ChatGPT Performance Analysis on Annotation Tasks

Leonardo Matone

Abstract

Many Natural Language Processing (NLP) applications require labeled data which is both time consuming and expensive. Projects of large scope usually require many manual coders, and niche areas of study require these coders to have a degree of training or knowledge of the field. The advance of Large Language Models (LLMs) in the past decade have led to huge improvements in language processing and comprehension. Chatbots like ChatGPT are capable of outperforming annotators on specific datasets, and their zero-shot accuracy provide compelling evidence that LLMs could be the future of text annotation. This study will be focused on three things: (1) how well does ChatGPT compare to human coders on a multi-class problem, (2) how does it compare against other NLP approaches and (3) with slight prompt engineering and a few pre-training steps, does the ChatGPT approach offer efficiency benefits over the industry standards?

Keywords: ChatGPT, Text Annotation

1. Problem Overview

NLP tasks often require large amounts of labeled data. Modern classification approaches are heavily resource intensive and require labels to compute their gradients, and unsupervised models require labels to quantify their accuracy. ChatGPT has been demonstrated to outperform human coders on certain tasks (Gilardi et al.), but has not been compared against NLP approaches to the same task. This report is concerned with the following:

1. How well do ChatGPT's zero-shot and fine-tuned models hold up against an NLP model trained on the same task, with the same labeled data?
2. What forms should prompt-engineering on multi-class classification problems take?
3. How much fine-tuning can be generalized to an unseen domain?

LLMs have demonstated their ability to perform very well across domains. this project aims to explore the power of ChatGPT in annotation tasks by comparing their representative power against other NLP approaches. It also aims to evaluate the best methods of prompt-engineering and fine-tuning with a lens for unseen domains.

2. Data

The data for this project is a collection of reviews from New Orleans schools. The data was collected and annotated by the Tulane Economics Department, and consists of 500 student-annotated reviews, categorized into level 1 and level 2 categories, with level 2 denoting higher specificity and complexity. There are manual annotations from student-coders for the level 1 categories, which include: school-level features, physical environment, instruction & learning, school staff, overall quality, school culture, and resources. Manual coders were instructed to label each review with all applicable level 1 categories. The data also includes level 1 classifications from the Tulane Economics Department's NLP model for these 500 reviews. A version of this dataset exists in this report's GitHub repository.

There are two datasets which are merged by index for the purposes of this project. The first contains the annotations from the human coders and the NLP model, and the second contains the actual reviews and other relevant information per review. The dataset available in the project's GitHub is the merged dataset consisting of both sources.

3. Methods

To perform classification on the review dataset, we designed an interface layer to interact with OpenAI's API. Because ChatGPT is inherently creative and chooses its output stochastically based on the input, the format of the output must be standardized. In our testing, we did extensive prompt-engineering with the goal of encouraging accuracy but also standardizing output. Our initial interface layer would work sporadically, the stochastic nature of the output (both in terms of predictions and formatting) made this unreliable. Instead of penalizing incorrect response formats (as with a large enough token count on the review, ChatGPT seems to focus on directions less), we specify a limitation on stochastic responses in each API call using a ́temperatureánd ́top_pśetting of 0.1 and 0.2, respectively. Temperature denotes the creativity and randomness of the response. ChatGPT has been trained to handle a variety of tasks from code to text generation, and both require different character-level stochastic settings. By specifying the temperature as 0.1, we encourage the model be less random and more reliable, as is expected when ChatGPT produces code. Similarly, top_p represents the top token probability mass, the lower we set this, the less variance in our response.

Our interface layer allows for wiggle-room in how ChatGPT receives our unlabeled text and what prompt and categories it is given. The default prompt is:

```
"Which of these topics: {topics} are discussed in this review: {text}.
For all topics, respond in the format: "topic": 0 or 1."
```

Our initial comparisons will use no prompt engineering or fine-tuning in order to estimate ChatGPT's zero-shot capabilities in full. We will compare against the Tulane Economics Department's NLP model and human coders, and then begin optimizing this LLM approach.

4. Preliminary Experiments & Results

We have completed a series of small experiments concerned with establishing a functional input layer and several rudimentary analyses of the results against existing NLP approaches and human-coded annotations. Each is labeled 1-4 in this project's GitHub repository, under notebooks.

4.1 Preliminary Evaluation

As a first test, we used OpenAI's API to interact with ChatGPT and feed it a single review. We use the level 1 categories defined by the Tulane Econommics Department which consist of *school-level features*, *physical environment*, *instruction & learning*, *school staff*, *overall quality*, *school culture*, and *resources*. Below is the sample we passed to ChatGPT 3.5 and ChatGPT 4.0.

> (PLEASE DONOT SEND YOUR KIDS HERE) LOTS OF BULLYING AND FIGHTS I went here when I was a kid and it was bad and now it is worst as it has ever been I well adment the teachers are ok I loved Mrs.Long and Mrs.Right but there were lots of fights and lots of people in my class was below grade level and had 50's,40's,30's,and even F's SO I would just like everyone to know about this school.Thanks

This review is a good example of some of the NLP challenges which make learning from such data difficult. The formatting is inconsistent, at times the user is ambiguous, and some of the categories themselves are questionably general. From Table 1 below, it is evident that GPT is not going to always match our source of truth, the human annotators. Arguably, the results of GPT's labeling are comparable, which is what the existing literature would have us believe. In this case, Both GPT models annotated "Student discipline" as 1, which is implicit to a human reader from the review, but for whatever reason the human annotators did not flag it. Either way, the performance of both GPTs seems to leave something to be desired.

Table 1: ChatGPT 3.5/4.0 Predictions for 1 Review

Category	hum	3.5	4.0
Student discipline	0	1	1
Teacher quality	0	1	1
School culture	0	0	1
Building quality	0	0	0
School staff	1	0	1
Instruction	1	0	0
School safety	1	0	1

*results of preliminary test

4.2 Interface Layer

To make a prediction over all 500 of our labeled reviews, we designed an interface layer to query OpenAI's API and retrieve a prediction for each record. After combining the NLP/human coded dataset with the full dataset containing all review documents, we ran a test on a subset of 20 records. Each test has a real-world cost in OpenAI's payment heirarchy, so testing was limited in this stage. The cost scales with token counts, so this stage includes no prompt-engineering. On the level 1 categories, we compute the Jaccard similarities between the manually annotated reviews and ChatGPT 3.5 and ChatGPT 4.0's annotations, which you can see in Figure 1 (final report will have larger font and alpha to better see comparison):

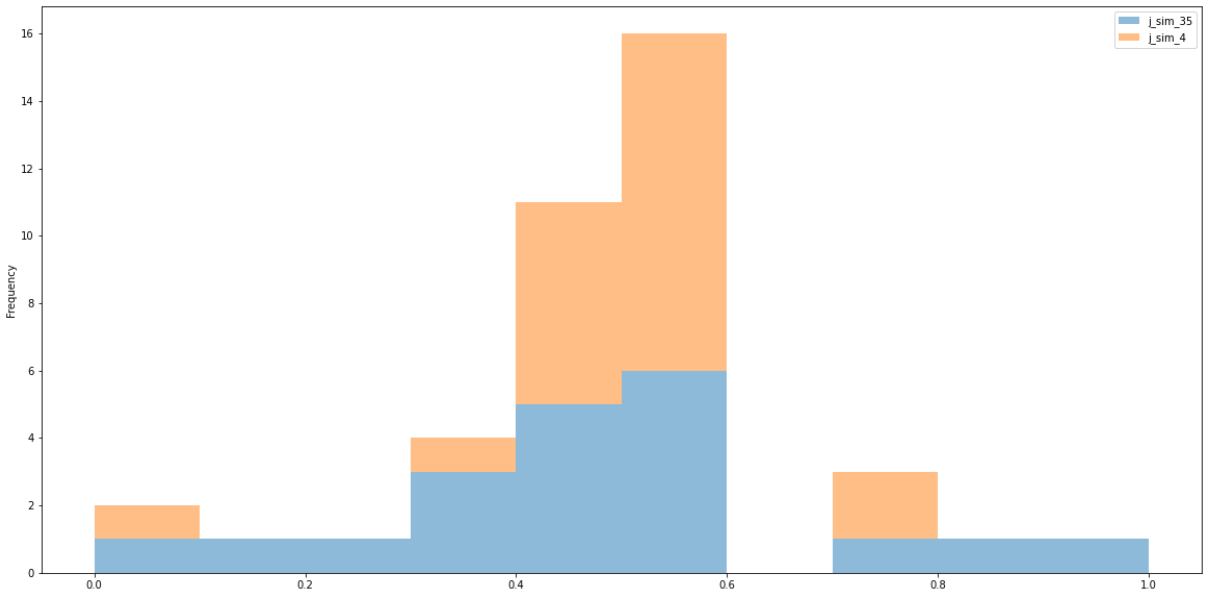

Figure 1: Jaccard Scores for 3.5 and 4.0

An interesting discovery in this experiment is the impact of tokens on the output of GPT 3.5 & 4.0. Both output different results based on the length and content of the review as they should, but often will make different guesses with subsequent calls for the same review. This is due to the stochastic nature of the output, which, even with our passed

temperature and top_p parameters, cannot be stifled permanently without fine-tuning or inflating our token counts to specify formatting in-depth. The final report will take steps to experiment with both of these alternatives.

4.3 Full Test

With our preliminary tests and interface layer complete, we ran ChatGPT 3.5 ad ChatGPT 4.0 on the full dataset of 500 reviews. This is a time-intensive and high-cost endeavor (taking 30 minutes and costing $5.00 in total), which we will aim to replicate at most 2 more times to compare prompt-engineering approaches. This first full test yielded unexpectedly low results, which are summarized in Table 2. The accuracy and precision of both LLMs is much lower than the NLP model, and while this is zero-shot learning, the existing literature would lead us to expect better results. Recall is slightly higher for both LLMs, as shown in Table 2. This is likely due to the increased variance in false negatives and false positives compared to the NLP model, shown in Figure 2. It is evident that the NLP model has been tuned to avoid false positives, and both GPT models try to get every category that could be in the review, perhaps defaultly prioritizing recall.

Table 2: Model Performance Comparison

Name	Jaccard Score	Recall Score	Precision Score	F1 Score	Accuracy Score
coded	0.995976	0.995976	0.995976	0.995976	1.000000
nlp_df	0.583252	0.623259	0.877005	0.704345	0.735844
predictions_35_df	0.479884	0.653679	0.672789	0.607320	0.603334
predictions_40_df	0.511953	0.692795	0.683396	0.647616	0.620868

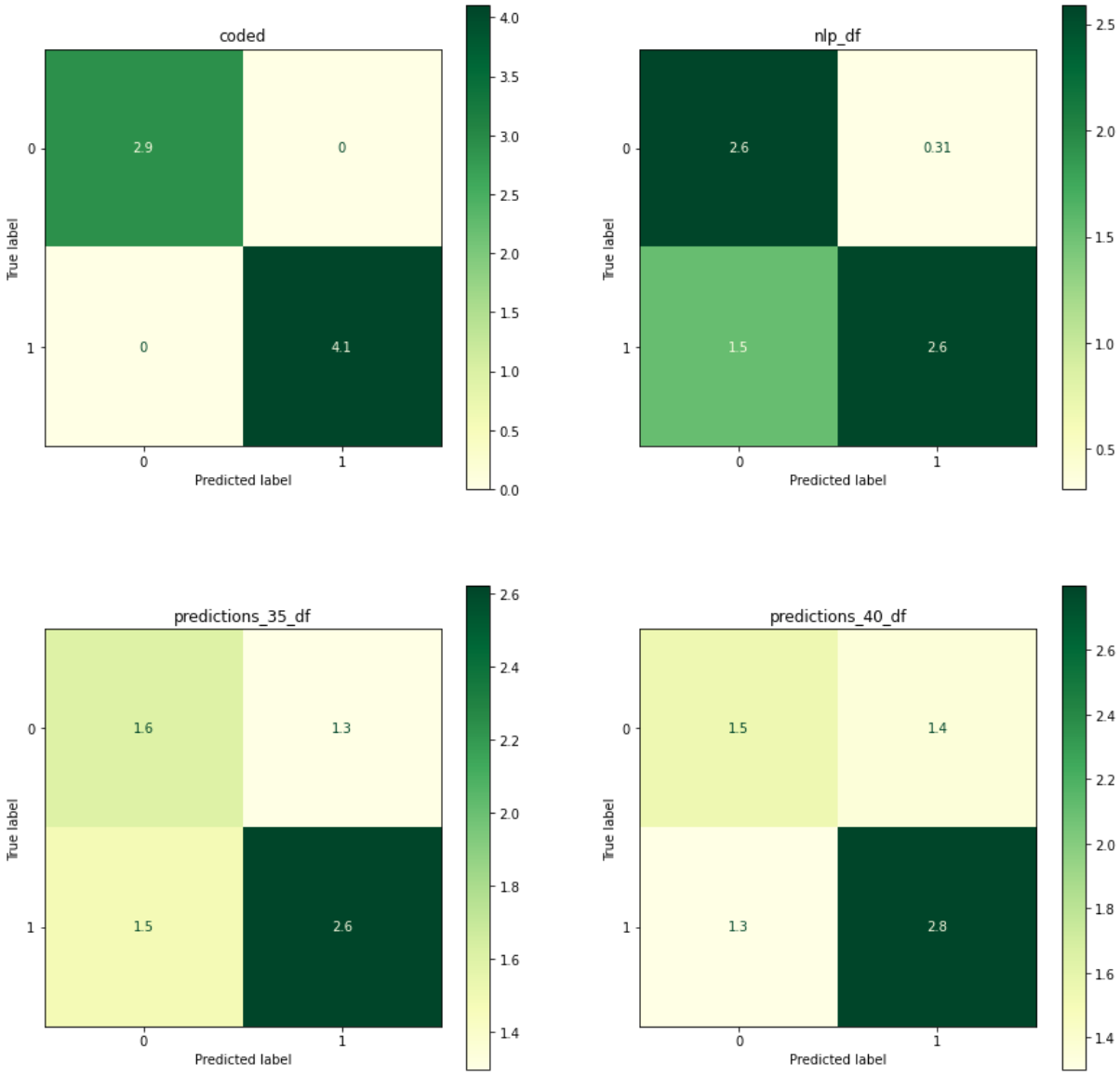

Figure 2: Averaged Confusion Matrices

4.4 Category Analysis

Figure 2 illustrates the averaged confusion matrices across all 500 predictions. We studied each category's confusion matrices independently to evaluate where OpenAI's strengths and weaknesses lie across 3.5 and 4.0. We found that GPT's interpretation of labels in its prompt fall into three categories: unclear comprehension, clear comprehension, and somewhat-clear comprehension. The deciding factors for each of these categories are the ambiguity and/or expressiveness of the label's terms.

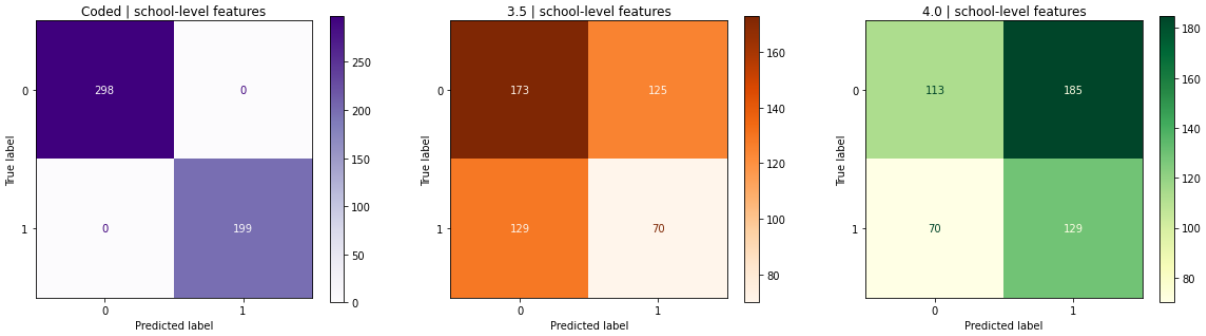

Figure 3: Unclear Comprehension (School-Level Features)

Unlike the NLP model, ChatGPT has virtually the same amount of information the human coders received. It understands the name of the label, and attempts to understand which reviews contain the label solely based on the language embedding it accumulated from training. Thus, more ambiguous labels like "school-level features" or "resources" suffer in accuracy and experience confusion, even between 3.5 and 4.0. This is evident in Figure 3, where 3.5 and 4.0 have wildly different understandings of what "school-level features" means, exhibiting different trends in how they label and varying wildly across false positives and false negatives.

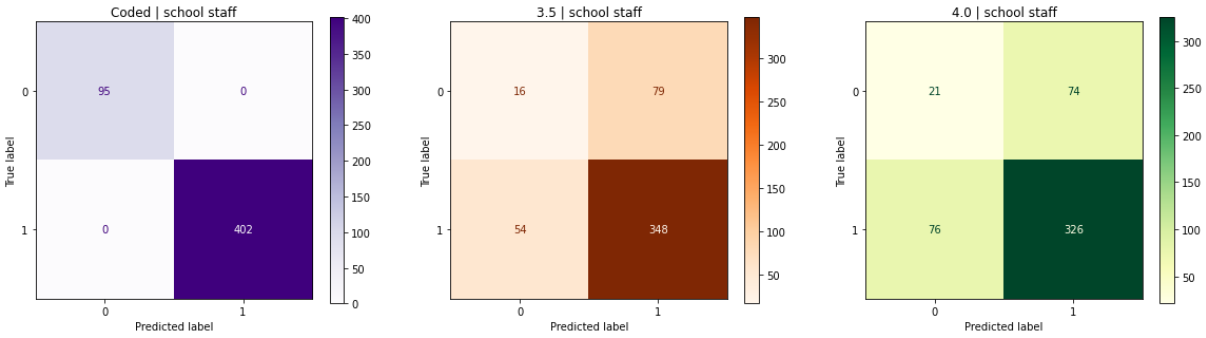

Figure 4: Clear Comprehension (School Staff)

As illustrated in Figure 4, "School Staff" is an example of a more interpretable label, with far less variance across false negatives and false positives. Figure 5 illustrates somewhat-clear comprehension. While both models tend to label positively like the coders, there is greater variance in false positives and negatives.

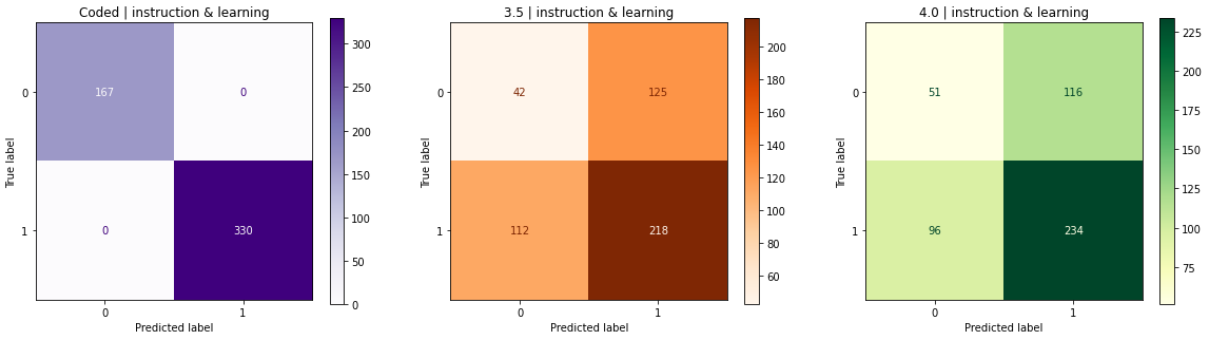

Figure 5: Somewhat Clear Comprehension (Instruction & Learning

5. Related Work

Considering the relative youth of ChatGPT, there have been a considerable number of papers written studying the use of ChatGPT as an annotator. The primary example of which has already been cited. *Gilardi et al.* demonstrate that ChatGPT can outperform MTurk annotators. Running a similar experiment, they compared the results of ChatGPT's annotation against the human annotations from MTurk, using a gold-standard annotation from research assistants as their source of truth. Their experiments were conducted on a range of tasks, ranging from binary to multi-class classification. Their results indicate that ChatGPT is well suited to annotation tasks, but do not compare the approach of ChatGPT against existing NLP approaches. Additionally, they do not fine-tune or experiment with prompt-engineering. They utilize ChatGPT's zero-shot learning alone.

Dai et al. introduce a method to use ChatGPT for generating new text data, aimed at augmenting existing datasets to improve model training in text classification. It explores various data augmentation techniques at character, word, and sentence levels, emphasizing the importance of generating diverse and semantically consistent samples to enhance the robustness and performance of NLP models. While this is related to this project as both intend to augment and upgrade the availability of text data for learning, this is a different area altogether.

Wang et al. illustrate ChatGPT 3.5 as a cost-effective solution for data labeling, highlighting its potential to significantly reduce expenses compared to traditional human labeling while maintaining a high level of data quality. It explores the process of using ChatGPT 3.5 for labeling and the implications for cost, efficiency, and model performance, suggesting that leveraging ChatGPT 3.5 could work well for data annotation tasks. This article is a precursor to *Gilardi et al*, and while it is part of the relevant literature it is unrelated to this project's goals for the same reason.

Kuzman et al. focus on the domain of automatic genre identification. The study investigates ChatGPT's capabilities against traditional NLP models. By annotating text

data with genre labels and comparing the performance of ChatGPT with a multilingual transformer-based model, the paper explores the potential of large language models to streamline or even replace manual annotation processes in specific contexts, signifying a shift towards more efficient and automated methods of data labeling. This is very close to what we aim to accomplish with this project, but like *Gilardi et al*, they do not fine-tune or perform any prompt-engineering, instead studying the zero-shot learning capabilities of ChatGPT.

Huang et al. examine ChatGPT's effectiveness compared to human annotators in identifying implicit hate speech, a nuanced and complex task. Their findings delve into ChatGPT's potential advantages and its limitations, suggesting that while ChatGPT shows promise in automating the detection of such content, it may not fully grasp the subtleties that human annotators can discern. This paper illustrates the domain limitations on ChatGPT, and the fact that it struggles in certain environments due to its inherent bias.

6. Timeline

There are two more planned steps to this project. First, we will perform prompt-engineering and fine-tuning steps to examine how well ChatGPT holds up to NLP approaches. We will perform these steps with a lens for generalization on this task across domains, withholding domain-specific information. This is with the intention of our second step, which is the application of this ChatGPT to an unseen labeled domain to assess the generalization of a fine-tuned ChatGPT model.

7. References

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30). https://arxiv.org/abs/2303.15056

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, & Xiang Li. (2023). AugGPT: Leveraging ChatGPT for Text Data Augmentation. https://arxiv.org/abs/2302.13007

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, & Michael Zeng. (2021). Want To Reduce Labeling Cost? GPT-3 Can Help. https://arxiv.org/abs/2108.13487

Taja Kuzman, Igor Mozetič, & Nikola Ljubešić. (2023). ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification.

https://arxiv.org/abs/2303.03953

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In Companion Proceedings of the ACM Web Conference 2023. ACM.