Can GPT fairly judge vision-language instruction following models?

Rohan Wadhawan[1] **Tania Rajabally**[1] **Yashvi Mehta**[1] **Akshat Mehta**[1] **Shruthi Srinarasi**[1]

[1]Computer Science Department, UCLA {rwadhawan7,taniarajabally,yashvimehta, akshatmehta,shruthi223}@g.ucla.edu

Abstract

There exists a challenge of objectively determining the superior model when comparing two models. ChatGPT can be an effective judge for this, however, it has its biases. In this project, we aim to understand the specific prompts on which ChatGPT performs the task of choosing the better model effectively by conducting experiments. These experiments involve prompting ChatGPT with instructions consisting of text and OCR of images from different datasets and asking ChatGPT to generate a response about which model is better. Our analyses conclude that ChatGPT performs better when prompted with instructions and questions from the same dataset, however, its accuracy is dependent on the OCR generated in the dataset. Future works involve incorporating GPT-4 to reduce reliance on OCR for image instructions, potentially enhancing response accuracy. We make the code available at https://github.com/rohan598/Comsci-245-ST4

1 Introduction

In the age of multi-modal data, it has become increasingly important to process and understand visual data such as images and videos. There have been several developments in the artificial intelligence (AI) community in the space of multi-modal vision-language models or VLMs, which is able to process both visual and textual data simultaneously. The field of computer vision, especially, has witnessed the incredible impact of VLMs with the introduction of models like the ImageNet model in 2012. VLMs have far-reaching effects that go well beyond simply making image classification easier. These models have also found extensive use in other fields, such as caption generation for visual data and navigation assistance for autonomous vehicles.

To understand, quantify and compare the performance of these VLMs, there exist several different metrics. Popular metrics such as **ROUGE-L** and **BLEURT**, indeed, perform well, but primarily with measuring the lexical and semantic similarity between data.

With tasks that involve judging VLMs and comparing them to a given ground truth, this could prove to be inaccurate and pose a problem. For example, in Fig. 1., the given ground truth of the input image of a cat is "The photo shows a cat". On comparing the output between two VLMs, Model A and Model B, metrics like ROUGE-L and BLEURT seemingly choose the output that is more lexically and semantically similar to the ground truth rather than capturing the essence of the input data.

For judgement-based tasks, humans are undeniably the best possible judge. However, human-power is not only expensive but also not scalable. *What could be the closest alternative to humans that is more cost effective and more scalable?*

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

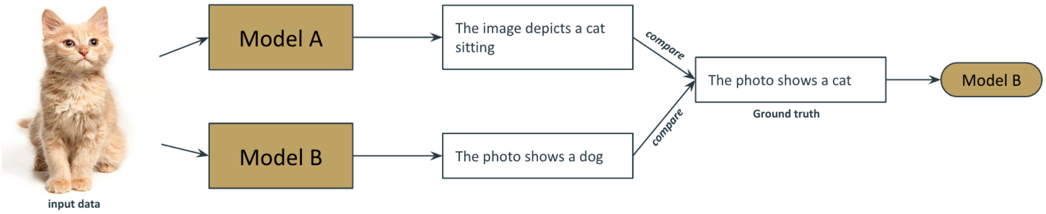


Figure 1: Judging two models only based on lexical and semantic similarity

With the introduction of ChatGPT, the world has seen a huge boom in the domain of large language models (LLMs) being utilized to receive human-like responses from machines. Hence, it could be possible that ChatGPT could replace humans for the task of model judgement. However, it has been observed that ChatGPT suffers from issues that could potentially disrupt fair judgement (e.g., primacy effect). Through this work, we hope to answer two main questions: *Is ChatGPT a specialized or generic judge?* and *What data is sufficient to make ChatGPT understand the premise of task?*.

It was observed that ChatGPT resulted in being a specialized judge rather than generic. ChatGPT also performed the best with one in-domain in-context example, illustrating that more information was not required for ChatGPT to understand the premise of a task.

The rest of the report is sectioned as follows: Section 2 focuses on VLMs, a few popular existing models and the tasks they can perform. Section 3 introduces the different related works and attempts to survey relevant literature. Section 4 details the experimental methodology and outlines the different components of the pipeline, and Section 5 describes the datasets used to test the working of ChatGPT as a judge. Section 6, 7, and 8 details the prompts generated to feed into ChatGPT, the metrics used for evaluation and the results of the experiments respectively. Finally Sections 9 and 10 outline the conclusions and potential future work.

2 Vision Language Models (VLMs)

VLMs refer to models that combine two modalities: vision and language. These models are able to process both text and image data at the same time. VLMs have gained a lot of popularity in recent times. The community has observed a surge in the number of VLMs, each built for different tasks. A few popular models include DALL-E [1], Vision-and-Language BERT (ViLBERT) [2], Guided Language to Image Diffusion for Generation and Editing (GLIDE) [3], and Contrastive Language-Image Pre-training (CLIP) [4].

DALL-E is a cutting-edge image generating model. It is an adaptation of the GPT architecture, which learns the statistical patterns in a language by pre-training it on a sizable corpus of text. Using a textual description as input, DALL-E creates an image that corresponds to the description. The encoder in the model extracts pertinent information from the text, while the decoder creates the image using a decoder-encoder architecture. **GLIDE**, a model based on DALL-E, is a diffusion model that is based on the idea of introducing randomized noise, sequentially.

ViLBERT is a model that can handle tasks requiring both textual and visual reasoning. Two images and written descriptions are used as the model's inputs. It is built using a transformer-based architecture which creates a combined representation of the text and the visuals that is utilized for a number of purposes, including visual entailment, visual reasoning, and visual question answering.

CLIP is able to identify images and the textual descriptions that correspond to them. The objective of the training process is to teach the model a joint representation of the image and the text by using a massive corpus of images and their textual descriptions. The combined representation is learned by CLIP using a transformer-based architecture and a contrastive loss function.

2.1 Vision-Language Tasks

Vision-language tasks are broadly classified into three different categories namely generation tasks, retrieval tasks and classification tasks.

Generation tasks consist of visual captioning (VC), question answering (VQA), commonsense reasoning (VCR) and visual generation (VG). VC refers to the generation of text captions or descriptions for a particular input image. VQA refers to the task of answering a question that is posed based on an input image. VCR refers to extracting the common-sense reasoning or understanding of a given visual input. Finally, VG refers to the generation of an image or any other visual output based on a given input text snippet.

Retrieval tasks include multi-modal machine translation (MMT), visual retrieval (VR) and vision-language navigation (VLN). MMT incorporates additional visual information while translating a description between languages. Images are retrieved via VR in response to a textual description. In VLN, an agent follows textual instructions to travel through a space.

Classification tasks comprise of two main subtasks: natural language for visual reasoning (NLVR) and multi-modal affective computing (MAC). NLVR checks the accuracy of a statement pertaining to a visual input. Similar to multi-modal sentiment analysis, MAC deduces visual affective activity from both inputs - visual and textual.

2.2 Visual Instruction Models

This work primarily focuses on the task of VQA and visual instruction models. To follow natural language instructions, several LLMs such as T5 and GPT-3 have undergone instruction-tuning, which has seen significant success. Extending this idea to computer vision, to improve zero- and few-shot generalization abilities, models such as OpenFlamingo [5], InstructBLIP [6], MiniGPT-4[7], and Large Language-and-Vision Assistant (LLaVA) [8] have been introduced.

To train autoregressive vision-language models, **OpenFlamingo** is an open-source version of DeepMind's Flamingo models. OpenFlamingo attempts to forecast the following text based on every text that has come before it as well as the most recent image. By attaching the dense cross attention modules to the layers of a frozen autoregressive language model, it becomes possible for the text tokens to focus on the associated images.

Employing the architecture of BLIP-2, **InstructBLIP** showed that instruction-tuned LLMs in conjunction with image inputs demonstrate significantly superior zero shot performance compared to multi-task networks or unimodal/multi-modal networks operating in the absence of instructions. InstructBLIP uses a frozen image encoder, frozen LLM, and a Q-former to extract informative visual features from the text.

MiniGPT-4 demonstrated strong abilities in image recognition. Only a small portion of the open-source language model Vicuna's parameters needed to be trained after the image encoder was integrated with it to create MiniGPT-4. MiniGPT-4 was made to work with ChatGPT, resulting in the creation of a high-quality data set of 3500 images and texts, which, after fine-tuning resulted in a more usable and computationally cheaper model.

Another unique model, **LLaVA** is a multi-modal assistant that combines the power of LLMs like GPT-4 and vision encoders like CLIP that comprehends and acts on multi-modal instruction.

3 Experimental Methodology

We have considered 2 models - Model A and Model B. Model A is mPLUG-Owl. mPLUG-Owl is a training paradigm that gives large language models (LLMs) multi-modal abilities. It involves a two-stage method for aligning image and text. mPLUG-Owl learns visual knowledge while maintaining and improving the generation abilities of LLM.

Model B is LLaVA v1.5. LLaVA stands for Large Language-and-Vision Assistant. It is an open-source project by Microsoft Research. LLaVA v1.5 represents the first end-to-end trained large multimodal model (LMM) that achieves impressive chat capabilities mimicking spirits of the multimodal GPT-4.

Our datasets consists of an instruction (question whose answer has to be derived from the image), the ground truth (actual expected response), Optical Character Recognition, Image Caption, Model A response, Model B response and the Response of which model performed better.

As shown in Figure 2, we give ChatGPT a prompt comprising of three parts. We first provide an explanation of what we expect ChatGPT to do with the information we are providing and how it is

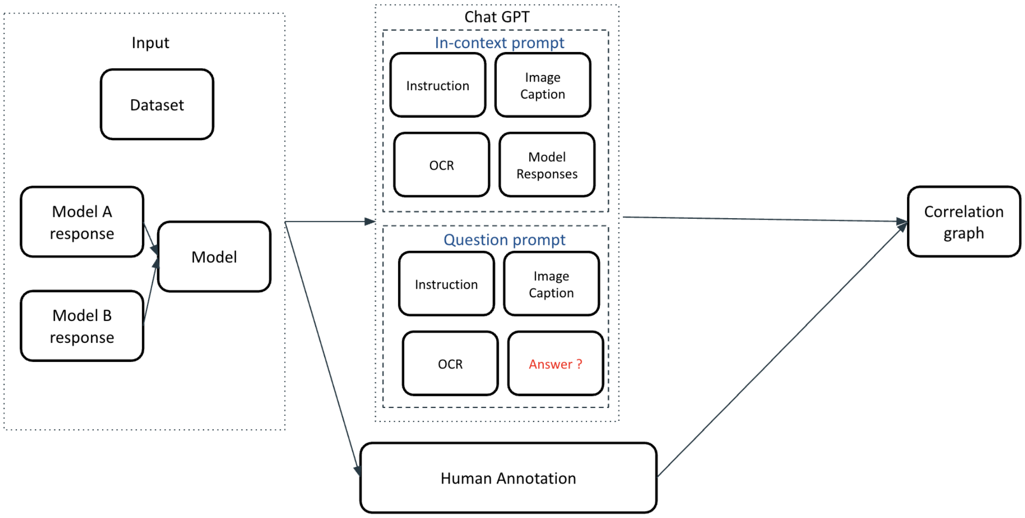


Figure 2: Experimental Methodology

supposed to analyse the image caption or OCR to get the expected response and compare it with the response of the models. We then provide incontext examples. An incontext example consists of an instruction i.e the question asked, the OCR and image caption of the image. Since these examples will be used to show ChatGPT what the expected output is, we even provided which model performed better in this case. We can vary the type of incontext examples depending to check how that affects the performance of ChatGPT. The last part of the prompt is the question prompt. We now provide all the same details as the incontext prompt but we expect ChatGPT to responds with the model which performed better. For this experimentation, we performed in domain as well as out of domain examples to check how this affected the performance.

4 Datasets

The experiment requires a kind of dataset which have concise, defined and extractive answers. This allows the correctness of the model responses to be binary. VQA datasets combine images with textual questions and answers, enabling models to learn the relationship between visual content and language. This multimodal aspect facilitates research and development in understanding the correlation between vision and language, enabling models to comprehend images in a more comprehensive manner. VQA datasets tend to be more diverse and complex in terms of the questions and answers associated with images. This diversity allows for a broader range of tasks, including reasoning, counting, spatial understanding, and more, compared to other datasets like COCO, which primarily focuses on object detection and segmentation. For these reasons, VQA dataset was used for our experiments.

We want to target three levels of difficulties in our experiments on the visual information capabilities of the two models being tested.

Level 1: Text sensing and extractive inferences from documents - Document data

Level 2: Questions on poster like images like with some text and some images - Infographic data

Level 3: Questions on images with little or no text, but contextual inferences - Scene test data

We have used 3 different datasets.

1. Infographic Visual Question Answering (InfoVQA)

Infographics convey information in a compact manner. The images contain textual and visual clues. They can comprise of graphs and charts along with textual information. The questions need to reason the document layout, understand the text, the graphs and graphical parts along with the data visualization. Figure 3 shows an example of a sample InfoVQA.

2. Document Visual Question Answering (DocVQA)

DocVQA consists of different types of documents containing information in different forms. The content of the document is extracted and used to respond to various questions. Figure 3 shows an example of DocVQA.

3. Scene Text Visual Question Answering (STVQA)

This dataset consists of high level semantic information that is present in the images. To answer the questions, it is important to be able to read the scene text depending on the question asked and the visual information provided. Figure 3 shows an example of STVQA.

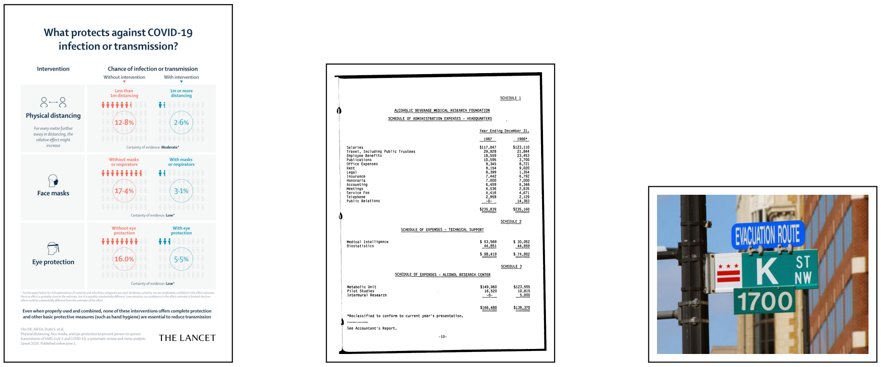
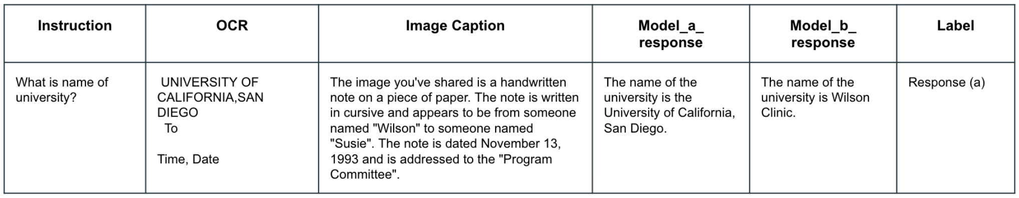


Figure 3: Examples of images from InfoVQA, DocVQA and STVQA datasets

Each of this datasets consisted of multiple columns:

- Instruction - The question whose answer we are looking for in these images
- Ground truth answer - This is the actual response to the question as interpreted from the corresponding image
- Optical Character Recognition of the Image - Since ChatGPT does not accept images as an input, we need to extract the information from the image which can then be provided as an input to ChatGPT. OCR only converts the text in the image.
- Image Captioning - Image captioning generates textual description of the image based on the visual content of the image from various perspectives.
- Model A answer - This captures the response obtained from Model A for the instruction.
- Model B answer - This captures the response obtained from Model B for the instruction.
- Response - Human annotation of which model performed better.

Figure 4 shows an example of a row in the dataset.



Instruction	OCR	Image Caption	Model_a_ response	Model_b_ response	Label
What is name of university?	UNIVERSITY OF CALIFORNIA,SAN DIEGO To Time, Date	The image you've shared is a handwritten note on a piece of paper. The note is written in cursive and appears to be from someone named "Wilson" to someone named "Susie". The note is dated November 13, 1993 and is addressed to the "Program Committee".	The name of the university is the University of California, San Diego.	The name of the university is Wilson Clinic.	Response (a)

Figure 4: Sample row in the dataset

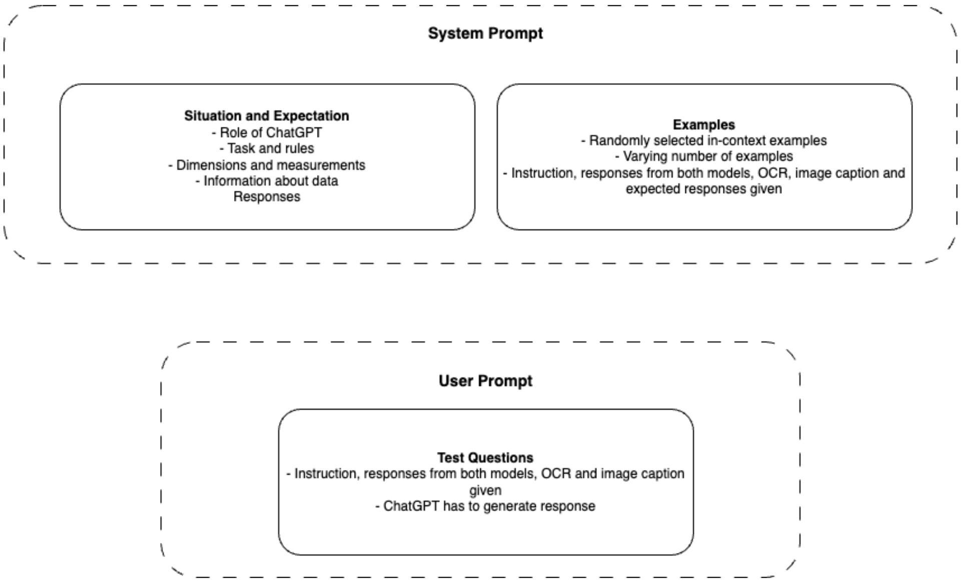

System Prompt
Situation and Expectation
- Role of ChatGPT
- Task and rules
- Dimensions and measurements
- Information about data
Responses
Examples
- Randomly selected in-context examples
- Varying number of examples
- Instruction, responses from both models, OCR, image caption and expected responses given
User Prompt
Test Questions
- Instruction, responses from both models, OCR and image caption given
- ChatGPT has to generate response

Figure 5: Prompts

5 Prompts

In the context of large language models, prompts refer to the input provided by users to generate responses. Concerning the project, defining the right prompts is necessary as it will help evaluate whether ChatGPT can be a fair judge. There are two parts to the parts - system prompts and user prompts. System prompts consist of task definition, that is, what is expected out of ChatGPT, what rules and parameters it needs to keep in mind while evaluating, and the varying number of in-context examples. The user prompt is the instruction whose response ChatGPT has to generate.

The task definition first has the response definition. The dimensions of measurements include accuracy, coherence, non-repetition, and factuality. The number of in-context variables can be varied from 0 to 3. It consists of the image caption, OCR of the image, instruction, response from model A, response from model B, and the expected response. For the user prompt, the same information is given except for the expected response which ChatGPT has to evaluate. It should be noted here that ground truth is not provided so that ChatGPT is not biased towards the ground truth and learns from the in-context example and generates the response.

2 types of prompt engineering have been carried out. They are in domain and out of domain. For the in-domain, both the in-context example and the question prompt are from the same dataset. However, in out-of-domain, both of them are chosen from different datasets, that is, if the in-context example is from docVQA dataset, then the question prompt is from STVQA or InfoVQA dataset. This prompting technique is used so that there is a clear understanding of whether ChatGPT is a specific judge or a generic judge.

6 Metrics

Spearman's correlation coefficient is used to calculate the correlation score between the human annotations (gold standard) and respective model results. Spearman's correlation coefficient is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where, ρ = Spearman's rank correlation coefficient
d_i = difference between the two ranks of each observation
n = number of observations

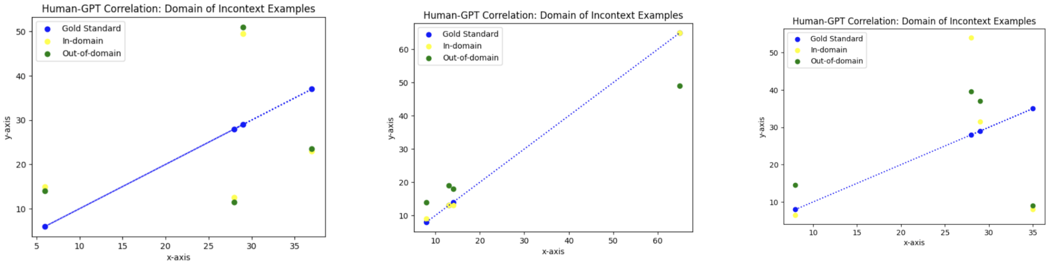


Figure 6: Gold standard - GPT output graph for 3 datasets with in-domain and out-of-domain incontext examples. *(Left: DocVQA, Center: InfoVQA, Right: STVQA)*

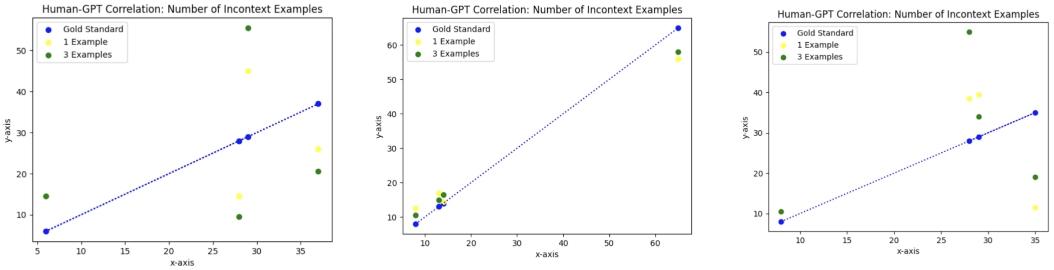


Figure 7: Gold standard - GPT output graph for 3 datasets with varying number of incontext examples. *(Left: DocVQA, Center: InfoVQA, Right: STVQA)*

7 Results

On conducting the experiments, there were two main observations:

1. GPT as a judge: As seen in Fig. 4., it is observed that with all three datasets, in-domain examples lead to a more optimal performance i.e., closer to ground truth. Through this, we infer that, although, GPT is intended to be a generic judge, in this case, it acts as a more specialized judge.

2. How much information is required to understand the premise of a task: As seen in Fig. 5., it is observed that one in-context example was sufficient for GPT to understand the premise of a task. Increasing the number of in-context examples did not improve the performance.

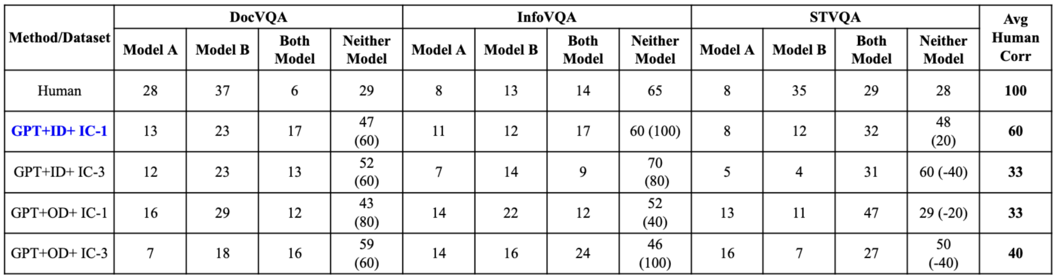
The same can also be observed in Table 1. The combination of GPT with in-domain examples and one in-context example outperform the other variations by at least 40%, with the average human correlation score being 60%.

With respective to the performance of both models and calculate the win ratio, we reward the outperforming model with 1 point for each outperformed example. If both or neither model outputs the correct answer, we reward both models with 0.5 points.

On calculating the rewards for both models at the end of 100 runs (100 data points), it was observed that ChatGPT shows a consistent win ratio of a model for the DocVQA and InfoVQA datasets. However, when it came to the STVQA dataset, the results were inconsistent i.e., ChatGPT outputted the wrong model more often. ß

8 Conclusions

From the results of the images, we can conclude that ChatGPT is a specific judge. ChatGPT needs to be prompted with very familiar and specific in-context examples to act as a more fair and reasonable judge of VLM models. While ChatGPT needs in-context examples to understand the task assigned to it, one example seems to be more than enough for it to work efficiently. Additional in-context examples are not incremental to its performance and would just put unnecessary load to the system.



Method/Dataset	**DocVQA**				**InfoVQA**				**STVQA**				**Avg Human Corr**
	Model A	**Model B**	**Both Model**	**Neither Model**	**Model A**	**Model B**	**Both Model**	**Neither Model**	**Model A**	**Model B**	**Both Model**	**Neither Model**	
Human	28	37	6	29	8	13	14	65	8	35	29	28	**100**
GPT+ID+ IC-1	13	23	17	47 (60)	11	12	17	60 (100)	8	12	32	48 (20)	**60**
GPT+ID+ IC-3	12	23	13	52 (60)	7	14	9	70 (80)	5	4	31	60 (-40)	**33**
GPT+OD+ IC-1	16	29	12	43 (80)	14	22	12	52 (40)	13	11	47	29 (-20)	**33**
GPT+OD+ IC-3	7	18	16	59 (60)	14	16	24	46 (100)	16	7	27	50 (-40)	**40**

Table 1: Average human correlation of GPT output with varied in-context examples

^{ID} in-domain examples, ^{OD} out-of-domain examples, ^{1} one in-context example, ^{3} three in-context examples

Since ChatGPT cannot take image input, OCR becomes very important for it to extract information. A better OCR format would substantially increase performance in judgement. Among the three datasets, ChatGPT has the best judgement accuracy between the two responses for InfoVQA. Since STVQA does not have a lot of text, there is no refined OCR to provide ChatGPT enough information to judge accurately. DocVQA has a decent OCR but might have more position, factual or abstractive questions which become tough for ChatGPT to judge. InfoVQA has a better balance of text and extractive tasks which ChatGPT can judge more accurately.

9 Future Work

The project currently tests only a small section of the dataset, but future plans include scaling up the testing to yield more accurate results. Further, the amount of visual information provided in the in-context examples can be varied to understand its dependency on it. While we have not delve into the issue of primacy effect in this paper, tests could be preformed to see if there is cognitive bias generated because of it. Lastly, since ChatGPT heavily relies on the OCR of the instruction to generate responses, GPT-4 can be inculcated which removes the need to provide the OCR of the image and delivers better results.

10 Literature Review/ Related Work

Understanding the landscape of research in vision language models requires an exploration of various dimensions, encompassing alignment with user intent, mitigation of biases, advancements in multimodal models, and innovative approaches in navigation tasks. This section contextualizes recent studies, underscoring their significance and interconnections.

[9] emphasizes the emergence of multimodal language models, integrating various data types such as images, text, and audio. Unlike traditional large language models primarily trained on textual data, multimodal LLMs, exemplified by GPT-4, encompass multiple data types, enhancing their understanding beyond text. These models exhibit human-level performance and offer extensive possibilities in high-value domains like multimodal robotics and document intelligence.

Efforts to align language models with user intent and enhance their behavior have been central to recent studies. [10] have investigated issues where language models generated untruthful or toxic content due to misaligned objectives. They emphasize the importance of refining language models with human feedback to mitigate such undesirable outputs. Notably, [10] introduces InstructGPT, a model fine-tuned with human feedback, aiming to improve the alignment of model behavior with desired outputs. This survey serves to bridge this gap by offering insights into social bias studies across these domains, presenting guidelines to mitigate biases in both unimodal and multimodal settings [11]. This approach underscores the significance of leveraging human-guided fine-tuning to steer language models towards intended outputs, thereby addressing concerns related to model behaviors not aligning with user expectations.

Efforts to address biases within machine learning models have gained substantial attention, particularly regarding social biases in training datasets. [12] investigates biases in transformer-based pre-trained models across NLP, CV, and VL domains [13]. While considerable studies have tackled biases in

NLP and CV individually, [12] underscores the limited focus on biases in VL models. This survey aims to bridge this gap by offering insights into cognitive bias studies across more domains, thereby contributing to the development of fairer AI models [14].

[15] pioneers a novel approach using language as a perceptual representation for vision-and-language navigation tasks. Unlike conventional methods relying on continuous visual features, this approach employs language descriptions derived from an agent's egocentric views for navigation. It explores scenarios involving GPT-4 prompts to synthesize trajectories and facilitate sim-to-real transfer, demonstrating the potential of language as a navigational perceptual space in low-data regimes. While existing works focus on language models for sub-task predictions based on world knowledge, [15] presents language's efficacy as a perceptual representation in navigation tasks.

Drawing from strategies for model alignment, bias mitigation, and innovative language-based navigation introduced in these papers, our study aims to investigate ChatGPT's judgment capabilities in evaluating vision-language instruction following models. Through experiments, we seek to discern the specialization of ChatGPT's judgment and ascertain the requisite data volume for effective comprehension of task premises in vision-language instruction following scenarios.

References

[1] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). *Zero-shot text-to-image generation.* In International Conference on Machine Learning (pp. 8821-8831). PMLR.

[2] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). *Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.* Advances in neural information processing systems, 32.

[3] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). *Glide: Towards photorealistic image generation and editing with text-guided diffusion models.* arXiv preprint arXiv:2112.10741.

[4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). *Learning transferable visual models from natural language supervision.* In International conference on machine learning (pp. 8748-8763). PMLR.

[5] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., ... & Schmidt, L. (2023). *Openflamingo: An open-source framework for training large autoregressive vision-language models.* arXiv preprint arXiv:2308.01390.

[6] Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., ... & Hoi, S. (2023). *InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning.* arXiv preprint arXiv:2305.06500.

[7] Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). *Minigpt-4: Enhancing vision-language understanding with advanced large language models.* arXiv preprint arXiv:2304.10592.

[8] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual instruction tuning.* arXiv preprint arXiv:2304.08485.

[9] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, Philip S. Yu (2023) *Multimodal Large Language Models: A Survey.* arXiv (Cornell University). https://arxiv.org/pdf/2311.13165.pdf

[10] Ouyang, L., Wu, J., Xu, J., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., John, S., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. J. (2022). *Training language models to follow instructions with human feedback.* arXiv (Cornell University). https://doi.org/10.48550/arxiv.2203.02155

[11] Chan, A. (2022). *GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry.* AI And Ethics, 3(1), 53–64. https://doi.org/10.1007/s43681-022-00148-6

[12] Nayeon L., Yejin B., Holy L., Samuel C., Wenliang D., and Pascale F. (2023) *Survey of Social Bias in Vision-Language Models.* arXiv (Cornell University). https://arxiv.org/pdf/2309.14381.pdf

[13] Liang, P. P., Wu, C., Morency, L., & Salakhutdinov, R. (2021). *Towards understanding and mitigating social biases in language models.* International Conference on Machine Learning, 6565–6576. http://proceedings.mlr.press/v139/liang21a/liang21a.pdf

[14] Xu, P., Zhu, X., & Clifton, D. A. (2023). *Multimodal Learning with Transformers: a survey.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–20. https://doi.org/10.1109/tpami.2023.3275156

[15] Pan, B., Panda, R., Jin, S., Feris, R., Oliva, A., Isola, P., & Kim, Y. (2023). *LangNav: Language as a perceptual representation for navigation.* arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.07889