commodity price forecasts, along with the economic explanations of these futures price movements, based on the text of these government reports. Futures prices are the price of a commodity delivered at a specified date in the future. By building the model, we improved our ability to understand the movement in these prices and causality reasons and bring machine learning and economics closer together.

Economists traditionally spend considerable time manually reviewing government reports to decipher the actual reasons behind commodity price movements. Additionally, the absence of domain-specific language models hampers the automated analysis of these reports. This research addresses these challenges by combining time-series forecasting with text-based analysis. This methodology not only enhances our understanding of price forecasts but also clarifies the underlying causal factors, thus paving the way for the development of reliable and interpretable machine learning models.

We experiment with various model architectures that integrate different combinations of forecasting and language models, aiming to effectively merge temporal and textual features for enhanced analytical capabilities. The forecasting models yield futures price predictions, while the language models elucidate the causality behind these movements. This project intersects the domains of social science and computer science, employing a fusion of economic analysis and natural language processing techniques to explore the underlying factors driving commodity price movements.

This task presents considerable complexity as the language model must adapt to the specific domain to enhance its reasoning capabilities. For instance, it is crucial for the model to effectively discern and economically interpret scenarios where prices increase despite rises in supply or declines in demand, by identifying how these factors uniquely impact the market.

II. Background

The original research idea comes from work done by [4] to explain the economic reasons for futures price movements based on news data. The work also only classified movements in prices into discrete bins instead of producing explanations that could be broader. We build on this work by incorporating expert analysis from government officials rather than news. While explaining movements is a recent task in economic prediction, it requires domain expertise. How to forecast commodities markets is one of the oldest questions in economics. The original work has long lagged behind standard economic literature technically and utilized esoteric classical econometric time-series techniques. A recent branch of the literature has attempted to add machine learning techniques [5]–[9], but these techniques tend to be poorly explained and lack validity in the eyes of economists. Adding an explainable component to the forecasts could make these forecasting techniques more reliable within mainstream economics. The question of how government reports impact commodity markets and contain unique information has been of perennial interest to economics going back to the 1970s with most papers finding information content but some finding declining new information as markets become more efficient [10]–[20]. A budding literature also exists combining NLP-analysis with economic analysis, but this literature has mostly been led by economists and thus lacks good implementations of recent techniques in deep learning [21]–[23]. Our research combines these separate strands of literature to utilize the information in government reports to explain forecasts made by state-of-the-art transformer-based systems that combine text and time-series data.

Our study focuses on fusing multi-modal features into single- and multi-objective models [24]–[27] by integrating time series and textual data to simultaneously generate numerical predictions and their corresponding rationales. While there has been significant research on exploring multi-modality with language and other modals such as sight, sound, print, images, video, music [28], an opportunity exists when it comes to using language with Time-Series data. This integration between the two modals allows for a deeper understanding of the data, thereby generating accurate, more robust long-term forecasts, and complementary causality explanations for the movements in the forecasts - that enhance the explainability, reliability and usability of the forecasts. To do this, we explore state-of-the-art models for each modal individually. The vanilla transformer architecture [29] for time-series forecasting was implemented by [30] wherein they propose a general-purpose foundation model for uni-variate probabilistic time series forecasting. Using this as motivation, we create a custom n-layered encoder-decoder structure to fit our commodities' time-series data. For causality reasoning, we draw inspiration from success in Reasoning with fine-tuned Large Language Models (LLMS) using Instruction

Tuning [31]. For this task, we choose language models of varying sizes - GPT-2-Medium [32] (355M parameter) and LLAMA-2-7b [33] (7B parameters). For the GPT-2 model, we train on all parameters, while for LLAMA-2 model, we leverage a technique called parameter-efficient fine-tuning, proposed by [34] - that allows us to limit fine-tuning parameters and memory usage while achieving comparable performance to full-parameters fine-tuning. We specifically chose the PEFT with LLAMA model because it adapts well to multiple domains [35]. For the final model architectures, we fuse cross-modal features by using Multi-Objective Learning [36] and Cascaded Model Learning [37].

III. Data

The development of the dataset represents a significant contribution to this project. Data was sourced from the Energy Information Administration (EIA), which regularly publishes reports and datasets on the dynamics of the oil and natural gas markets, including supply and demand. Given the absence of any pre-existing dataset that amalgamates the historical records of these reports, we were compelled to construct our own dataset. This involved a combination of scraping from the EIA website and developing scripts to bulk download PDF copies of the reports. [I. Syntax] To prepare the data for analysis, we conducted basic data preprocessing using regular expressions. This involved removing HTML tags and comments, non-meaningful numbers, special characters as well as CSS code from the text, to ensure the cleanliness and usability of the data.

We examined the histories of three distinct text-based reports: the weekly natural gas report, the weekly petroleum report, and the monthly short-term energy outlook. Each report contains narrative text that explains past price and market movements and the reasons behind these changes. Notably, the short-term energy outlook encompasses both oil and natural gas, featuring extensive forecasting discussions. Additionally, we downloaded futures prices from the EIA website and linked these prices with the corresponding reports to construct a comprehensive time-series dataset.

To obtain ground-truth labels for training our model on the reasoning behind price movements, we utilized the ChatGPT API. We input documents from the Energy Information Administration (EIA) that detailed the economic reasons for fluctuations in oil and natural gas prices into ChatGPT. We then prompted ChatGPT to emulate a financial economist and generate explanations for the previous period's price movements based on the current report. ChatGPT provided three potential reasons for each report. These reasons were manually reviewed by an economist with expertise in commodity futures forecasting, and following this verification, the data was adopted as ground truth labels for our model training.

Prior to model training, we partitioned the text and numerical features into train and test sets using an 80:20 ratio. The training data was used for model training, and then models were evaluated on test set. All the metrics present in this report (after Data section) are evaluated on this out-of-sample, test set. Each data point in the train set was converted to the instruction format, in a way that "instruction", "input" and "target(reason)" appeared as the input to the model in the training phase. For the test set, "target(reason)" was excluded from model inputs, and the output was expected to be the target reason.

Table I provides descriptive statistics about the created dataset.

TABLE I: Data Sources

Source of Data	**Features**			
	Type of Data	***Start***	***Docs***	***Size***
NG Reports	HTML Text	2001	1136	1.5M words
Oil Reports	Text	2011	533	580K words
STEO Reports	Text	1997	325	1M words
Futures Prices	Numerical	1983	NA	17,891 prices

IV. Methods

A. Baselines

While the task at hand focuses on explainable forecasting combining time-series and text data, our baseline models evaluate the two modals separately. For the time-series forecasting, we start with 3 simple models: a) average of history, b) previous time-step value, and c) ARIMA with 1-integrated, 5-Auto regressive & 0 Moving Average features. Furthermore, we have added 2 more baselines to predict next week's future prices based on historical prices - d) adapting GPT-2 model architecture for a regression task [38], & e) an Autoformer architecture (decomposition architecture with autocorrelation) [39].

We implement zero-shot learning using Llama2, by prompting the model to generate a causality reason for the movement between prices on the basis of EIA reports. The prompt included an instruction for the task, historical & future commodity price and the future week's report as inputs.

B. Improvements

We follow a similar approach as we did for the Baseline Models - we evaluate each modal separately. For forecasting prices, [II. Semantics] we implement two improved models - a) Time Series Transformer model, & b)Supervised Fine-Tuned Model using a RoBERTa-base model with a regression objective. Both the models undergo hyper-parameter tuning to fit our data. We further implemented hyper-parameter tuning with our baseline models - GPT2 and Autoformer, but dropped them due to poor performance in comparison to our improved models.

For predicting monthly future prices for both natural gas and oil together from the same report, the supervised

fine-tuning trainer (SFT) based on RoBERTa was modified to perform dual regression and was improved further by hyper-parameter tuning, with learning rate of 0.00001 and 5 epochs.

For causality reasons, [III.Language Modeling] we implemented a) Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation(LoRA) [40] on Llama-2-7b model, & b) Full Fine-Tuning on GPT-2-M model, for predicting the reason for the future price movements. For LLAMA-2, we experimented with different batch sizes, steps, r, alpha, dropout & target modules, to find the optimum values - limiting trainable parameters to roughly 12.4% of the total number of parameters (8M). For GPT-2 model, we implemented a full model fine-tuning approach - updating all the trainable parameters (355M). We further experimented with the number of epochs, batch size, learning rate & weight decay to increase the performance. A summary of improved models alongside hyper-parameters is in Figure 1.

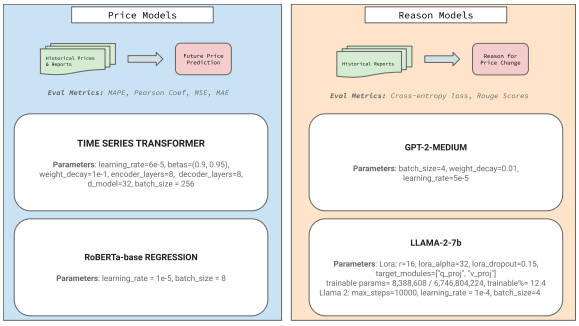

Fig. 1: Diagram of Improved Models and Parameters.

C. Model Extensions

[IV. Applications] For the model extensions, we developed models capable of simultaneously handling both modalities. Evaluations of forecasting accuracy and the generation of causality reasons were conducted jointly. Figure 2 provides a flow chart of the two model architectures we implemented. We utilize 2 model architectures - Cascaded & Joint. Both architectures leverage fine-tuned improved models [from Part B].

In our Cascaded model architecture, we integrate and assess combinations of improved Price Models and Reason Models, totaling four distinct model variations. This approach utilizes EIA reports as inputs for the RoBERTa-base Regression model, and historical sequences of price data for the Transformer Time Series model, to forecast future commodity prices. These forecasts then serve as inputs to the Reason Models, which generate explanations for the price movements between historical and predicted future prices.

Figure 3 provides the model card for the final model, which is the cascading model combining a time-series transformer with Llama2.

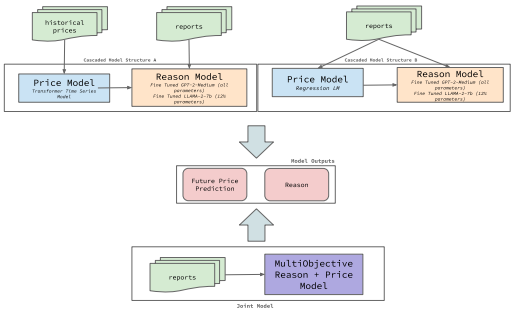

Fig. 2: Flow Chart for Extension.

Model Card

Cascading Multi-objective Time-series Model+Llama2

- **Model Details**
  - **Developed by:** Dana Golden, Dikshya Mohanty, and Khushboo Singh
  - **Model date:** 5/6/2024
  - **Model version:** 1
  - **Model Type:** Cascading model for text and time series forecasting
  - Model combines Time-series transformer with Llama2. PEFT LoRa is used to increase training speed with an r of 16, an alpha of 8, and dropout of .15.
  - **Contact:** dana.golden@stonybrook.edu, dikshya.mohanty@stonybrook.edu, khushboo.singh@stonybrook.edu
- **Intended Use**
  - The model is primarily intended to analyze energy commodities to provide a first level of analysis when creating forecasts
  - The model is primarily designed for commodity analysts and energy market economists
  - The model should not be utilized to make trades or to create a trading system. Caution should also be used applying the model to other commodities
- **Factors**
  - Usage may vary if used on data other than government reports, other commodities, or other time periods.
- **Metrics**
  - Forecasting was evaluated on MSE, MAPE and MAE while reasons were evaluated by Rouge scores
- **Training and Evaluation Data**
  - The model was trained based on futures prices as well as text data from EIA reports.
  - These datasets were chosen because Energy commodities are highly volatile, and the EIA outlook program has a rich history of providing information on market movements.
  - EIA Reports for multiple commodities were standardized for consistency.
- **Ethical Considerations:** This model utilizes data obtained from the EIA reports, which were systematically scraped from the official EIA website. While this data is publicly accessible and intended for broad public use, the scraping process was conducted with consideration to comply with the terms of use of the EIA website. Additionally, employing public data for private benefit raises certain ethical considerations.
- **Quantitative Analysis** For forecasting, MAPE scores on the evaluation set were 0.26 while for reasoning Rouge1 scores were 0.59.
- **Caveats and Recommendations:** Users should approach the model's predictions with caution, recognizing that it is unlikely to provide sufficiently accurate forecasts for trading purposes. This model is not intended for use in making financial decisions that carry significant risks. Instead, it should be viewed as an analytical tool to aid understanding of market trends and dynamics.

Fig. 3: Model Card Final Model.

In our joint model architecture, a fine-tuned GPT-2-Medium model is adapted to multiple objectives (next-word prediction & regression) to simultaneously generate a price forecast and causality reason. A simpler model architecture was selected both to facilitate easy modifications for generating various outputs and to enable comparisons with the Lora-trained Llama model. While the Llama model is fine-tuned on only a subset of parameters, the

GPT-2 model undergoes fine-tuning across all parameters. This disparity in the number of trainable parameters raises the question of whether a less complex model could potentially outperform a more sophisticated one.
Based on our manual analysis, we assumed that retaining the initial 2,600 words of each report would yield optimal results with minimal information loss. All computational models and experiments presented in this paper were executed on A100 GPUs to ensure high-performance processing capabilities and reproducibility of results.

V. Results

For evaluating the forecasting models, we used mean-squared error, mean-absolute error, and mean-absolute percentage error. For evaluating the reason models, since it's a text generation task, we used Rouge Scores (Rouge1, RougeL & ROUGE-Lsum) [41].

A. Forecasting

A summary of evaluation metrics by commodity and model is provided in this table II. In this case, the SFT model had the best performance.

TABLE II: Model Performance Forecasting

Model &Commodity	Evaluation Metrics		
	MSE	***MAE***	***MAPE***
Oil Average	496.74	18.47	.630
NG Average	5.273	1.543	.4122
Oil last timestep	10.998	2.41	.039
NG last timestep	.179	.268	.179
Oil ARIMA	25.15	2.799	–
NG ARIMA	.1645	.266	–
Oil+NG TST	.097[a]	.25	.26
Oil Autoformer	29.9	4.61	–
NG Autoformer	3.12	1.28	–
Oil+NG SFT	.017[a]	.095	.22
Oil+NG Joint Model	.046[a]	.212	.294

[a]Trained on both simultaneously.

B. Reasons

Figure 4 provides a bar chart of forecasting metrics for different models. Figure 5 represents out-of-sample predicted vs. actual prices for the time-series transformer model on oil data. Figure 6 represents the same for the SFT model. The SFT performs quite well but seems to be overfitting the data.
Table III provides Rouge scores for each language models while Figure 7 provides a graphical representation of Rouge scores for the models. The cascaded architecture utilizing a combination of fine-tuned RoBERTa SFT model for forecasting prices and Llama-2 with PEFT for generating reasons outperforms all other models.

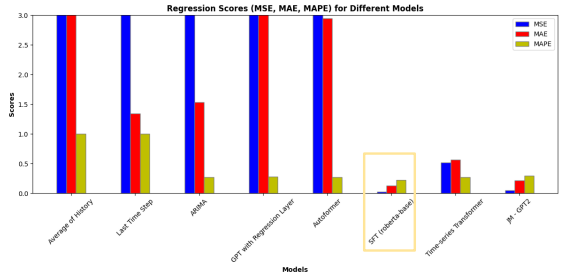

Fig. 4: Forecasting Evaluation Metrics by Model.

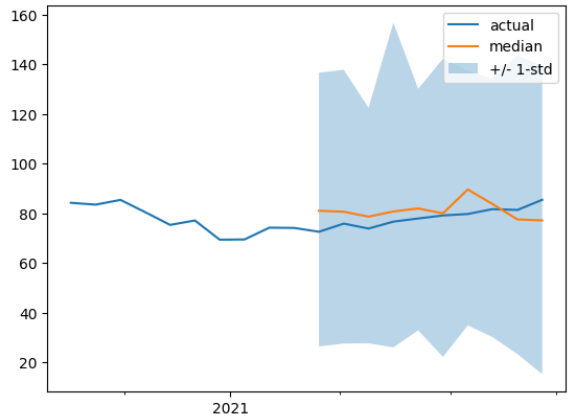

Fig. 5: Time-Series Transformer Predicted Vs. Actual.

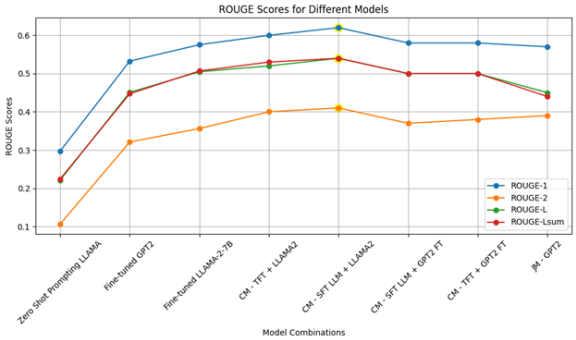

Fig. 7: Rouge by Model.

Figure 8 showcases an exemplary output from the final model for natural gas, where the reasons provided align with those stated in the report and conform to economic rationale. This output is contrasted with the explanations generated by the Llama2 zero-shot model when exposed to actual price movements. In contrast, Figure 9 displays a suboptimal output from the final model, illustrating several encountered challenges. These issues encompass the provision of information not present in the report, erroneous explanations based on economic theory, and explanations that were either nonsensical or mere repetitions of reported price movements. Additionally, the model repeatedly produced a specific response, indicating

TABLE III: Model Performance Explanation

Model &Commodity	Evaluation Metrics		
	Rouge1	***RougeL***	***Lsum***
Llama2 Zero-shot Oil	.268	.088	.201
Llama2 Zero-shot NG	.326	.244	.247
Llama2 Zero-shot STEO	.349	.214	.242
Llama2 PEFT	.57	.5	.5
TST+Llama2 FineTuned	.59	.519	.52
SFT+Llama2	.62	.53	.53
GPT2 Joint	.57	.45	.44
TST+GPT2-M FineTuned	.58	.50	.50
SFT+GPT2-M	.58	.49	.5

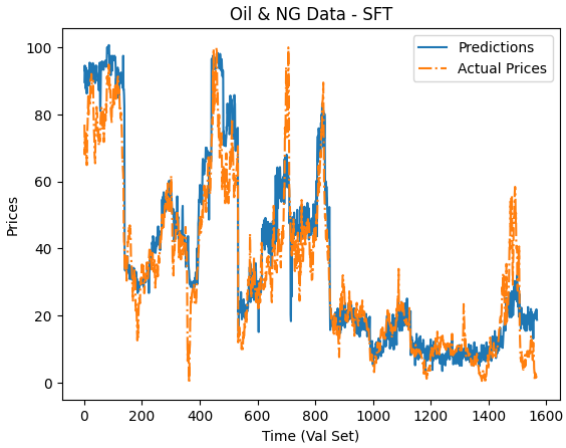

Fig. 6: SFT Predicted Vs. Actual.

a limitation in its response variability.

Reason for Movement: August 11th 2022

- Cascading TST+Llama2: "Today's natural gas price increased from last week due to volatile spot prices in the western United States driven by demand fluctuations in California and the Pacific Northwest, with SoCal Citygate and PG&E City"
- Zero-shot Llama2: "Based on the information provided in the report, the reason for the change in next week's natural gas price from this week's price is an increase of 6 cents from $7.83 per million British thermal units (MMBtu) to $7.89/MMBtu. This increase can be attributed to higher demand for natural gas in the electric power sector due to warmer-than-normal temperatures across much of the United States."

Fig. 8: Example Good reason. Report.

Reason for Movement

"Today's price decreased from last week's price due to a decrease in U.S. crude oil imports and refinery inputs, leading to lower supply and higher demand, respectively."

Fig. 9: Example Bad reason.

VI. Conclusions

In this study, we attempted to combine time series and text data for forecasting future prices and generating a causality reason, by experimenting with different model architectures. We found that the cascaded architecture leveraging Time Series Transformer and Llama2 had the best performance.

In future work, we aim to expand our methodology to additional commodities and incorporate real-time news data for dynamic forecasting. We plan to build on the current research, and explore a) a multi-task multi-objective approach that integrates both textual and time-series data, b) enhance our framework with a fusion layer that combines these features, & c) implement cross-modality by incorporating cross-attention layers, facilitating more effective interaction between the text and time-series modalities.

References

[1] Energy-Information-Administration, "Natural gas weekly update,"

[2] Energy-Information-Administration, "Weekly petoleum report,"

[3] Energy-Information-Administration, "Short-term energy outlook (steo),"

[4] A. R. P. Sarah Mouabb, Evgenia Passari, "The origins of commodity price fluctuations,"

[5] Y. H. Gu, D. Jin, H. Yin, R. Zheng, X. Piao, and S. J. Yoo, "Forecasting agricultural commodity prices using dual input attention lstm," *Agriculture*, vol. 12, no. 2, p. 256, 2022.

[6] L. Gifuni, *NLP for analysis and forecasting of crude oil prices*. PhD thesis, University of Glasgow, 2023.

[7] G. Hegde, V. R. Hulipalled, and J. Simha, "Price prediction of agriculture commodities using machine learning and nlp," in *2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, pp. 1–6, IEEE, 2021.

[8] C. T. Cortez, S. Saydam, J. Coulton, and C. Sammut, "Alternative techniques for forecasting mineral commodity prices," *International Journal of Mining Science and Technology*, vol. 28, no. 2, pp. 309–322, 2018.

[9] J.-T. Bernard, L. Khalaf, M. Kichian, and S. McMahon, "Forecasting commodity prices: Garch, jumps, and mean reversion," *Journal of Forecasting*, vol. 27, no. 4, pp. 279–291, 2008.

[10] M. K. Adjemian, "Quantifying the wasde announcement effect," *American Journal of Agricultural Economics*, vol. 94, no. 1, pp. 238–256, 2012.

[11] G. D. Bunek *et al.*, *Characterizing the Effect of USDA Report Announcements in the Winter Wheat Futures Market Using Realized Volatility*. PhD thesis, Montana State University-Bozeman, College of Agriculture, 2015.

[12] L. H. Ederington, F. Lin, S. C. Linn, and L. Yang, "Eia storage announcements, analyst storage forecasts, and energy prices," *The Energy Journal*, vol. 40, no. 5, pp. 121–142, 2019.

[13] B. Falk and P. F. Orazem, "A theory of future's market responses to government crop forecasts," 1985.

[14] J. Huang, T. Serra, and P. Garcia, "The value of usda announcements in the electronically traded corn futures market: A modified sufficient test with risk adjustments," *Journal of Agricultural Economics*, vol. 72, no. 3, pp. 712–734, 2021.
[15] B. Karali, O. Isengildina-Massa, S. H. Irwin, M. K. Adjemian, and R. Johansson, "Are usda reports still news to changing crop markets?," *Food Policy*, vol. 84, pp. 66–76, 2019.
[16] S. C. Linn and Z. Zhu, "Natural gas prices and the gas storage report: Public news and volatility in energy futures markets," *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, vol. 24, no. 3, pp. 283–313, 2004.
[17] S. Ye and B. Karali, "The informational content of inventory announcements: Intraday evidence from crude oil futures market," *Energy Economics*, vol. 59, pp. 349–364, 2016.
[18] J. Ying, Y. Chen, and J. H. Dorfman, "Flexible tests for usda report announcement effects in futures markets," *American Journal of Agricultural Economics*, vol. 101, no. 4, pp. 1228–1246, 2019.
[19] W. Yun, "Predictability of wti futures prices relative to eia forecasts and econometric models," *JOURNAL OF ECONOMIC RESEARCH-SEOUL-*, vol. 11, no. 1, p. 49, 2006.
[20] X. Li, W. Shang, and S. Wang, "Text-based crude oil price forecasting: A deep learning approach," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1548–1560, 2019.
[21] Q. Chen, "Stock movement prediction with financial news using contextualized embedding from bert," *arXiv preprint arXiv:2107.08721*, 2021.
[22] S. Baker, N. Bloom, S. J. Davis, and M. C. Sammon, "What triggers stock market jumps?," tech. rep., National Bureau of Economic Research Cambridge, 2021.
[23] X. Tang and N. Lei, "Research on cpi prediction based on natural language processing," *arXiv preprint arXiv:2303.05666*, 2023.
[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[25] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.
[26] R. Zhou and Y.-D. Shen, "End-to-end adversarial-attention network for multi-modal clustering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14619–14628, 2020.
[27] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, pp. 478–487, PMLR, 2016.
[28] J. Wu, W. Gan, Z. Chen, S. Wan, and S. Y. Philip, "Multimodal large language models: A survey," in *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256, IEEE, 2023.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[30] K. Rasul, A. Ashok, A. R. Williams, H. Ghonia, R. Bhagwatkar, A. Khorasani, M. J. D. Bayazi, G. Adamopoulos, R. Riachi, N. Hassen, *et al.*, "Lag-llama: Towards foundation models for probabilistic time series forecasting,"
[31] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, *et al.*, "Instruction tuning for large language models: A survey," *arXiv preprint arXiv:2308.10792*, 2023.
[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[33] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
[34] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950–1965, 2022.
[35] A. Gema, L. Daines, P. Minervini, and B. Alex, "Parameter-efficient fine-tuning of llama for the clinical domain," *arXiv preprint arXiv:2307.03042*, 2023.
[36] F. Liu, X. Lin, Z. Wang, S. Yao, X. Tong, M. Yuan, and Q. Zhang, "Large language model for multi-objective evolutionary optimization," *arXiv preprint arXiv:2310.12541*, 2023.
[37] G. Heitz, S. Gould, A. Saxena, and D. Koller, "Cascaded classification models: Combining models for holistic scene understanding," *Advances in neural information processing systems*, vol. 21, 2008.
[38] A. M. Medina and J. A. H. Álvaro, "Using generative pre-trained transformers (gpt) for electricity price trend forecasting in the spanish market," *Preprints*, March 2024.
[39] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 22419–22430, 2021.
[40] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
[41] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.