

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

Owen He, Axel Adonai Rodriguez-Leon, Arnav Taduvayi, and Matthew Louis Mauriello

1

Introduction

Training a model requires accurately labeled data

There are significant costs to generate a large amount of labeled data.

Large multimodal models (LMMs) are capable of analyzing and describing images.

2

Purpose & Research Question

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?

3

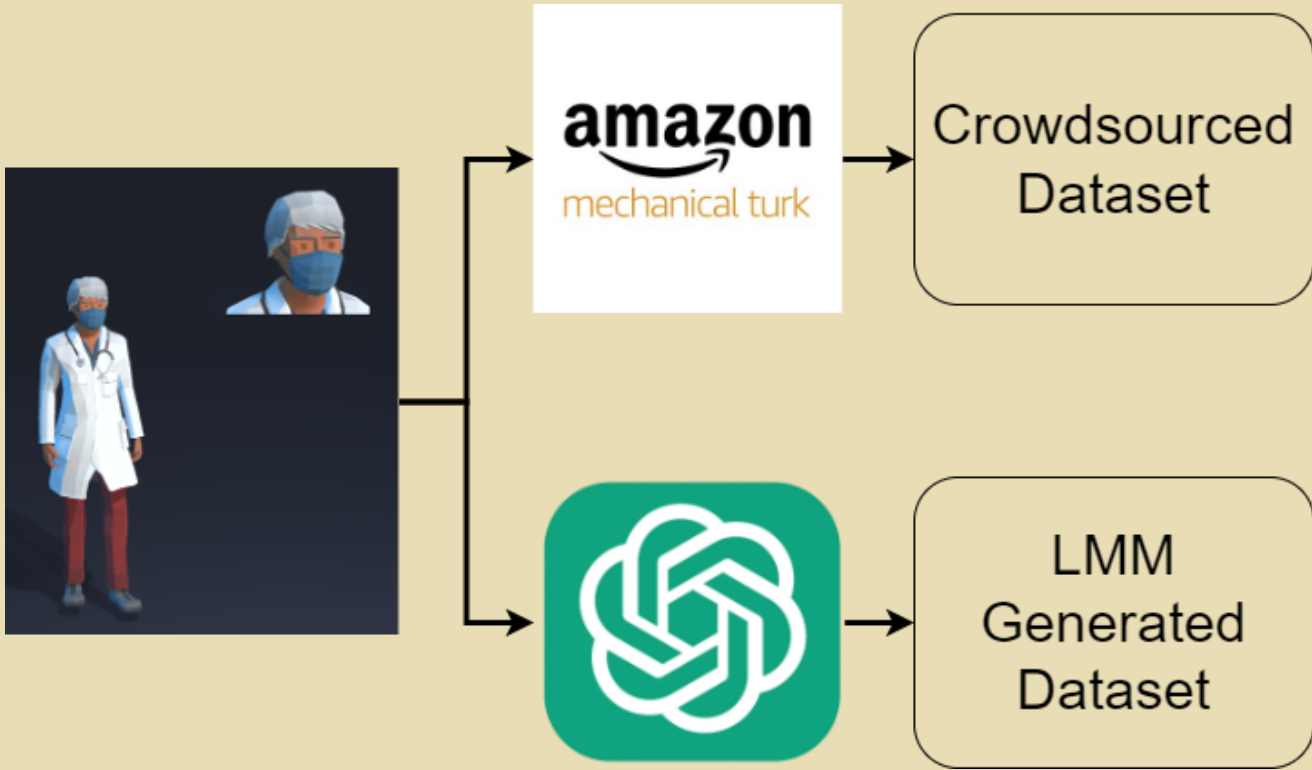
Methodology

Benchmarking Task:
Describing a screenshot of a randomly generated 3D character model

Crowdsourced workers were asked a question about the following:

- 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?

The image was processed through a LMM with a prompt similar to the questions in the crowdsourced group



Machine Learning Model:
Created using random forest classifier with the average weighted F1 score as our comparison primary metric. Trained models with varied test/train datasets with 80/20 split over 100 randomized trials

Input modification (RQ2):
Crowdsourced data was passed through a LLM (GPT-4) to test whether there was improvement in F1 score over the original crowdsourced dataset when used as the test split when training the model.

Dataset contains **500** randomly generated characters each with an AMT description, a LMM description, and a LLM modified AMT description.

Example of one character:

“The character is wearing a white lab coat over a blue shirt with a stethoscope around the neck, burgundy pants, and gray shoes. The attire suggests that the character is a medical professional, likely a doctor or a medical researcher.”

GPT-4V LMM description

“They are wearing a blue mask on their face and a white jacket with a stethoscope around their neck. They have on red pants. They have on gray shoes. This person is a doctor or veterinarian. They are possibly a surgeon due to the mask or maybe due to the pandemic they are wearing it. They seem professional and experienced.”

AMT Description

“They are wearing a blue mask on their face, a white jacket, and red pants. A stethoscope is around their neck, and they have on gray shoes. This person is likely a doctor or veterinarian, potentially indicating they are a surgeon.

AMT Description after being filtered by GPT-4 LLM

4

Results

Random Forest Model Metrics:

		Average Weighted F1 Scores of Train/Test Datasets (Std. Dev)				% Change	
Category	Number of 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
Shoes	14	0.267 (0.032)	0.423 (0.038)	0.269 (0.030)	0.283 (0.029)	58.110	5.152
Hair	17	0.060 (0.021)	0.053 (0.029)	0.056 (0.021)	0.050 (0.020)	-11.444	-10.664
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	-
Fat	3	0.408 (0.045)	0.349 (0.041)	0.331 (0.040)	-	-14.484	-

1. AMT→GPT refers to % change from AMT/AMT to GPT/GPT

2. AMT→mAMT refers to GPT/AMT to GPT/mAMT

- Significant increase in F1 score in the shoe category for LMM-generated data, improving score by 58.11%
- By converting the test set with an LLM categories of shirt, pants, and shoes increased by 6.54%, 1.55%, and 5.152%, respectively
- All physical attributes resulted in a decrease of F1 score when using the LMM dataset
- For attributes which increased F1 scores when using the LMM dataset, a LLM modified test split as opposed to the AMT test split increased as well.

5

Discussion/Future Work

Attributes that decreased in the conversion of the dataset from AMT to LMM were deemed as “less visible”

A fairer comparison might be to have a greater sample size for the LMM generated dataset due to cost differences

- A single LMM generated description = \$0.02
- An AMT generated description = \$2.00

The benchmarking task is complex and would not be representative of all image annotation tasks.

Website: <https://sensifylab.cis.udel.edu/>