Using LLMs as Data Annotator

Ansh Kumar Dev

1 Overview

The project aims to estimate the performance and reliability of using large language models (LLMs) as data annotators. We are addressing the challenge of annotating customer reviews based on their references to various product features. Traditionally, this task involves multiple human annotators, but as the volume and complexity of the data increase, more time and resources are required to achieve a consistently reliable annotation. This is where LLMs come into play, providing a fast and efficient alternative for data annotation.

In this project, we focus on annotating multiple labels for given reviews using LLMs, a process that can be time-consuming for human annotators. Initially, the review data is manually annotated by an expert. Subsequently, the same data is annotated using the LLM. We then compare the performance of the LLM with the expert-annotated data to assess its reliability.

We evaluate the inter-annotator agreement between the two data sets using Cohen's kappa score to analyze this reliability. Finally, we implement a basic classifier to demonstrate the usability of the annotated data.

2 Methodology

2.1 About the data

The data created for this project consists of a review dataset that includes a total of 100 reviews, each labeled with six categories. These six multi-class labels are: 'Battery Life,' 'Build Quality and Durability,' 'Comfort and Fit,' 'Features and Functionality,' 'Sound Quality,' and 'Value for Money.' The product reviews have been gathered from various customers through web scraping. The purpose of creating this dataset is to analyze the product features highlighted by customers in their reviews, providing deeper insights beyond just star ratings. This approach aims to enhance the understanding of the product across different aspects.

2.2 Human annotations

The initial task involved manually annotating the data. Each review has been annotated by a domain expert, ensuring that the generated annotations are reliable. In this dataset, each multi-class label is treated as a binary classification and is checked for its relevance in the review. The annotations adhere to a set of guidelines that the annotators follow while labeling the data. Annotation guidelines consist of rules and instructions that specify how data should be labeled or annotated.

2.3 Large Language Model as annotator

In this part, we utilize large language models (LLMs) to replicate the work of human annotators. The same guidelines are provided to the LLMs for annotating the data. This approach not only accelerates the annotation process but also addresses the challenge of having insufficient human annotators. However, a key question arises: Is the data annotated by the LLMs reliable? We will explore this concern in the subsequent analysis section.

For the task of label annotation by LLMs, the models used are GPT-4 and GPT-4o-mini. These models were selected because they are among the top models provided by OpenAI.

3 Experiments

The data annotated by human experts is compared with that annotated by a Large Language Model (LLM). First, we analyze the frequency of each category in both the human-generated dataset and the LLM dataset. Next, we present a side-by-side comparison of the LLM models with the human-annotated data. The image below shows a bar graph illustrating the frequency of labels identified in the review text for both human and LLM annotations. This graph highlights how often the LLM selects a particular label and how closely these selections

align with the human-annotated data. From Figure 1, it is evident that the labels "feature and functionality" and "value for money" show a significant difference in the number of annotations. The reasons for this discrepancy will be discussed further in the paper.

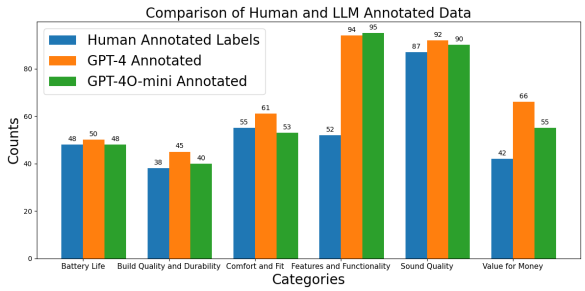

Figure 1: Comparing Human and LLM Annotated data

The performance of each model is assessed by calculating precision scores for both the human-annotated data and the data annotated by the large language model (LLM). This process involves comparing each model's results to the expert-annotated data to determine their accuracy. The table below presents the precision scores for all labels, based on the annotations from the two different models. In this context, the gold labels refer to the annotations made by human experts.

Category	GPT-4	GPT-4o-mini
Battery Life	0.94	0.92
Build Quality and Durability	0.76	0.73
Comfort and Fit	0.92	0.87
Features and Functionality	0.549	0.547
Sound Quality	0.97	0.91
Value for Money	0.65	0.71

Table 1: Precision matrix for LLM annotated data

When using large language models (LLMs) as annotators, it's essential to evaluate the reliability of their annotations. We utilize Cohen's Kappa score for this assessment. This score measures the level of agreement between human annotations and those from LLMs. We calculate this score for various LLMs to determine which models perform best for the task. A higher Kappa score indicates greater consistency across different annotators, demonstrating that the method is reliable and that the interpretations of the data are uniform.

Category	GPT-4	GPT-4o-mini
Battery Life	0.92	0.84
Build Quality and Durability	0.61	0.60
Comfort and Fit	0.88	0.68
Features and Functionality	0.13	0.11
Sound Quality	0.63	0.36
Value for Money	0.53	0.63

Table 2: Cohen's kappa scores for LLM and Human annotated data across different categories

4 Analysis and Results

When comparing the precision scores, it is evident that both models demonstrate very similar precision when compared to human annotations. The scores show minimal variation among the models, indicating that the choice of model for annotation is not a significant factor—unless a small change in precision could greatly impact the user base. Overall, GPT-4 outperformed the other models in nearly all aspects. Precision can fluctuate with different models and can be optimized by providing more detailed annotation guidelines. This could be explored as a separate research project.

We now have the precision scores, but to assess reliability, we will consider Cohen's Kappa score. The Cohen's Kappa score for the 'Battery Life' and 'Comfort and Fit' labels is very high for the GPT-4 model annotations, indicating a strong inter-annotator agreement. This suggests that LLMs can serve effectively as annotators for this type of data.

There are differences in the scores of the two models, with one model exhibiting higher precision than the other. The Features and Functionality model has the lowest score, indicating that it cannot be reliably used for annotating these labels. The remaining labels fall within the weak to moderate agreement range, meaning they are not highly reliable for certain tasks. One possible way to achieve higher scores is to create more detailed annotation guidelines.

Once we achieve a reliable data score, LLMs can be used for annotation. For tasks like annotating Battery life labels, the current approach works well. However, for other labels, improving inter-annotator agreement is essential; otherwise, human annotations remain necessary. The project also demonstrates how these labels can train basic classifiers, with the code included in the project files.