Are large language models good annotators?

Jay Mohta*, Kenan Emir Ak*, Yan Xu, Mingwei Shen
Amazon.com
{jaymoht, kenanea, yanxuml, mingweis}@amazon.com

Abstract

Numerous Natural Language Processing (NLP) tasks require precisely labeled data to ensure effective model training and achieve optimal performance. However, data annotation is marked by substantial costs and time requirements, especially when requiring specialized domain expertise or annotating a large number of samples. In this study, we investigate the feasibility of employing large language models (LLMs) as replacements for human annotators. We assess the zero-shot performance of various LLMs of different sizes to determine their viability as substitutes. Furthermore, recognizing that human annotators have access to diverse modalities, we introduce an image-based modality using the BLIP-2 architecture [17] to evaluate LLM annotation performance. Among the tested LLMs, Vicuna-13b [34] demonstrates competitive performance across diverse tasks. To assess the potential for LLMs to replace human annotators, we train a supervised model using labels generated by LLMs and compare its performance with models trained using human-generated labels. However, our findings reveal that models trained with human labels consistently outperform those trained with LLM-generated labels. We also highlight the challenges faced by LLMs in multilingual settings, where their performance significantly diminishes for tasks in languages other than English.

1 Introduction

Machine Learning (ML) models are widely deployed for various applications like sentiment analysis [24], machine translation [26], paraphrase equivalence detection [9], and many more. In the traditional machine learning scenario used for deploying these models, ML practitioners typically begin by training the model on a designated training dataset, and subsequently employ a test dataset to assess the model’s performance. However, assembling a comprehensive set of labeled data is a time-consuming process that necessitates expertise in the relevant domain.

In order to avoid labeling such huge datasets, the NLP community started to explore the use of transfer learning to improve the few-shot performance of the models. Models like BERT [7], RoBERTa [18], T5 [22] are pretrained on a large corpus of unlabeled data and then fine-tuned on downstream tasks. These models performed state-of-the-art on various open-source benchmarks like GLUE [30] and SuperGLUE [31].

In a recent study conducted by Brown et al. [3], it was demonstrated that GPT-3, a massive language model with 175 billion parameters, exhibits zero-shot performance on par with state-of-the-art supervised fine-tuned models. However, the access to GPT-3 may be restricted. To overcome this limitation, the open-source community released numerous competitive language models such as LLama [28], Alpaca [27], and Vicuna [34]. These works demonstrated that they were able to achieve similar performance to that of ChatGPT on multiple benchmarks [34, 11]. Even though these models have state-of-the-art zero-shot performance, deploying these models is expensive due to their substantial size.

*Equal contribution
Given the impressive zero-shot performance of LLMs, we explore their potential as annotators in this study. Recent work by He et al. [13] and Ding et al. [8] have demonstrated use of GPT-3 as annotator with only text based inputs. However, a critical question remains: can open-source LLMs effectively serve as replacements for human-annotated data? Our aim is to understand whether we can leverage these open-source models to generate labels cost-effectively, enabling the training of smaller models. This approach can offer a dual advantage: it significantly reduces the cost associated with label generation and eliminates the need for time-consuming and costly human annotation. By distilling data generated by these models, we can effectively train smaller models.

In this study, we conduct benchmarking analyses of various open-source models to evaluate their effectiveness as annotators for our proprietary datasets, as well as for publicly available datasets such as Hateful Memes [15] and Multimodal IMDB (MM-IMDB) [2]. Specifically, we conduct benchmarking experiments involving the following models: Vicuna (13b, 7b) [34], LLama (13b, 7b) [28], InstructBLIP (13b, 7b) [6]. For evaluating the multi-lingual aspect of these large language models we test the performance on XNLI dataset [5].

Our investigation indicates that Vicuna 13b performs reasonably well on various annotation tasks compared to the other tested models. It also exhibits the unique capability to provide reasons for classification, a feature often absent in human-labeled data. Conversely, we observed these models encountering challenges in multilingual settings. Considering that public datasets and internally annotated datasets have image information we also study the impact of images on annotation performance via a multimodal InstructBLIP architecture. Our findings indicate that the usefulness of images varies depending on the specific task, with their effectiveness contingent on the nature of the task at hand. To assess the effectiveness of LLMs as annotators, we substitute human-annotated data with LLM-annotated data in supervised model training, resulting in a consistent decline in performance.

The remaining sections of this paper are structured as follows: In Section 2 we provide an overview of the background on LLMs and multimodal LLMs. Section 3 outlines our approach. Section 4 presents the key findings of this study. We conclude the paper in Section 5.

2 Background

2.1 Large Language Models

Brown et al. [3] showcased the remarkable capabilities of GPT-3, a large language model with 175 billion parameters. GPT-3 achieved zero-shot performance comparable to state-of-the-art fine-tuned models. Subsequent to this milestone, a wave of language models emerged, including PaLM [4], LLama [28], Falcon [21], and numerous others. These models have consistently demonstrated exceptional zero-shot performance, rivaling that of finely-tuned models across entire datasets.

Ouyang et al. [20] showed that while these language models can achieve impressive zero-shot performance on targeted benchmarks, they often encounter difficulties in accurately comprehending and adhering to human instructions when prompted. To address this challenge, they introduced a reinforcement learning framework with fine-tuning guided by human feedback, resulting in the development of the InstructGPT model. However, recent research by Zhou et al. [36] has demonstrated that conventional fine-tuning of models using chat instruct datasets can yield performance comparable to models fine-tuned with the reinforcement learning and human feedback-based approach.

With the release of LLama weights there has been various fine tuned versions of LLama models like Alpaca and Vicuna. It has been shown that by Zheng et al. [34] these models perform comparable to that ChatGPT and many larger models. In this work, we delve into the assessment of models zero-shot performance across various annotation tasks.

2.2 Multimodal Large Language Models

Human perception is inherently reliant on the integration of sensory inputs from various modalities, such as text, sound, and vision. In the field of multimodal learning, there have been various studies involving pretrained models trained on multiple types of data [19] [23] [16]. These models have demonstrated their effectiveness in various tasks, including visual question answering [35] and reasoning that combines both visual and textual information [25].
While large-scale pre-training and fine-tuning have been successful in creating general-purpose language models, building versatile vision-language models is challenging due to the diverse input distributions and tasks that arise from incorporating visual input. Recently introduced BLIP-2 model [17] has showcased impressive performance across a diverse range of tasks involving both images and language. The incorporation of the Query Transformer (Q-Former) component within this model has resulted in substantial enhancements in capturing interactions between text and image.

InstructBLIP delivers a comprehensive study on fine-tuning vision-language models, building upon pretrained BLIP-2 models. Through training on 13 various datasets, InstructBLIP achieves state-of-the-art performance on various vision-language tasks, surpassing the performance of BLIP-2 and larger Flamingo [1] models. Both BLIP-2 and InstructBLIP relies on a frozen LLM such as Vicuna and focuses on mapping visual information to LLM embeddings.

3 Approach

In this section, we outline our methodology for investigating the research question of whether large language models (LLMs) can effectively function as annotators. We provide a visual representation of the annotation process in Figure1 for both unimodal and multimodal scenarios.

3.1 Network

In this study, we perform benchmarking on various language models to evaluate their applicability for annotation purposes. Specifically, we assess several open-source LLMs, such as LLama (13b, 7b), Vicuna (13b, 7b), and OpenLLama 13b, for their ability to encode text-based information.

To incorporate the image modality for further comparison and leverage the additional modalities in the open-source and internally annotated datasets, we adopt InstructBLIP, which utilizes a Vicuna model. InstructBLIP adopts BLIP-2 architecture for its effective utilization of image features through Vision Transformer (ViT) [10] and Q-Former. The output of the Q-Former is subsequently mapped to the LLM’s embedding space and can be leveraged by the LLM for multimodal tasks.

For image encoder, we employ the ViT-G/14 [33], which uses $14 \times 14$ image patches, resulting in a feature dimension of $(HW/14^2 \times N)$, where $H$ and $W$ represent the height of the image, and $N$ is the dimension of the image features. The Q-Former, pretrained in representation and generative learning stages, enables the extraction of a consistent number of output features from the image encoder, regardless of the input image’s resolution. Q-Former module, takes input from both image and text embeddings and extracts the most relevant information from the image based on the accompanying text embedding. The output of the Q-Former is then mapped to LLM space which can be utilized by the LLM.
3.2 Inference & Network Details

During inference, we employed a generation configuration with specific settings. With a maximum token limit of 1024, we utilized the `do_sample` option, which allows the model to generate responses independently. We intentionally refrained from employing beam search to promote diversity in the generated outputs. To balance creativity and coherence in responses, we set the temperature parameter to 0.8, controlling the randomness of the generation process. Furthermore, we applied a top-p (nucleus) sampling technique with a threshold of 0.9, ensuring that the model focuses on the most probable tokens in each step of generation.

The Vicuna version utilized in InstructBLIP is an older iteration, specifically v1.1, in contrast to the latest release, Vicuna-v1.5. In Vicuna-v1.5, there has been a transition to the Llama2 as the base model, which is anticipated to significantly boost overall performance. Moreover, Vicuna-v1.5 has expanded the context length versions to 16K through linear RoPE scaling, effectively enabling the model to handle longer sequences.

For our multimodal experiments, we resize all images to a resolution of $224 \times 244$, while Q-Former extracts a token size of $32 \times 768$. Following this, image features are mapped to the language feature space using a linear layer, enabling the LLM to comprehend the context effectively.

To categorize the model’s outputs into different classes, we employed a regex-based matching following Dai et al. [6].

3.3 Supervised Model

To assess the substitutability of LLMs for human annotators, we train a well-established multimodal classifier known as MultiModal BiTransformers (MMBT) [14]. MMBT utilizes a BERT architecture for embedding multimodal data. In this framework, image embeddings, derived from a Residual Neural Network (ResNet) [12], are mapped into the text token space. The transformer, enriched with positional and modal-type embeddings, is designed to classify image-text pairs into distinct classes. By incorporating a supervised model, our objective is to facilitate a meaningful comparison with LLMs and underscore the distinctions that emerge when substituting human-annotated training data with LLM-generated annotations.

4 Experiments

4.1 Datasets

We employ multiple datasets specifically curated for multimodal classification tasks. These datasets are chosen to encompass a diverse range of domains and challenges, allowing us to thoroughly evaluate the performance of LLMs as annotators across different contexts.

**MM-IMDB.** The MM-IMDB dataset is designed for movie genre classification, relying on movie plots, poster images, and various metadata. This dataset poses a multi-label classification challenge, featuring 23 genre labels. For our annotation performance evaluation of Large Language Models (LLMs), we specifically focused on the top two genres from this dataset: drama and comedy. For our inference experiments, we utilize the development set, which comprises 1,946 image-text pairs.

**Hateful Memes.** The Hateful Memes dataset is primarily aimed at identifying hateful content within memes. This task involves binary classification, determining whether the content is hateful or not. Despite its binary nature, detecting hateful content can be challenging as most of the samples include sarcasm. To assess the performance of LLMs, we conducted experiments using the “test_unseen” set, which consists of 2,000 image-text pairs.

**Internally Annotated Datasets (IAD).** Our internally annotated datasets consists of two binary classification tasks, namely IAD-1 and IAD-2.

**XNLI.** The XNLI dataset is designed to evaluate the cross-lingual generalization capability of natural language understanding models. XNLI is an extension of the MultiNLI (MNLI) [32] dataset, which consists of sentence pairs labeled for three categories: entailment, contradiction, and neutral. We measure the annotation performance on three different languages French, Dutch and English.

---

2Hugging Face `do_sample`
4.2 Zero-shot experiments

In Figure 2, we present our results for each task. We include the performance of the supervised model to provide a reference point for interpreting the results generated by the LLMs. For internally annotated datasets, we consistently present results relative to a supervised baseline to adhere to regulations. As shown in Figure 2, Vicuna-7b-v1.5 exhibits a surprising outcome, surpassing the supervised model by approximately 2%. However, it can only annotate about 60% of the samples during inference. We would also like to highlight that these LLMs, especially Vicuna, exhibit the capability to generate reasons for classifications, a characteristic typically absent in human annotations.

Our findings reveal a noteworthy disparity in performance between Llama models and Vicuna models, with the latter essentially representing a fine-tuned iteration of the former. Our investigation suggests that the substantial improvement observed in Vicuna models compared to Llama models can be predominantly attributed to the fine-tuning process on the instruct dataset, as detailed in Zheng et al. [34]. Specifically, the refinement through fine-tuning empowers Llama models to align more closely with the provided instructions within the prompts, resulting in significantly improved accuracy and, consequently, more adept zero-shot annotation performance.

Intuitively, the inclusion of the image modality should yield a positive improvement in annotation tasks. However, when examining Figure 2a and Figure 2b, it’s evident that InstructBLIP encounters challenges in outperforming its unimodal counterparts, such as Vicuna-7b and Vicuna-13b. Nevertheless, it’s important to note that InstructBLIP has a significantly higher answer rate compared to other methods. Thus, it’s not straightforward to conclude that the addition of the image modality consistently enhances annotation performance. This observation may be attributed to the way the image modality is integrated, as InstructBLIP relies on mapping the image modality to the LLM’s embedding space, rather than incorporating pre-training with the image modality.
4.3 Comparing human-annotated and LLM-generated labels for supervised model training

In this section, we train the MMBT classifier as described in Section 3.3, using both human labels and LLM labels, and subsequently compare their performances. The LLM labels utilized for these evaluations are derived from the InstructBLIP-7b and InstructBLIP-Flan-t5-xxl for Hateful Memes and MM-IMDB, respectively due to their good performance and high answer percentage identified in Section 4.2. We convert the output of this model into labels for classification and employ them to train the MMBT classifier.

As shown in Table 1, there is a significant reduction in performance when transitioning from human labels to LLM-generated labels. This decline can be attributed to the inherent noise present in the labels generated by LLMs, leading to a degradation in performance. These findings underscore the continued necessity for human-labelled data in training machine learning models, indicating that the label annotation process cannot be entirely automated with the advent of LLMs.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model trained with human labels</th>
<th>Model trained with LLM labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hateful Memes</td>
<td>0.64</td>
<td>0.60</td>
</tr>
<tr>
<td>MM-IMDb</td>
<td>0.84</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 1: Comparison of models trained with human Labels and LLM Labels on different datasets.

4.4 Multi-lingual performance

In this section, we delve into the multi-lingual capabilities of the Vicuna-13b v1.5 model, recognized as the top performer for annotation tasks, as highlighted in section 4.2. As shown in Table 2, the model’s performance significantly diminishes when handling languages other than English. This observation can primarily be attributed to the fact that the model’s pre-training dataset is predominantly composed of English data, encompassing approximately 90% of the entire pre-training dataset Touvron et al. [29]. We believe that fine-tuning the model on these language-specific datasets may aid in improving its performance, but we defer such an analysis to future work.

<table>
<thead>
<tr>
<th>Task</th>
<th>Language</th>
<th>Accuracy</th>
<th>% of pretraining data</th>
</tr>
</thead>
<tbody>
<tr>
<td>XNLI</td>
<td>French</td>
<td>0.41</td>
<td>0.16%</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>0.42</td>
<td>0.17%</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>0.53</td>
<td>89.7%</td>
</tr>
</tbody>
</table>

Table 2: Vicuna-13b v1.5 language wise performance

5 Conclusion

In this study, we conducted a benchmark of various large language models, specifically focusing on their performance across diverse annotation tasks. Our primary goal was to explore the feasibility of utilizing annotations generated by LLMs for training smaller models, thereby reducing reliance on human label collection. The need for human annotations is often associated with significant costs, time constraints, and the necessity for domain expertise. Among the tested language models, Vicuna-13b gives good performance across a spectrum of downstream tasks. Recognizing that human annotators leverage different modalities, such as images, we incorporated these modalities into the annotation process through the BLIP-2 architecture. However, we observed that the utility of images varied depending on the specific task at hand. To assess the viability of LLM-generated labels, we substituted them for human labels and trained a supervised model and observed consistent decrease in performance. This underscores the current challenges and limitations in fully replacing human-generated labels with those generated by LLMs. We anticipate that our insights will empower machine learning practitioners to make informed decisions regarding the incorporation of LLMs as annotators. Furthermore, we hope that this work will stimulate further research to enhance the annotation performance of LLMs, ultimately bringing them closer to the quality of human-generated labels.
References


A  Prompts used for different datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hateful Memes</td>
<td>Is the text above hateful?</td>
</tr>
<tr>
<td>MM-IMDB</td>
<td>Identify the genre of this movie? Options: drama, comedy or both. Provide short answers.</td>
</tr>
</tbody>
</table>

Table 3: Prompts for Hateful Memes and MM-IMDB datasets.

B  Case studies

![Input Image](image1.png)

When two beautiful women fall for the unluckiest guy in the world his life takes on delicious complications. Identify the genre of this movie? Options: drama, comedy or both? Provide short answers.

Ground Truth

It's a comedy movie.

InstructBLIP-7b-v1.5

The genre of this film is likely a romantic comedy. The plot description suggests a lighthearted and humorous tone, with elements of romance and relationships between the three main characters.

Vicuna-13b-v1.5

This film could be classified as a romantic comedy. The genre typically combines elements of both romance and comedy, and the description you provided suggests that the film will likely focus on the relationship between the two women and the unlucky guy, with a mix of humor and drama.

InstructBLIP-13b

Both

Vicuna-7b-v1.5

Comedy

InstructBLIP-7b

Comedy

InstructBLIP-Flan-T5-11b

Comedy

Figure 3: This example shows that all models except InstructBLIP 13b responds correctly. We also observe that Vicuna models are able to provide justifications for classification as well.
Chris Hunter is a sly newsreel reporter. While in Shanghai doing reports on the Chinese-Japanese war, he meets pilot Alma Harding. At first she doesn't trust him, but by a trick he manages to get her hired as his assistant. During an adventurous expedition through the jungles of South America he manages to change her view of him. Identify the genre of this movie? Options: drama, comedy or both? Provide short answers.

Ground Truth: It's both comedy and drama.

This movie is a combination of both drama and adventure. It has elements of a romantic comedy, but the main focus is on the adventurous expedition through the jungles of South America.

Vicuna-13b-v1.5: The genre of this movie is drama and adventure.

Vicuna-7b-v1.5: The genre of this movie is drama and adventure.

InstructBLIP-13b: Drama

InstructBLIP-7b: Drama

InstructBLIP-Flan-T5-11b: Comedy

Figure 4: This example shows Vicuna-13b is only able to answer the question correctly with proper reasons while the other fails to provide correct label.