Transformers but not the Kind You're Thinking of: Commodity Market Forecasting using Natural Language Processing

Dana Golden [1 3] Dikshya Mohanty [2] Khushboo Singh[2]

[1]Stony Brook Economics [2]Stony Brook Computer Science [3]This material is based upon work supported by the National Science Foundation under Award No. 2125295 (NRT-HDR: Detecting and Addressing Bias in Data, Humans, and Institutions). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Motivation

- Commodity markets are highly volatile but essential to macroeconomics
- Commodity markets are more dependent on physical factors impacting supply and demand
- EIA releases commodity market reports providing information on markets
- LLMs can be used to forecast prices and reasons for forecasts based on reports

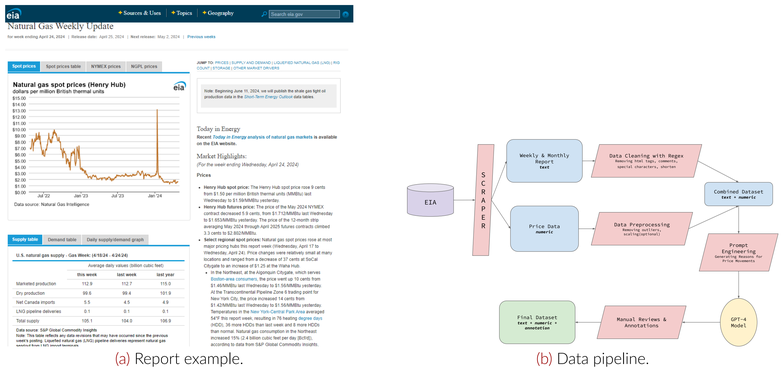
Research Overview

- Combine time-series forecasting with natural language processing to create explainable forecasts of movements of energy
- Utilize modern LLM and transformer-based techniques to forecast time series and provide text-based reasons for forecasts
- Develop cascading model to take in futures data and reports separately and feed forecast to reason model
- Develop joint model combining forecasting and explanation into one task
- Compare forecasting performance with other state-of-the-art time series forecasting techniques
- Evaluate models on performance of explanations

Data

Commodity Area	Commodities	Reports	Tokens	Earliest Date
Agriculture	20	2021	19.4 M	1999
Metals	27	853	12.2 M	1997
Energy	13	615	12.3 M	1997

Table 1. Data Size by Area



(a) Report example.

(b) Data pipeline.

Figure 1. Data processing and example.

Models and Methodology

Figure 2. Baseline models.

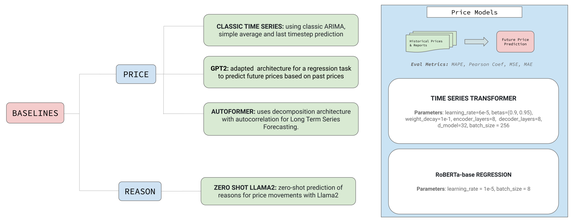
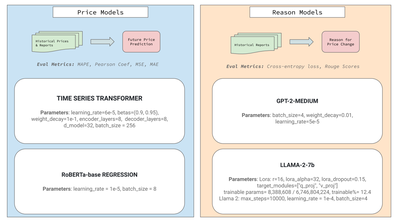


Figure 3. Improved models.



Model Performance for Forecasting

- Supervised fine-tuned transformer outperformed all other methods for forecasting including traditional time-series forecasting
- LLaMA3 performed best for generating reasons

Figure 4. Predicted vs. Actual Live Cattle.

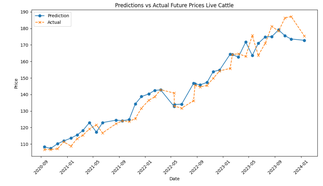


Figure 5. Predicted vs actual Crude Oil.

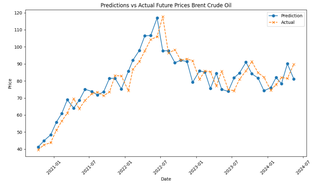
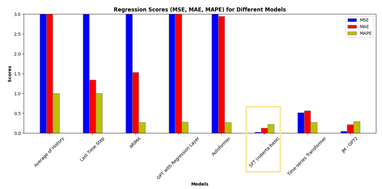


Table 2. Model Performance Forecasting

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.

Figure 6. Errors by model.



Results: Reason Model

Figure 7. Rouge Scores by Model.

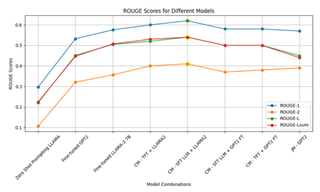


Table 3. Model performance reason model.

Model &Commodity	Evaluation Metrics *Rouge1*	*RougeL*	*Lsum*
LLaMA3 Zero-shot Oil	.268	.088	.201
LLaMA3 Zero-shot NG	.326	.244	.247
LLaMA3 Zero-shot STEO	.349	.214	.242
LLaMA3 PEFT	.57	.5	.5
TST+LLaMA3 FineTuned	.59	.519	.52
SFT+LLaMA3	.62	.53	.53
GPT2 Joint	.57	.45	.44
TST+GPT2-M FineTuned	.58	.50	.50
SFT+GPT2-M	.58	.49	.5

Reason Example

Figure 8. EIA Report.

Prices

- **Henry Hub spot price:** The Henry Hub spot price rose $1.62 from $7.89 per million British thermal units (MMBtu) last Wednesday to $9.51/MMBtu yesterday.
- **Henry Hub futures prices:** The price of the September 2022 NYMEX contract increased $1.042, from $8.202/MMBtu last Wednesday to $9.244/MMBtu yesterday. The price of the 12-month strip averaging September 2022 through August 2023 futures contracts climbed $1.3 cents to $7.545/MMBtu.
- **Select regional spot prices:** Natural gas spot prices rose at most locations this report week (Wednesday, August 10, to Wednesday, August 17). Week-over-week increases at major pricing hubs ranged from 63 cents at Malin, located on the California-Oregon border, to $1.89 at SoCal Citygate in Southern California.
  - The price at PG&E Citygate in Northern California rose $1.28, up from $9.14/MMBtu last Wednesday to $10.42/MMBtu yesterday. In northern California, PG&E's maintenance schedule includes ongoing work on the Redwood pipeline and Buckeye station through the end of the month. The Redwood pipeline delivers natural gas from Malin, Oregon, to the San Francisco Citygate. PG&E expects available pipeline capacity to be between 70% and 90% for the next few weeks.
  - The price at SoCal Citygate in Southern California rose $1.89, up from $10.21/MMBtu last Wednesday to $12.10/MMBtu yesterday. In the Southwest, El Paso Natural Gas Company rescheduled pipeline remediation on Line 1100 from Wenden, Arizona, to Ehrenberg, Arizona, but continues to report that 450 million cubic feet per day (MMcf/d) of natural gas is currently unavailable due to the on-going repair of Line 2000. In the Permian Basin of New Mexico, a major region of production that supplies natural gas to California, production fell 0.3 billion cubic feet per day (Bcf/d) week over week, or 6%, according to data from PointLogic. Temperatures across California remain high. The California ISO issued a Flex Alert on Tuesday urging

- Good Reason: Cascading TST+LLaMA3:
  "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"
- Bad Reason: Cascading TST+LLaMA3:
  "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."

Conclusions and Next Steps

- Combined economic and natural language processing insights to create explainable forecasting method for futures market
- Combined time-series transformer and LLaMA3 to create cascading model, feeding forecasts into model of reasons
- Used GPT-2 to create joint model, jointly predicting price movements and reasons
- Extensions and Next Steps:
  - Implement multi-task, multi-objective learning with time series and text features
  - Create fusion layer with time series and text features
  - Implementing cross-modality by adding cross-attention layers across both modals
  - Incorporating real-time news data
  - Adding Metals and agricultural data, adding other outlook programs
  - Apply techniques to counterfactual analysis