From Crowdsourcing to Large Multimodal Models: Toward Enhancing Image Data Annotation with GPT-4V

Owen He, Ansh Jain, Axel Adonai Rodriguez-Leon, Arnav Taduvayi, *Matthew Louis Mauriello

Department of Computer and Information Sciences, College of Engineering, University of Delaware
210 South College Avenue
Newark, Delaware 19716 USA
mlm@udel.edu

Abstract

This paper investigates the potential of Large Multimodal Models (LMMs), specifically GPT-4 with Vision (GPT-4V), to automate data labeling tasks often performed through crowdsourcing. Recent studies have evaluated the performance of Large Language Models (LLMs) for data annotation; however, there has been little study of the performance of LMMs for complex visual annotation. Our work compares the performance of a model trained on GPT-4V data with models trained on crowdsourced data from Amazon Mechanical Turk. We address two research questions: how might the performance of LMMs compare to crowdsourced workers in data labeling, and whether LLMs can modify input data to enhance model accuracy? Using a benchmark task involving detailed descriptions of human character models, we employ a random forest classifier to assess performance. Our results indicate that while GPT-4V offers promising capabilities, the modification of the input through the LMM yields marginal improvements, highlighting both the potential and limitations of automated data annotation using systems like GPT-4V.

Introduction

Machine Learning (ML) models rely heavily on vast amounts of accurately labeled data to perform effectively. Often this labeling process uses crowdsourcing, a method that leverages the input of numerous individuals to manually annotate data (Vaughan 2017). Recent advances in Large Multimodal Models (LMMs) have demonstrated their ability to perform accurately in many comprehension tasks, as noted in developed vision-language benchmarks (Yu et al. 2023; Li et al. 2024). Unlike Large Language Models (LLMs) which generate from text-based content, LMMs can infer and generate text from diverse inputs, including images, audio, and video.

GPT-4 with Vision (GPT-4V), developed by OpenAI, represents a significant leap in the capabilities of large-scale multimodal models (Roumeliotis and Tselikas 2023). Integrating both vision and language understanding, GPT-4V can process and interpret complex visual data, enabling it to perform tasks that would require a nuanced understanding of images in conjunction with text. LMMs enhances the potential for automated data annotation, a potentially efficient alternative to traditional crowdsourcing methods.

This study aims to evaluate GPT-4V's ability to accurately describe an image by comparing ML models trained using the generated text data to models trained on data generated via crowdsourcing where crowd workers completed the same image annotation task. Specifically, our work seeks to address two key research questions:

- RQ1: How might the performance of LMMs compare to crowdsourced workers in an image annotation task?
- RQ2: Can LLMs further modify input data to enhance model accuracy?

Related Works

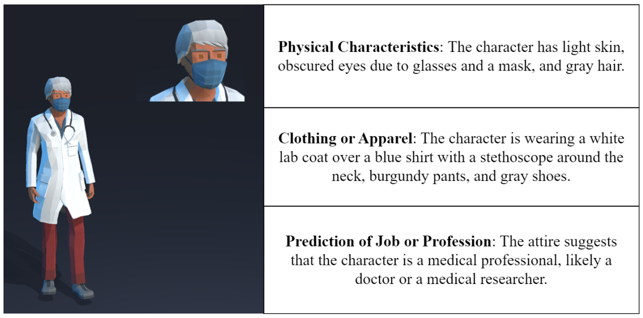
Data annotation through the use of LLMs has been extensively studied in recent years, primarily due to the increase in popularity of commercial models such as OpenAI's ChatGPT. LLMs such as GPT-4 have demonstrated greater data labeling accuracy than crowd workers in significant datasets (Gilardi, Alizadeh, and Kubli 2023; He et al. 2024). Furthermore, combining crowd-sourced data labels with LLM-generated labels may improve the overall accuracy of the models (He et al. 2024). Our work explores how LMMs might perform on a vision-based task to evaluate whether these performance observations hold true across tasks.

Methods

To conduct our study, we first establish a complex visual task that can be completed using both data generation methods (i.e., crowdsourcing and GPT-4V). Here, we describe the steps for the selection of the ML model used for our evaluation and how the LMM prompts were generated and used.

Benchmarking Task

The task for our study involved asking viewers to write detailed descriptions of a human character presented as a screenshot of a 3D model that encompasses a variety of characteristics. These included physical features such as height, skin color, and eye color, as well as other attributes like clothing and hair types. We used an off-the-shelf character creator for the Unity3D game engine, the Advanced People Pack 2 (Lenk 2020), to generate screenshots of characters used in our subsequent annotation tasks.



Physical Characteristics: The character has light skin, obscured eyes due to glasses and a mask, and gray hair.

Clothing or Apparel: The character is wearing a white lab coat over a blue shirt with a stethoscope around the neck, burgundy pants, and gray shoes.

Prediction of Job or Profession: The attire suggests that the character is a medical professional, likely a doctor or a medical researcher.

Figure 1: Character image paired with GPT-4V's annotations using prompts that approximate crowd instructions.

Crowd Workers

Crowd workers on Amazon Mechanical Turk (Turk 2012) were given an image of a randomly generated character model (Figure 1, left) and the following questions:

- *Q1: How might you describe the physical attributes (i.e., not their clothing or apparel) of the person displayed in the image?*
- *Q2: How might you describe the clothing and apparel of the person displayed in the image?*
- *Q3: How might you describe the look, profession, or bearing of the person displayed in the image?*

Though all questions were designed to be open-ended, Q1 and Q2 would provide more specific details about the character's visuals. Q3 sought additional descriptors, as more creative freedom was given based on their interpretation. To ensure the quality of the response, master qualifications and a 90% task approval rate were required to complete the annotation tasks. In addition, these descriptions were reviewed by another set of master-qualified crowd workers. Qualified workers could complete as many tasks as they liked. In total, 504 images were annotated and reviewed by 136 unique workers to create the final dataset used in our experiments.

Prompting

Due to the complex nature of this task, we attempted to follow established prompt engineering practices as described in previous literature to the best of our ability. This involved defining a clear goal, providing context, and refining through multiple iterations (Marvin et al. 2024; White et al. 2023). We finalized the following prompt by taking into account the necessity to maintain similarity between the prompt and the questions given to crowd workers.

> **Prompt:** *”You will be given an image of a character. Describe it in the following three categories, physical characteristics (skin color, eye color, hair color, height, and weight), clothing and apparel, prediction of job or profession. When describing the character, do not consider details about the artstyle of the image, the setting, or the character's posture. Do not add sub categories to the output.”*

Usage of a LLM with the output of the LMM was necessary to maintain consistent output in a usable format. In this case, we follow a generalized template format supported by existing literature (White et al. 2023):

> **Prompt:** *”I am going to provide you a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders listed. Preserve the formatting and overall template that I provide. This is the JSON template, {’Physical_Characteristics’: INPUT, ’Clothing’: INPUT, ’Job_Prediction: INPUT*”}

These two prompts were applied sequentially. The first prompt was given to GPT-4V and applied to images previously annotated by crowd workers. The second prompt was then given to the GPT-4 LLM with GPT-4V's output to format the data for downstream computation.

Model	Average Weighted F1 Score
Random Forest	0.769 (0.032)
Gradient Boosted Trees	0.705 (0.031)
Logistic Regression	0.652 (0.034)
Support Vector Machine	0.605 (0.033)
K-Nearest Neighbors	0.456 (0.038)
Multinominal Naive Bayes	0.429 (0.037)
Guassian Naive Bayes	0.320 (0.037)

Table 1: Average weighted F1 scores with standard deviations of various models trained in the shirt category

Classification Model

We evaluated several ML models to identify the best fit for this classification task (Table 1). The models used were created using an 80/20 split for training and testing, respectively, and evaluated over 100 randomized trials. To maintain relevance, categories such as shirt, pants, and shoes were trained using Q2 and Q3, while physical attributes such as height and skin color were trained using Q1. GPT provided similar responses after formatting and was trained similarly. We used the weighted F1 score, which ranges from 0 to 1, and measures a model's accuracy by considering both precision and recall to determine the best model. A score of 1 indicates perfect precision and recall, while 0 indicates no learning. As a result, we opted to use a random forest classifier since it provided the highest average cross-categories. We will use this model to evaluate RQ1.

Input modification

We also hypothesized that processing crowdsourced data through an LLM to remove irrelevant words and enhance its similarity to LMM-generated data would yield improved results when evaluated against a dataset generated by GPT (RQ2). Thus, we designed a third prompt and processed the crowd worker data (Q2 and Q3) through this prompt:

> **Prompt:** *”You are a helpful assistant that rewords descriptions of character models to more accurately represent a description generated by a large language model. Remove any details that do not fit in the categories of clothing and job prediction.”*

		Average Weighted F1 Score [Train/Test]				% Change	
Category	Classes	AMT/AMT	GPT/GPT	GPT/AMT	GPT/mAMT	AMT→GPT	AMT→mAMT
Shirt	15	0.769 (0.033)	**0.779** (0.028)	0.560 (0.043)	**0.596** (0.046)	1.332	6.537
Pants	10	0.535 (0.027)	**0.549** (0.017)	0.480 (0.016)	**0.487** (0.016)	2.673	1.551
Accessory	13	**0.387** (0.030)	0.380 (0.034)	**0.318** (0.034)	0.309 (0.034)	-1.766	-2.796
Hat	4	**0.689** (0.033)	0.643 (0.033)	**0.576** (0.037)	0.566 (0.032)	-6.614	-1.893
Hair	17	**0.060** (0.021)	0.053 (0.029)	**0.056** (0.021)	0.050 (0.020)	-11.444	-10.664
Shoes	14	0.267 (0.032)	**0.423** (0.038)	0.269 (0.030)	**0.283** (0.029)	58.110	5.152
Beard	11	**0.248** (0.027)	0.242 (0.025)	**0.224** (0.025)	0.215 (0.024)	-2.139	-4.296
Height	3	**0.412** (0.048)	0.339 (0.042)	0.333 (0.047)	-	-17.658	-
Neck Length	3	**0.345** (0.039)	0.335 (0.045)	0.314 (0.050)	-	-2.968	-
Head Size	3	**0.450** (0.047)	0.349 (0.043)	0.326 (0.039)	-	-22.479	-
Muscles	3	**0.333** (0.045)	0.312 (0.044)	0.302 (0.039)	-	-6.248	-
Thin	3	**0.315** (0.036)	0.311 (0.040)	0.335 (0.048)	-	-1.115	-
Fat	3	**0.408** (0.045)	0.349 (0.041)	0.331 (0.040)	-	-14.484	-
Hair Color	12	**0.293** (0.041)	0.227 (0.033)	0.219 (0.035)	-	-22.326	-
Eye Color	7	0.401 (0.043)	**0.416** (0.037)	0.375 (0.038)	-	3.827	-
Skin Color	9	**0.179** (0.033)	0.178 (0.032)	0.121 (0.028)	-	-0.559	-

Table 2: Average weighted F1 scores and standard deviations for each category. AMT → GPT refers to the % change between AMT/AMT and GPT/GPT, while AMT → mAMT refers to the % change between GPT/AMT and GPT/mAMT.

Results and Discussion

The F1 scores between AMT/AMT and GPT/GPT remained similar for most categories, the major differences being shoes, height, head size, and hair color. Crowdsourced data for depicting the character's physical characteristics resulted in better performance for height, head size, and hair color; converting to GPT-generated data resulted in a 17.66%, 22.48%, and 22.33% improvement, respectively. Meanwhile, there was a significant increase in F1 score in the shoe category for LMM-generated data, improving accuracy by 58.11%. This may signal a tendency for LMMs to be more descriptive in this aspect of clothing than humans when presented with an image. It should be noted that categories pertaining to physical traits, such as height and head size, have fewer unique classes compared to a category like shoe, which has 14. This draws some significance away from the large percentage change values in those categories as the F1 scores are low initially. When considering performance across categories, LMM generated data appears to struggle with physical attributes more than crowdsourced data but performs on par or better in apparel-related categories.

In modifying the AMT test data through GPT-4, we saw minor performance increases and decreases over the models trained with LMM data. In particular, the categories of shirt, pants, and shoes increased by 6.54%, 1.55%, and 5.152%, respectively. It is important to highlight that in the categories that decreased, we often deemed these as ”less visible” attributes; for example, it may be difficult to discern accessory types as differences between them are much less pronounced than in a category such as shirt.

Limitations and Future Work

As early work, this experiment provides insight into how LMMs may be used for complex image annotation; however, limitations are numerous and include dataset size and balance, as well as task complexity and number of subclasses. In addition, the task we evaluated was rather abstract and would not be applied to all image labeling tasks. Finally, we stopped some experiments early for cost and exploratory reasons. Thus, more work is needed before drawing additional conclusions about using LMMs for descriptive annotations, but these preliminary results highlight an interesting area for future study.

To facilitate a straightforward comparison, we maintained the same sample size for both AMT and GPT datasets, however, the cost for generating the AMT dataset was significantly greater than the GPT dataset. As a result, a more fair comparison may involve a greater sample size for the GPT dataset, allowing for more training data in creating those models. We urge future work in LMMs for complex annotation tasks to consider the following: a cost performance analysis involving datasets of different sizes with balanced classes, improvements in prompting to ensure greater similarity to the crowd sourced task, and potential rework of the task to apply to a broader image labeling perspective.

Acknowledgements

We note that Arnav Taduvayi contributed to this project while a student at Odyssey Charter High School, and Ansh Jain contributed to this project while a student at Caravel Academy and later the University of Maryland. We would like to thank Alina Christenbury, Emily Taylor, and Noah Hodgson for their contributions to the early engineering and data collection while students at the University of Delaware that led to this experiment. We also would like to thank the University of Delaware's Undergraduate Research Program for supporting Owen He and Axel Adonai Rodriguez-Leon while they contributed to this research.

References

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

He, Z.; Huang, C.-Y.; Ding, C.-K. C.; Rohatgi, S.; and Huang, T.-H. K. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM.

Lenk, A. 2020. Advanced People Pack 2 | Characters | Unity Asset Store.

Li, B.; Ge, Y.; Ge, Y.; Wang, G.; Wang, R.; Zhang, R.; and Shan, Y. 2024. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13299–13308.

Marvin, G.; Hellen Raudha, N.; Jjingo, D.; and Nakatumba-Nabende, J. 2024. *Prompt Engineering in Large Language Models*, 387–402. ISBN 978-981-99-7999-8.

Roumeliotis, K. I.; and Tselikas, N. D. 2023. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6): 192.

Turk, A. M. 2012. Amazon mechanical turk. *Retrieved August*, 17: 2012.

Vaughan, J. W. 2017. Making better use of the crowd: how crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.*, 18(1): 7026–7071.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. arXiv:2308.02490.