Reconstruction of the Economic Policy Uncertainty Using Large Language Models

Jacky Yeh

This version: November 2023

0.0.1 Result

What utilize OpenAI's ChatGPT-3.5-turbo API to test the predictability and feasibility of the NLP foundation model in constructing textual indices, with our main focus on the EPU index. In our experiments, we employed, but were not limited to, three approaches: zero-shot prompting, few-shot prompting (3, 6, 8 shots), and fine-tuning. Essentially, one-shot prompting involves purely asking questions to ChatGPT without providing any examples. On the other hand, few-shot prompting does involve providing some examples, which can often result in more accurate and thorough answers from ChatGPT. Below are the news articles that were used in few-shot prompting.

Table 1: news for few-shot prompt

澳洲央行昨日意外調高基準利率，隔夜拆款利率從原
【丁威　台北報導】國內消費市場不景氣，國內 KTV 龍頭之一的錢櫃（8359）今年前 4 月業績也衰退近 2 成，為了力抗景氣「寒流」，錢

Notes: ADD NOTES HERE

prompt

System:
I am an economist working on constructing Taiwan's Economic Policy Uncertainty Index (EPU index). My primary goal is to classify wheter a news should be excluded when constructing EPU index in Taiwan. There are two criteria I'm considering to exclude a news.
Criterion1:
The main idea of the news is either historical accounts or abstract subjective inferences, which won't impact Taiwan's economics for sure. Hence, this kind of news should be excluded.
Criterion2:
There main idea of the news is not related with Taiwan. For example, the people or companies mentioned in the news have nothing to do with Taiwan or the events in the news don't actually happen within Taiwan. I will excluded the news as well.

Notice that you can first justify wheter there is a person, company or event in news related to Taiwan. If there isn't any, it should be excluded with high probability.

{few shot examples}

Human:
Help me complete the classification task identifying whether the given news should be excluded.
News: {one new from test set}

Output Instructions:
The output should be formatted as a JSON instance that conforms to the JSON schema below.
As an example, for the schema "properties": "foo": "title": "Foo", "description": "a list of strings", "type": "array", "items": "type": "string", "required": ["foo"] the object "foo": ["bar", "baz"] is a well-formatted instance of the schema. The object "properties": "foo": ["bar", "baz"] is not well-formatted.
Here is the output schema: **{json schema}**
Besides, don't forget to escape a single quote in the reason section and be aware of your reasoning's token length.

Furthermore, we explored additional prompting strategies, including the Chain of Thought (CoT) method proposed by the Google Research Brain Team. CoT has demonstrated a broad impact, and in prior studies, its effectiveness has been shown in tasks related to mathematical derivation and commonsense reasoning. In Table 3, we illustrate our endeavor to incorporate step-by-step reasoning into our approach.

Table 2: CoT reasoning in few-show prompt

First, the news is related to Taiwan's economics as it mentions Taiwan is one of the countries which have experi
First, the news is related to Taiwan's economics as it points out the relation between the quality of voluntary military and the T

Notes: ADD NOTES HERE

The intuition behind CoT is analogous to solving a math problem. When faced with a mathematical challenge, our approach involves thinking step by step, linking these steps together to arrive at the final answer. Applying this idea of chaining, we break down our task into three components: Economy, Policy, and Uncertainty—the fundamental concepts of EPU.

Despite the disappointing results, the concept used to identify whether a news item is related to EPU appears to be more complex and abstract. Consequently, it's highly challenging to provide the reasoning part in few-shot examples, prompting the need for ChatGPT to break down the important concepts and make more logical statements or reasoning. We hope future literature can further unveil the hidden potential of this prompting technique and develop a more concrete strategy for the EPU construction task.

Table 3: ChatGPT answering

〔編譯許世函／綜合報導〕中國央行週四晚間出乎外界意料宣布降息一碼（．二五%）。美國聯準會（Fed）主席柏南克則表示，目前美國不
【陳文蔚　台北報導】原以為大選紛擾將影響房地產市場，不過國有財產局發現，

Notes: ADD NOTES HERE

So far, we've illustrated various prompt engineering concepts, and one might understand the benefits of providing some examples. However, can we input as many examples as we want without any restrictions? The answer is quite straightforward—no. There is a limit to restrict the input context length for ChatGPT. In our experiment, to utilize the research fund in the most efficient way, we used the base 3.5 model with 4k tokens as long as we could. For longer news, we had to switch to the 16k model, which comes with a much higher cost. It's worth mentioning that the 3.5 model is not the most powerful model provided by OpenAI; there is a 4.0 version renowned for its better understanding not only of textual data but also images, known for its multimodal capability.

When it comes to using a Large Language Model, there's always a two-choice problem: whether to go for fine-tuning or simply conduct prompt engineering. Although, thanks to the pre-trained capability, ChatGPT is already ready-to-use in many tasks, we also gave fine-tuning a shot.

It's crucial to emphasize that there are many aspects of fine-tuning techniques. Traditionally in the NLP domain, it refers to training the model on top of pre-trained model parameters. There might be concerns that the parameters could get polluted when there is a lot of noise in the personalized data of the downstream task. In our case, we used the fine-tuning API provided by OpenAI, and most importantly, we don't know how it works. As our main goal is to explore the possibility of ChatGPT, this is the only way to perform fine-tuning.

One benefit of fine-tuning is that as long as we have the fine-tuned model, there is no need to provide few-shot examples in each request. In other words, we can use zero-shot prompting to replace few-shot prompting with ease. Nonetheless, the cost is expensive because the workflow is intensive, and it's not as straightforward as asking a question and receiving a response. There is always a trade-off, and we are curious about which approach is more suitable for constructing the EPU index.

The result of fine-tuning can't compete with the best-performing few-shot prompting, especially in the case where the provided training instances are only labels (without reason). We observed that the training process reaches a bottleneck when approaching 100 steps, and the loss also seems to have no room for improvement. Interestingly, when we add reasoning to our training instances, meaning we force ChatGPT to answer with a label plus a reason, the loss seems to improve, but the evaluation metrics don't show much difference. This is because we have limited knowledge of the process, and only a few parameters can be tuned, such as the epoch and learning rate. We decided to conclude our exploration at this point.

In conclusion, the few-shot prompting with six examples performed the best. This outcome sheds light on how powerful the foundation model is, revealing that even when provided with only six examples, it can yield substantial improvements. Although in the current state ChatGPT3.5 can't outperform deep learning models when constructing the EPU index, it provides an alternative that strikes a balance between training time and result performance.

Table 4: Evaluation Metrics for Test Set Contained in 7000 News Articles

	(1)	(2)	(3)	(4)
	micro_f1	macro_f1	weighted_f1	precision_0
8 shot with reason	0.707	0.705	0.799	0.707
6 shot with reason	0.672	0.674	0.762	0.677
3 shot with reason (fine-tuned 1000)	0.555	0.598	0.63	0.58
6 shot no reason	0.54	0.585	0.621	0.566
zero shot with reason (fine-tuned 1000)	0.417	0.579	0.59	0.472
zero shot no reason (fine-tuned 1000)	0.372	0.588	0.589	0.438
zero shot with reason	0.401	0.58	0.588	0.459
zero shot no reason	0.415	0.572	0.587	0.469

Notes: ADD NOTES HERE

Figure 1: confusion matrix of best performing model

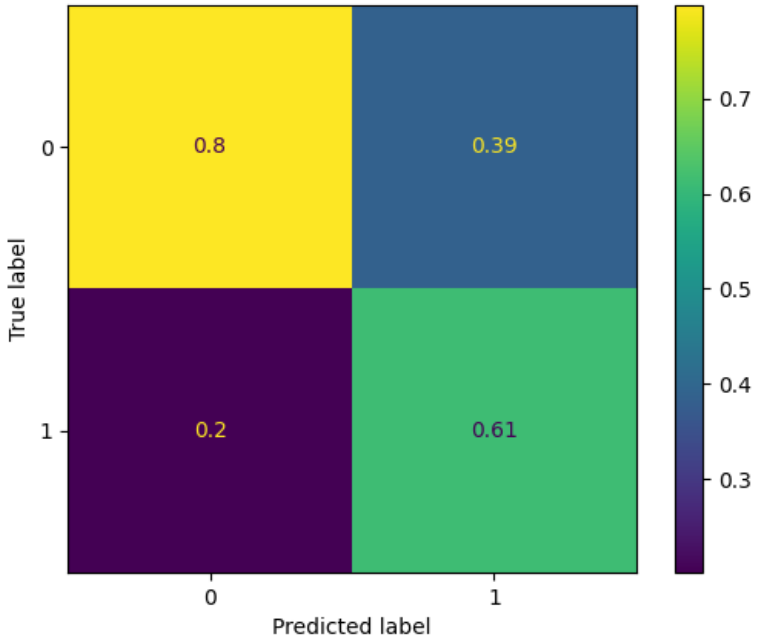

Figure 2: loss of fine tuning without reason

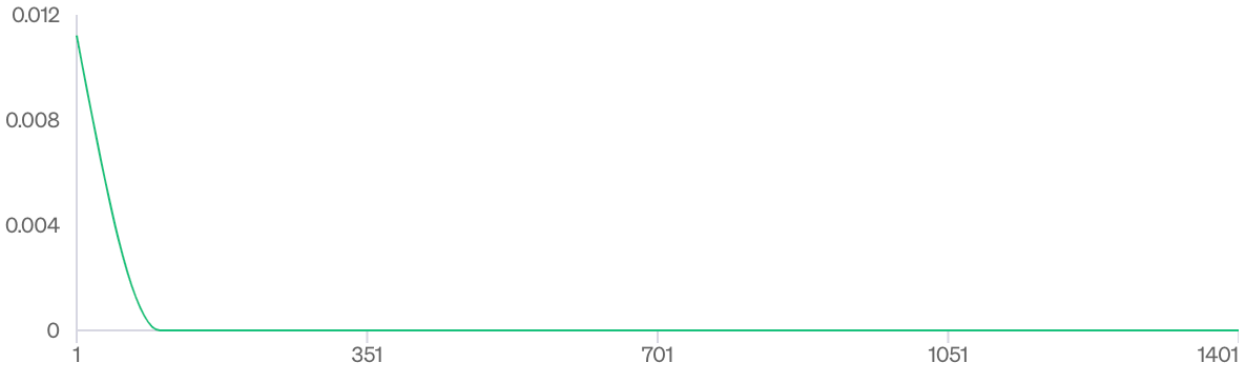

Figure 3: loss of fine tuning with reason