LLaVAR: Enhanced Visual Instruction Tuning for Text-Rich Image Understanding

Yanzhe Zhang[1]*, **Ruiyi Zhang**[2], **Jiuxiang Gu**[2], **Yufan Zhou**[2], **Nedim Lipka**[2], **Diyi Yang**[3], **Tong Sun**[2]
[1]Georgia Tech, [2]Adobe Research, [3]Stanford University

Abstract

Instruction tuning unlocks the superior capability of Large Language Models (LLM) to interact with humans. Furthermore, recent instruction-following datasets include images as visual inputs, collecting responses for image-based instructions. However, visual instruction-tuned models cannot comprehend textual details within images well. This work enhances the current visual instruction tuning pipeline with text-rich images (e.g., movie posters, book covers, etc.). Specifically, we first use publicly available OCR tools to collect results on 422K text-rich images from the LAION dataset. Moreover, we prompt text-only GPT-4 with recognized texts and image captions to generate 16K conversations, each containing question-answer pairs for text-rich images. By combining our collected data with previous multi-modal instruction-following data, our model, **LLaVAR**, substantially improves the LLaVA model's capability on text-based VQA datasets (up to 20% accuracy improvement) while achieving an accuracy of 91.42% on ScienceQA. The GPT-4-based instruction-following evaluation also demonstrates the improvement of our model on both natural images and text-rich images. Through qualitative analysis, LLaVAR shows promising interaction (e.g., reasoning, writing, and elaboration) skills with humans based on the latest real-world online content that combines text and images. We make our code/data/models publicly available at `https://llavar.github.io/`.

1 Introduction

Instruction tuning [1, 2] improves generalization to unseen tasks by formulating various tasks into instructions. Such open-ended question-answering capability fosters the recent chatbot boom since ChatGPT [2]. Recently, visual instruction-tuned models [3–5] further augment conversation agents with visual encoders such as CLIP-ViT [6, 7], enabling human-agent interaction based on images. However, possibly due to the dominance of natural images in training data (e.g., Conceptual Captions [8] and COCO [9]), they struggle with understanding texts within images [10]. However, textual understanding is integral to humans' daily visual perception.

Fortunately, recognizing texts from images is accessible based on OCR tools. One naive way to utilize this is adding recognized texts to the input of visual instruction-tuned models [11], which increases the computation (longer context lengths) without fully leveraging the encoding capability of visual encoders. To this end, we propose to enhance the visual instruction-tuned model end-to-end by collecting instruction-following data that requires an understanding of texts within images.

*Collaborations through Adobe University Gift Program.
[2]`https://openai.com/chatgpt`

Preprint. Work in progress.

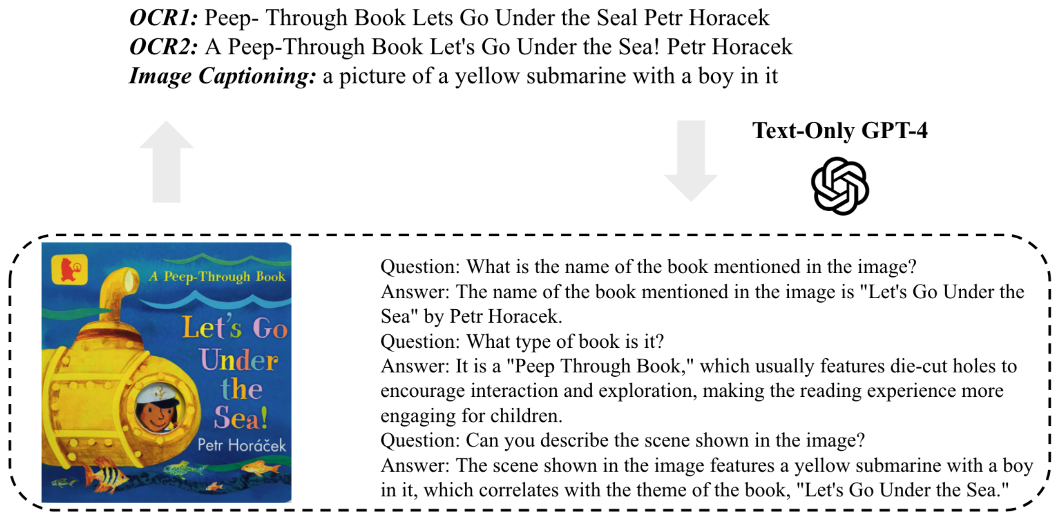


Figure 1: The process of collecting high-quality instruction-following data.

Specifically, we first collect 422K noisy instruction-following data using text-rich[3] images by combining manually written instructions (e.g., "Identify any text visible in the image provided.") and the OCR results. Such large-scale noisy-aligned data effectively improve the feature alignment between the visual features and the language decoder. Furthermore, we prompt text-only GPT-4 [12] with OCR results and image captions to generate 16K conversations, where each conversation can be multiple turns of question&answer pairs, as high-quality instruction-following examples. This process requires GPT-4 to denoise the OCR results and develop specific questions to create complex instructions based on the input (Figure 1).

To evaluate the effectiveness of collected data, we use noisy and high-quality examples to augment the pretraining and finetuning stages of LLaVA accordingly. We name our model **LLaVAR**, signifying the LLaVA (Large Language and Vision Assistant) that can **R**ead. Compared to the original LLaVA, we also experiment with scaling the input resolution from 224^2 to 336^2 to encode small textual details better. Empirically, we report the results on four text-based VQA datasets following the evaluation protocol from [10] together with the finetuning results on ScienceQA. Moreover, we apply GPT-4-based instruction-following evaluation on 30 natural images from COCO [9, 3] and 50 text-rich images from LAION [13]. Furthermore, we also provide the qualitative analysis (e.g., on posters, website screenshots, and tweets) to test more complex instruction-following skills. To sum up, our contributions are:

- We collect 422K noisy instruction-following data and 16K high-quality instruction-following data. Both are shown to be effective in augmenting visual instruction tuning.
- Our model, LLaVAR, significantly enhances text understanding within images while slightly improving the model's performance on natural images.
- The enhanced capability enables our model to provide end-to-end interactions based on various forms of online content that combine text and images.
- We open-source the training and evaluation data together with the model checkpoints.

2 Related Work

Instruction Tuning Following natural language instructions is the key capability for an agent to interact with real-world users. Instruction tuning starts from collecting human-preferred feedback for human written instructions [1] or formulating multi-task training in a multi-task instruction-following manner [2, 14]. However, large, capable instruction-tuned models are usually close-sourced and serve

[3]In this work, we use the phrase "text-rich images" to describe images that have text in them. In contrast, we refer to images without text as "natural images."

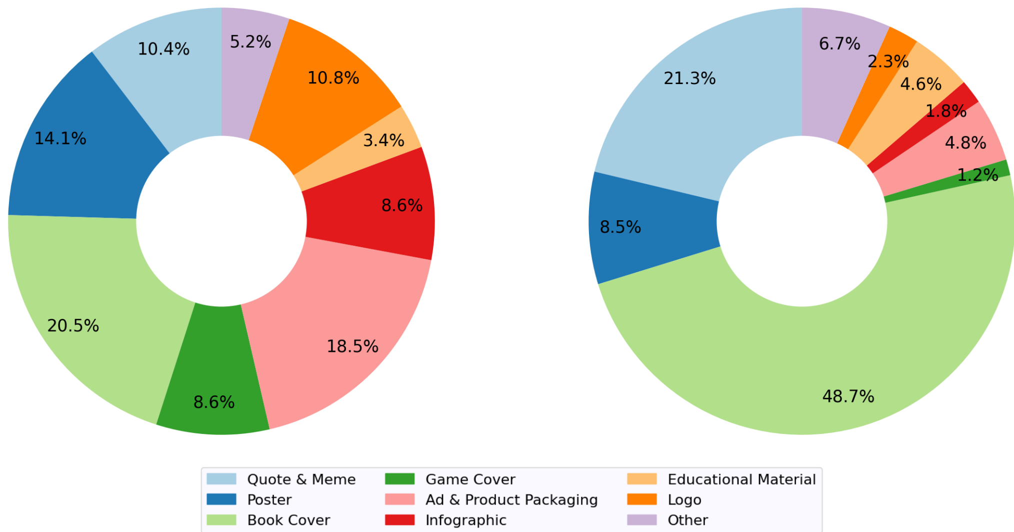


Figure 2: CLIP-based categorization of our collected images. The left one refers to images used to collect noisy data, and the right one refers to images used in GPT-4 prompting. Both pie charts are based on 10K sampled images from the corresponding datasets.

as commercial APIs only. Recently, Alpaca [15, 16], Vicuna [17], and Baize [18] start the trend of generating high-quality instruction-following data based on LLMs such as GPT-3.5/ChatGPT/GPT-4 and finetuning the open-sourced LLaMA model [19]. However, the evaluation of instruction-following capability remains challenging. While GPT-4 has demonstrated superior evaluation capabilities [20], it still has apparent drawbacks, including biases toward response length[18] and lack of robustness to the order or examples [21]. Following [17, 3, 22], we use GPT-4-based instruction-following evaluation in this work.

Multimodal Instruction Tuning Recently, instruction tuning has been expanded to the multimodal setting, including image, video [23, 24], and audio [25, 26]. In particular, for image-based instruction tuning, MiniGPT4 [27] employs ChatGPT to curate and improve the detailed captions for high-quality instruction-following data. LLaVA [3] and generates multimodal instruction-following data by prompting text-only GPT-4 with captions and object's bounding boxes. LLaMA-Adapter [28, 11] uses COCO data for text-image feature alignment and utilizes textual data only for instruction tuning. mPLUG-owl [29] combines more than 1000M image-text pairs for pretraining and a 400K mixture of text-only/multimodal instruction-following data for fine-tuning. However, according to [10], most of these models struggle with accomplishing tasks that require OCR capability. InstructBLIP [30] transforms 13 vision-language tasks (including OCR-VQA [31]) into the instruction-following format for instruction tuning. Cream [32] applies multi-task learning that includes predicting masked texts in images. In this work, we select LLaVA as our baseline, which is the most data-efficient and powerful model, and demonstrate the effectiveness of our proposed pipeline.

3 Data Collection

Starting from the LAION-5B [13] dataset [4], our goal is only to keep images that are text-rich. Considering documents usually contain plenty of text, we first obtained a binary classification dataset by combining natural images and document data. Subsequently, we trained an image classifier using a DiT [33] base backbone, which was fine-tuned on the RVL-CDIP dataset [34]. Hopefully, such a classifier can predict whether an image contains text or not. We first build a subset by selecting images with a predicted probability greater than 0.8 while also satisfying $p(\text{watermark}) < 0.8$ and $p(\text{unsafe}) < 0.5$ [5]. The derived subset is noisy due to the limitation of the classifier. To further

[4]`https://huggingface.co/datasets/laion/laion-high-resolution`

[5]Both probabilities are from the LAION dataset's metadata.

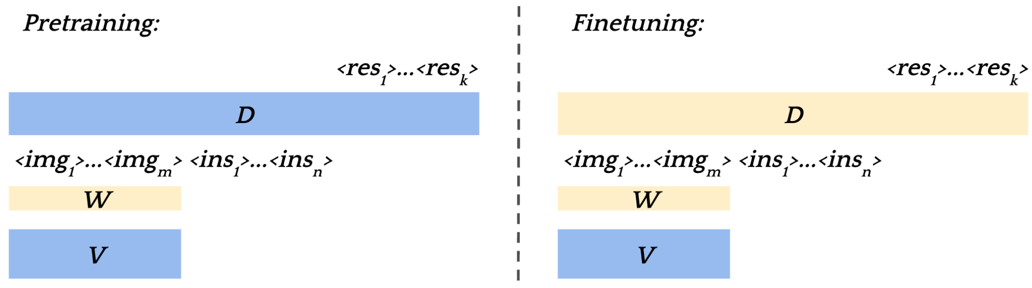


Figure 3: The model training process for visual encoder V, projection matrix W, and language decoder D. ***Blue blocks*** denote frozen modules, and ***yellow blocks*** denote trainable modules. The training input is image tokens (<*img*>) and instruction tokens (<*ins*>), while the target is response tokens (<*res*>).

Data	Image	Instruction	# Conv	Avg Ins Len	Avg Res Len
LLaVA pretraining	CC3M	CC3M	595K	15.9	15.4
$\text{R}_{\text{pretraining}}$	LAION	PaddleOCR	422K	17.2	48.8
LLaVA finetuning	COCO	GPT-4	158K	15.9	93.1
$\text{R}_{\text{finetuning}}$	LAION	GPT-4	16K	15.1	40.5

Table 1: Summary of data statistics. $\text{R}_{\text{pretraining}}$ and $\text{R}_{\text{finetuning}}$ denote the extra pretraining/finetuning data we collected. Average instruction and response length are calculated after LLaMA tokenization.

clean up the data and incorporate human judgment, we randomly sampled 50K images and clustered them into 100 clusters based on `CLIP-ViT-B/32` visual features. After inspecting the clustering results, we carefully select 14 clusters (See Figure 8 in Appendix for examples.) containing diverse text-rich images ranging from posters, covers, advertisements, infographics, educational materials, and logos. As a reference, we provide a CLIP [7]-based categorization (See Appendix A for details.) to illustrate the distribution of used images for both two types of data we collected in Figure 2. We also summarize and compare our collected data with LLaVA's data in Table 1.

Noisy Instruction-following Data Using the clustering model as the classifier, we collect 422K images that belong to the 14 preferred clusters. To balance the examples from different categories, we keep at most 52K examples for one cluster. We run all images through PaddleOCR [6]. Note that running OCR on the original resolution (e.g.,1024^2) might recognize small fonts that are not visible by visual encoders like CLIP ViT [6, 7] (up to 336^2). To ensure the recognition of visible fonts while maintaining OCR accuracy, we perform OCR on the resized image (the short edge is resized to 384 pixels) to extract the text. Then, based on the geometric relationships between the recognized words, we apply specific rules[7] to merge the words and obtain a text paragraph. As a robust instruction-following model should react similarly to instructions with similar meanings, we reword "Identify any text visible in the image provided." into ten distinct instructions (Table 7 in Appendix). We then create a single-turn conversation for a given image by (i) randomly sampling an ***input instruction*** and (ii) using the recognized texts as the desired ***output response***. Such instruction-following data is noisy due to the relatively limited performance of OCR tools on diverse fonts and colorful backgrounds.

GPT-4-based Instruction-following Data Compared to high-quality instruction-following data, there are mainly two issues for the noisy data collected above. (i) The responses should contain organized sentences instead of raw OCR results with missing words and grammar errors. (ii) The instructions should be diverse, suitable, and specific to the given image instead of monotonously asking for all visible texts. To address these issues, we follow [3] to generate instruction-following data by prompting text-only GPT-4 [12] with OCR results and captions.

[6] `https://github.com/PaddlePaddle/PaddleOCR`

[7] `https://github.com/JaidedAI/EasyOCR/blob/f454d5a85d4a57bb17082c788084ccc64f1f7397/easyocr/utils.py#L643-L709`

	Res	ST-VQA	OCR-VQA	TextVQA	DocVQA
BLIP-2 [35] †	224^2	21.7	30.7	32.2	4.9
OpenFlamingo [36] †		19.3	27.8	29.1	5.1
MiniGPT4 [27] †		14.0	11.5	18.7	3.0
LLaVA [3] †		22.1	11.4	28.9	4.5
mPLUG-Owl [29] †		29.3	28.6	40.3	6.9
LLaVA ‡	224^2	24.3	10.8	31.0	5.2
LLaVAR		30.2 (+5.9)	23.4 (+12.6)	39.5 (+8.5)	6.2 (+1.0)
LLaVA ‡	336^2	28.9	11.0	36.7	6.9
LLaVAR		39.2 (+10.3)	23.8 (+12.8)	48.5 (+11.8)	11.6 (+4.7)

Table 2: Results (accuracy %) on text-based VQA. We use † to refer to results fetched from [10] and ‡ to refer to our reproduced results. The accuracy metric used by [10] only counts for whether the ground truth appears in the response.

	ST-VQA	OCR-VQA	TextVQA	DocVQA
(1) LLaVA	28.9	11.0	36.7	6.9
(2) LLaVA + $R_{pretraining}$	36.7	26.1	46.5	9.6
(3) LLaVA + $R_{finetuning}$	34.1	21.6	43.6	9.5
(4) LLaVA + $C_{pretraining}$	35.4	27.0	45.6	9.2
(5) LLaVA + $N_{finetuning}$	34.1	25.9	43.3	10.2
(6) LLaVAR	39.2	23.8	48.5	11.6

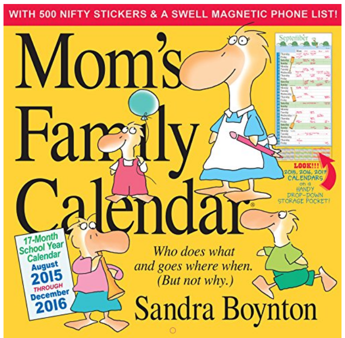
Table 3: Ablation Study on text-based VQA. All results are from 336^2-based models. $R_{pretraining}$ and $R_{finetuning}$ denote the extra pretraining/finetuning data we collected. $C_{pretraining}$ refers to using captions instead of OCR results as responses during pretraining. $N_{finetuning}$ refers to using written questions + raw OCR results instead of GPT-generated QA for finetuning.

It is challenging to prompt GPT-4 with fragmented OCR results with a few words to generate nontrivial instructions. To this end, we carefully select 4 out of the previously mentioned 14 clusters (The 3rd, 4th, 6th, and 9th clusters in Figure 8) to collect images with enough visible and coherent sentences. As shown in Figure 2, such filtering dramatically increases the percentage of book covers and quote images. We randomly select 4K examples from each cluster (no overlap with images used for noisy instruction-following data), yielding a total of 16K images. Following prior work [15, 16, 3], we provide the visualization of verb-noun pairs for instructions generated by GPT-4 in Appendix Figure 10. For those instructions with no verb-noun pair, we demonstrate the frequency of objects being asked in Appendix Figure 9.

Furthermore, based on the system message and two in-context few-shot examples ([37], shown in the Appendix B), we ask GPT-4 to generate conversational data based on OCR results and image captions (Figure 1). The generated questions are used as ***input instructions***, and answers are used as ***output responses***. Concretely, for a given image, we first provide two OCR results from EasyOCR and PaddleOCR, which can complement each other. To illustrate the visual elements other than texts within the image, we also provide the image captioning result. To prevent the caption from focusing on the text, we use OCR bounding boxes to mask the text and then use the inpainting to refill in the mask before using generation captions with BLIP-2 [35]. Note that the generated captions sometimes contain hallucination, which could come from the training data of the captioning model or the "fuzzy" shapes created by masking/inpainting. We leave generating more detailed and knowledge-enhanced captions [38] for future work.

4 Model Architecture and Training

Architecture We use the same model architecture as LLaVA. For the visual encoder V, we use `CLIP-ViT-L/14` for 224^2 resolution and `CLIP-ViT-L/14-336` for 336^2 resolution. The grid



Question: Who wrote this book?	
GT	Sandra Boynton
(1)	The book was written by Sandra Boynton.
(2)	The book was written by Sandra Boyton. It is a children's book that features a family calendar with an accompanying activity guide. The book provides information (...) The cover of the book displays a picture of the family from the calendar, which helps visually represent the topic of the book.
(3)	The book is written by Sandra Byington.
(4)	The book was written by Sandra Boynton.
(5)	With 500+ Nity StickerS &aSwEllMagnetic Phone List! Moms 291573016171800 FAMII (Pocket) Calendar (...) 2016 SANDRA BOYNTON
(6)	The book was written by Sandra Boynton.

Figure 4: Ablation study based an example from OCR-VQA. GT refers to ground truth, and (1) - (6) are different model variants from Table 3. We replace the excessively long response with (...).

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Human [39]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [39]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [39]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [28]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [40]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [40]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
LLaVA [3]	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4 [3] (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	92.53
Chameleon (GPT-4) [41]	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72	86.54
LLaVAR	91.79	93.81	88.73	90.57	88.70	91.57	91.30	91.63	91.42

Table 4: Results (accuracy %) on Science QA dataset. All baseline results are from [3, 41]. The categories are denoted as NAT: natural science, SOC: social science, LAN: language science, TXT: text context, IMG: image context, NO: no context, G1-6: grades 1-6, G7-12: grades 7-12.

features before the last Transformer layer are then transformed into the word embedding space of the language decoder through a trainable projection matrix W. Vicuna-13B [17], a LLaMA-based [19] instruction-tuned language model, is used as the language decoder D.

Training Similarly, we follow the two-stage training design of LLaVA (Figure 3). The training objectives of both stages are the same: generate ***output responses*** (<*res*>) for the ***input instructions*** (<*ins*>). The transformed image tokens (<*img*>) are added either before or after the first input instruction. (i) During the first pretraining stage, only the projection matrix W is trained for feature alignment. Since the decoder D is frozen, the training tolerates noisy data. We combine the 595K pretraining data from LLaVA with our 422K noisy instruction-following data in the pretraining stage. (ii) Both the projection matrix W and the language decoder D are trained during the finetuning stage, where we merge our 16K instruction-following data into the 158K instruction-following data from LLaVA as the training set. Note that the visual encoder is frozen throughout the whole training period, which might restrict the performance of text recognition as CLIP is trained for general-purpose text-image alignment. Better choices of the visual encoder [42] or further fine-tuning CLIP-ViT [29] might further benefit the visual understanding capability, which we leave as future work.

5 Experiments

We use the same training parameters as LLaVA[8], except (1) We set the maximum sequence length to 1024 during pretraining. (2) We first pad any given image to a square shape before resizing it to

[8] `https://github.com/haotian-liu/LLaVA`

	Res	Conversation	Detail	Complex	Read
LLaVA (Original)	224^2	83.1	75.3	96.5	-
LLaVA	336^2	83.9	78.2	95.3	87.9
LLaVAR	336^2	84.5	78.9	96.5	91.7

Table 5: Relative scores (w.r.t. text-only GPT-4) for instruction-following questions, where the first three dimensions are based on natural images, the last dimension ("Read") is based on text-rich images. In the first row, we show the original results (224^2-based) fetched from [3]. We report our reproduced LLaVA on 336^2 resolution for a fair comparison.

Question: Based on the title and the image on the cover, what can be inferred about the content of "Bo's Lasting Lessons" and its potential target audience?		Relative Score
GPT-4 (Text-Only)	From the title and image, it can be inferred that "Bo's Lasting Lessons" is a book about leadership lessons taught by a legendary coach, likely in the context of sports. The image of men running suggests that the book might be focused on a specific sport, such as football or track and field. The target audience could be individuals interested in coaching, sports, or leadership in general, looking to learn from the experiences and teachings of a successful coach.	-
LLaVA	The book "Bo's Lasting Lessons: The Timeless Teachings of Bo Schembechler" is written by Bo Schembechler and Jim Harshaw. Given the title and the image on the cover, which shows a football coach running across the field with his players in pursuit, it can be inferred that the content of the book is likely focused on football and the strategies, techniques, or principles that can help players improve their performance and achieve success in the sport. The potential target audience for this book could be football players, coaches, and fans who are interested in learning more about the game and the experiences and insights of a seasoned coach like Bo Schembechler.	75%
LLaVAR	"Bo's Lasting Lessons" appears to be a book that shares the legacy coach Bo Schembechler's teachings on leadership and timeless fundamentals. The image on the cover, featuring a football coach leading his team, further supports this inference. The book might be targeted towards sports enthusiasts, particularly fans of football, and people interested in learning about leadership principles and fundamentals from a renowned coach's perspective. The audiobook, read by John H. Mayer and featuring a foreword and conversation with John U. Bacon, suggests that it could also appeal to audiences who prefer listening to books or want to gain more insights from the author and other experts.	129%

Table 6: An example for our collected evaluation data, corresponding to the "**Read**" dimension of GPT-4-based instruction-following evaluation. The corresponding image is Figure 11 in Appendix.

the desired input size, avoiding some image content being cropped during the preprocessing. For both resolutions (224^2, 336^2), we reproduce the original LLaVA for a fair comparison. The GPT-4 model used in this work refers to the `gpt-4-0314` version, while the cost for collecting finetuning data is around \$300. Temperature is set to 1.0 for training data generation, 0.7 for evaluation data generation, and 0.2 for GPT-4-based evaluation. All experiments are run on NVIDIA A100 80GB GPU. During the evaluation, the temperature is set to 0.9 for training text-based VQA, 0.7 for GPT-4-based instruction-following evaluation, and 0.2 for other qualitative demonstrations.

5.1 Quantitative Analysis

Text-based VQA Following the evaluation protocol in [10], we evaluate the performance of LLaVAR on four text-based VQA datasets: ST-VQA [43], OCR-VQA [31], TextVQA [44], and DocVQA [45], representing various domains (see Appendix C for more details). We present the results of baseline models and our models in Table 2. Note that InstructBLIP includes OCR-VQA in its training sets, making it incomparable with our settings. In both two resolutions and all four

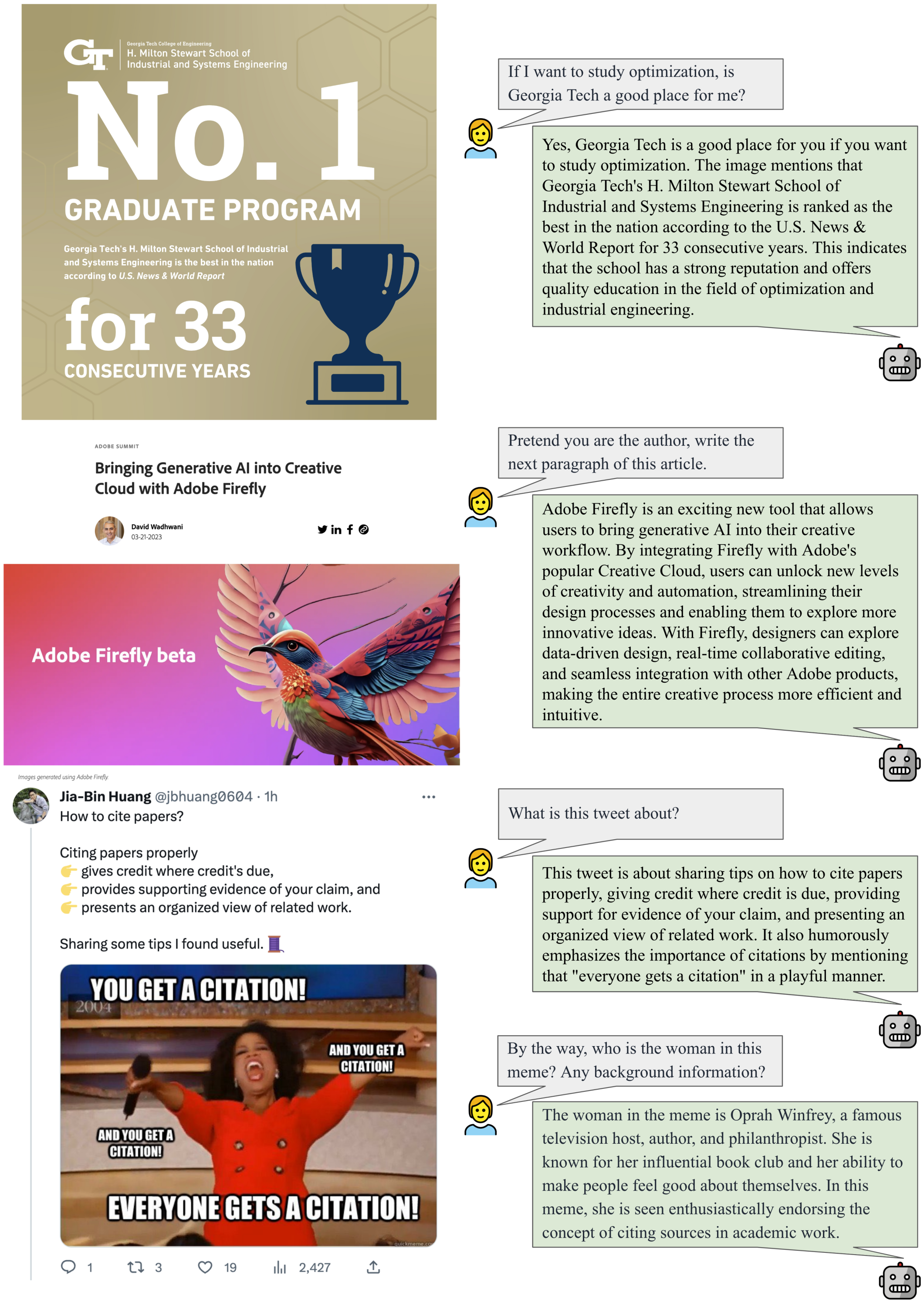


Figure 5: Transferred instruction-following capability of LLaVAR.

datasets, LLaVAR substantially improves the LLaVA baseline, demonstrating our collected data can bring robust improvement. Moreover, the improvement is more significant in the 336^2 resolution compared to 224^2, indicating the collected data might bring larger improvement in even higher resolutions. Our best model, 336^2-based LLaVAR, performs best in 3 out of 4 datasets evaluated.

H: When will this movie be released?
L: The movie "The Call of the Ottoman" starring Daniel Craig is expected to be released on February 14th, 2023.
R: The movie will be released on January 13.
H: Who starred in the movie?
L: The movie stars Tom Cruise and Cobie Smulders.
R: Tom Hanks starred in the movie.
H: What is this movie about?
L: The movie "A Man Called Ove" is a comedy-drama film directed by Hannes Holm. It tells the story of a grumpy old man named Ove, who is played by Rolf Lassgård. Ove is a widower who lives a solitary life in a small apartment. (...)
R: The movie is about Tom Hanks, who is called Ottol, and it is an international bestseller based on a true story. The movie will be released on January 13th.
H: Can you describe the given image?
L: The image is a movie poster featuring a man and a cat. The man, who appears to be a character from the movie "The Ugly Truth," (...)
R: The image features a man standing in front of a house with a cat beside him.
H: Who directed the film?
L: The film was directed by Florian Henckel von Donnersmarck.
R: The film was directed by Tom Hanks, who made his directorial debut with this movie.

Figure 6: Qualitative Comparison. We color the human inputs (***H***) in blue, LLaVA's responses (***L***) in pink, and LLaVAR's responses (***R***) in green. Both are 336^2-based models. We replace the excessively long, hallucinated response with (...).

Note that this is not a fair comparison. Some key different factors include different language decoders, different resolutions, and different magnitudes of text-image training data.

Ablation Study We further report the result of the ablation study in Table 3 and Figure 4. **(i)** Based on variants (2) and (3), we find that collected data can benefit the pretraining stage ($R_{\text{pretraining}}$) and finetuning stage ($R_{\text{finetuning}}$) separately while being complementary to each other in most cases [9]. More importantly, enhancing the pretraining stage alone achieves the second-best overall performance, indicating the potential to boost textual detail understanding without dependence on GPT-4-generated high-quality data. **(ii)** Using pretraining images, we obtain $C_{\text{pretraining}}$ by replacing the pretraining instructions with questions & captions, the same pattern as LLaVA. As variant (4) is not as good as (2), we can conclude that OCR is more advantageous than captions. **(iii)** We further validate the value of GPT-4 generated data by generating noisy finetuning data ($N_{\text{finetuning}}$), similar to pretraining data. Variant (5) achieves comparable accuracy as variant (3). However, as shown in Figure 4, such noisy finetuning data hurts the instruction-following capability: (5) responds with all recognized texts while ignoring the questions. Overall, our ablation study confirms the necessity of our pipeline.

ScienceQA Starting from our pretrained LLaVAR (336^2-based, without finetuning), we also report the results of further finetuning on the ScienceQA dataset [39] in Table 4, which is a multimodal multi-choice QA dataset covering diverse domains. Our motivation is that some images in this dataset contain text descriptions and tables that require textual understanding within images. The LLaVAR model finetuned on ScienceQA achieves an average accuracy of 91.42%, better than LLaVA (90.92%), while the biggest improvement comes from natural science questions (+1.43%).

GPT-4-based instruction-following evaluation Following [3], we report the GPT-4 evaluation results on instruction-following questions in Table 5. (i) **Natural Images**: 90 questions based on

[9]Since the metric only consider the recall, it might favor variant (2)(4)(5) due to their longer outputs.

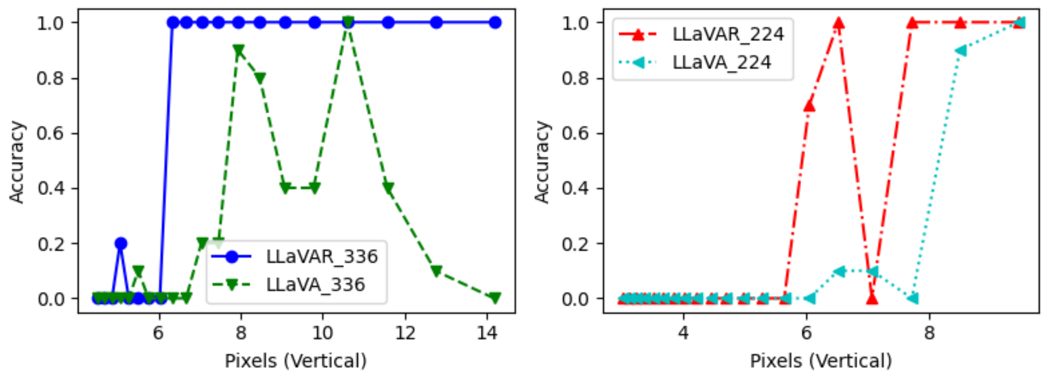


Figure 7: Case study of the recognizable font size. We plot the results for the 336^2-based models on the left and the 224^2-based models on the right.

30 COCO validation images, including three aspects: conversation, detail description, and complex reasoning. This aims at testing whether our collected data will hurt, maintain or improve the model's performance on natural images. First of all, using a higher resolution brings improvement (+2.9) in the performance of detail description, which is intuitive. Furthermore, LLaVAR achieves a better trade-off and increases the performance of all three aspects (+1.6 on average). (ii) **Text-Rich Images**: Similar to collecting the finetuning data, we leverage 50 text-rich images from LAION to collect instruction-following questions based on OCR results and human annotation. We then collect responses from our trained model and use GPT-4 to calculate the relative score w.r.t GPT-4 responses. We provide an example in Table 6 and add this as an extra dimension "Read" to the GPT-4-based evaluation Table 5. Our model demonstrates a more significant (+3.8) improvement on this axis.

5.2 Qualitative Analysis

We use a recent movie poster [10] to demonstrate the difference between LLaVA and LLaVAR regarding interaction with humans based on text-rich images. LLaVA, without augmenting the textual understanding within images, suffers from hallucination while answering these questions. Some mentioned movies, like "A Man Called Ove" and "The Ugly Truth", are real movies, suggesting the language decoder is hallucinating its internal knowledge while the visual encoder cannot encode helpful information. Alternatively, LLaVAR can correctly answer many of the provided questions with **faithful** information, which is clearly grounded in the image. However, there are still some limitations, such as the spelling error "ottol". Also, the final question asks for information that is not observable from the given poster, where an expected response should express such uncertainty instead of giving concrete answers. However, both models fail to answer it correctly.

5.3 Case Study: Recognizable Font Size

By scaling the poster in Figure 6, we provide a case study on the recognizable font size on the top of the question, "When will this movie be released?". We calculate the number of vertical pixels for the ground truth "January 13th" in the scaled posters and estimate the accuracy for each scale based on ten trials (Fig 7). (i) For our model LLaVAR, it can no longer recognize the ground truth while its vertical size is less than 6 pixels. Meanwhile, the 336^2-based version provides better robustness as it works consistently well for any scale greater than 6 pixels. (ii) For the baseline model LLaVA, surprisingly, it achieves a certain level of correctness while the ground truth is between 8 and 10 pixels with poor performance on larger scales (e.g., 14 pixels). This suggests that LLaVA, without specific training to recognize texts, still recognizes texts at specific scales with particular contexts. However, the lack of robustness prevents it from better performance in understanding text-rich images.

[10] `https://www.imdb.com/title/tt7405458/`

5.4 Transferred Instruction-following Capability

According to the dataset statistics (Table 1) and visualization (Figure 10), our collected instruction-following data is not as diverse and substantial as LLaVA. This can be attributed to the relatively limited information given GPT-4 compared to five different human-written captions used in LLaVA. The content of text-rich images is also less diverse than natural images. While using more complex in-context examples can definitely stimulate generating more complicated instruction-following examples, it can also multiply the cost. In Figure 5, we demonstrate the transferred instruction-following capability of LLaVA, potentially from both the LLaVA data and the Vicuna backbone. While the extra data we add mainly focuses on understanding the visible texts within images, LLaVAR manages to build its reasoning, writing, and elaboration skills based on the top of its text recognition capability in an end-to-end manner. This allows users to interact with various online content based on simple screenshots.

6 Conclusion

In this work, we enhance visual instruction-tuned models in terms of their capability to read texts in images. Using text-rich images from the LAION dataset, we collect 422K noisy instruction-following examples using OCR results only and 16K high-quality instruction-following data based on text-only GPT-4. These two sets of data are leveraged to augment the pretraining stage and finetuning stage of LLaVA accordingly. Our model, LLaVAR, demonstrates superior performance in understanding texts within images and following human instructions on both prior benchmarks and real-world online content. Moreover, our analysis shows that the same augmented data is more effective with higher resolution. Also, using noisy instruction-following examples to augment pretraining essentially boosts the model performance without prompting GPT-4. For future work, we encourage exploration of (i) better image selection criteria or domain reweighting strategy [46] and (ii) more data-efficient and cost-efficient ways to enhance multimodal instruction-following datasets.

References

[1] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[4] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[5] Chunyuan Li. Large multimodal models: Notes on cvpr 2023 tutorial. *ArXiv*, abs/2306.14895, 2023.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[10] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Cheng lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models, 2023.

[11] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023.

[12] OpenAI. Gpt-4 technical report, 2023.

[13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[14] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022.

[15] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2022.

[16] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[17] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

[18] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data, 2023.

[19] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[20] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

[21] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators, 2023.

[22] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.

[23] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[24] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2023.

[25] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jia-Bin Huang, Jinglin Liu, Yixiang Ren, Zhou Zhao, and Shinji Watanabe. Audiogpt: Understanding and generating speech, music, sound, and talking head. *ArXiv*, abs/2304.12995, 2023.

[26] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities, 2023.

[27] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

[28] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention, 2023.

[29] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

[30] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[31] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[32] Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. Cream: Visually-situated natural language understanding with contrastive reading model and frozen large language models, 2023.

[33] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. *Proceedings of the 30th ACM International Conference on Multimedia*, Oct 2022. doi: 10.1145/3503161.3547911. URL `http://dx.doi.org/10.1145/3503161.3547911`.

[34] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval, 2015.

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[36] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. URL `https://doi.org/10.5281/zenodo.7733589`.

[37] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[38] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning, 2022.

[39] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022.

[40] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2023.

[41] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *ArXiv*, abs/2304.09842, 2023.

[42] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels only, 2022.

[43] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Minesh Mathew, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sep 2019. doi: 10.1109/icdar.2019.00251. URL `http://dx.doi.org/10.1109/ICDAR.2019.00251`.

[44] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. doi: 10.1109/cvpr.2019.00851. URL `http://dx.doi.org/10.1109/CVPR.2019.00851`.

[45] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2020.

[46] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining, 2023.

[47] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[48] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[49] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[50] Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. Judging a book by its cover, 2016.

Appendix

A

CLIP-based categorization Based on the observation of selected clusters, we divide the images used into 8 categories. For each category, we use one or multiple words as labels.

- **Quote & Meme**: “quote”, “internet meme”.
- **Poster**: “movie poster”, “podcast poster”, “TV show poster”, “event poster”, “poster”,
- **Book Cover**: “book cover”, “magazine cover”.

Instructions
Identify any text visible in the image provided.
List all the text you can see in the given image.
Enumerate the words or sentences visible in the picture.
Describe any readable text present in the image.
Report any discernible text you see in the image.
Share any legible words or sentences visible in the picture.
Provide a list of texts observed in the provided image.
Note down any readable words or phrases shown in the photo.
Report on any text that can be clearly read in the image.
Mention any discernable and legible text present in the given picture.

Table 7: Ten instructions asking for OCR results.

- **Game Cover**: "game cover".
- **Ad & Product Packaging**: "ad", "advertisement", "food packaging", "product packaging".
- **Infographic**: "chart", "bar chart", "pie chart", "scatter plot".
- **Educational Material**: "ad", "advertisement", "food packaging", "product packaging".
- **Logo**: "logo".

For each word, we use the following templates to achieve embedding-space ensembling [7]:

- "a photo of a {}."
- "a blurry photo of a {}."
- "a black and white photo of a {}."
- "a low contrast photo of a {}."
- "a high contrast photo of a {}."
- "a bad photo of a {}."
- "a good photo of a {}."
- "a photo of a small {}."
- "a photo of a big {}."

For each image, we calculate the similarity between the image and all words mentioned above using `CLIP-ViT-L/14`. If the highest similarity is less than 0.15, we then classify the image into **Other**, otherwise we classify into the "super class" (e.g., **Poster**) of the word (e.g., "movie poster") with the highest similarity.

B

System Message adapted from [3].

> *You are an AI visual assistant, and you are seeing a single image. What you see is provided with two OCR results and one image caption describing the information within the same image you are looking at. Image captions might include hallucinations, while OCR results are more accurate. Answer all questions with definite answers as you are seeing the image.*
> *Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.*
> *Include questions asking about the visual content of the image (e.g., the man, the sunset, the ocean.) and the texts contained in the image. Only include questions that have definite answers:*

(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any questions that cannot be answered confidently;
(3) DO NOT mention OCR or image caption in your questions and answers;
(4) DO NOT ask about information from captions while it looks unrelated to or contradicts OCR results.
Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the texts in the image, asking to discuss about the design of the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

Few-shot Examples

Human:
WHAT F YOUR BLESSINGS COME THROUGH RAINDROPS PDf Induded LAURASTOR Y DEVOHONAL BASED ON THE GRAAAMY VARD WIN ISONG mAT HAS INSEIRED MILLION
WHATIFYOUR BLESSINGS COMETHROUGHRAINDROPS PDF Included LAURASTORY A DEVOTIONAL BASEDON THE GRAMMY AWARD WINNI PISONGTHATHASINSPIREDMILLIONSE
a girl is standing in a field with a rainbow
GPT:
Question: What is the name of the devotional mentioned in the image?
Answer: The devotional is called "What If Your Blessings Come Through Raindrops" by Laura Story.
Question: What is special about the song mentioned in the image?
Answer: The song is special because it's a Grammy Award-winning song that has inspired millions of people."

Human:
One of the hardest things in life to accept is a called third strike Robert Frost te
One of the hardest things in life to accept is a called third strike Robert Frost quotefancy
a close up of a baseball glove
GPT:
Question: Why is the third strike hard to accept? Explain the quote to me.
Answer: The quote Öne of the hardest things in life to accept is a called third strikeättributed to Robert Frost is a metaphor for life situations. In baseball, a called third strike is when the umpire determines that a pitch was a strike, but the batter did not swing. This can be frustrating for the batter because they missed an opportunity to hit the ball due to either a lack of confidence or misjudgment.
In life, this metaphor refers to difficult moments where we might miss opportunities or face setbacks because of our own inaction or hesitation. The quote suggests that it is hard to accept these missed opportunities or challenges, especially when they result from our own choices or lack of action.

C

Details of evaluation dataset

- ST-VQA [43] contains 31791 questions that require understanding the scene text, based on images from COCO [9], Visual Genome [47], ImageNet [48], etc.
- TextVQA [44] contains 45,336 questions that need reading and reasoning about the text in images to answer, based on images from OpenImages [49].

- OCR-VQA [31] contains more than 1 million questions asking about information from book cover images [50].
- DocVQA [45] contains 50000 questions based on document images.

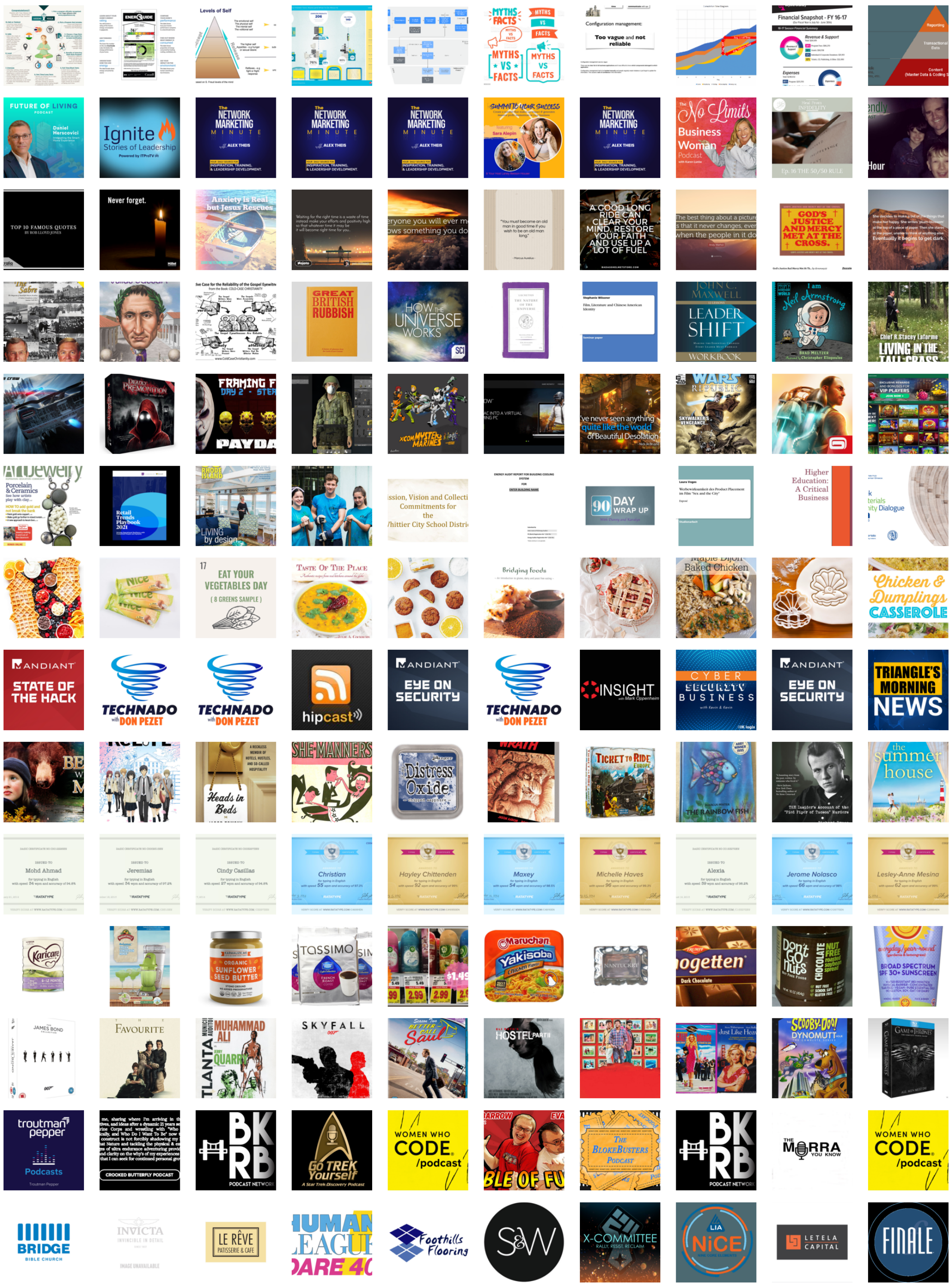


Figure 8: All 14 clusters we selected as text-rich images. Each row corresponds to one cluster, where we show ten randomly sampled examples before de-duplication.

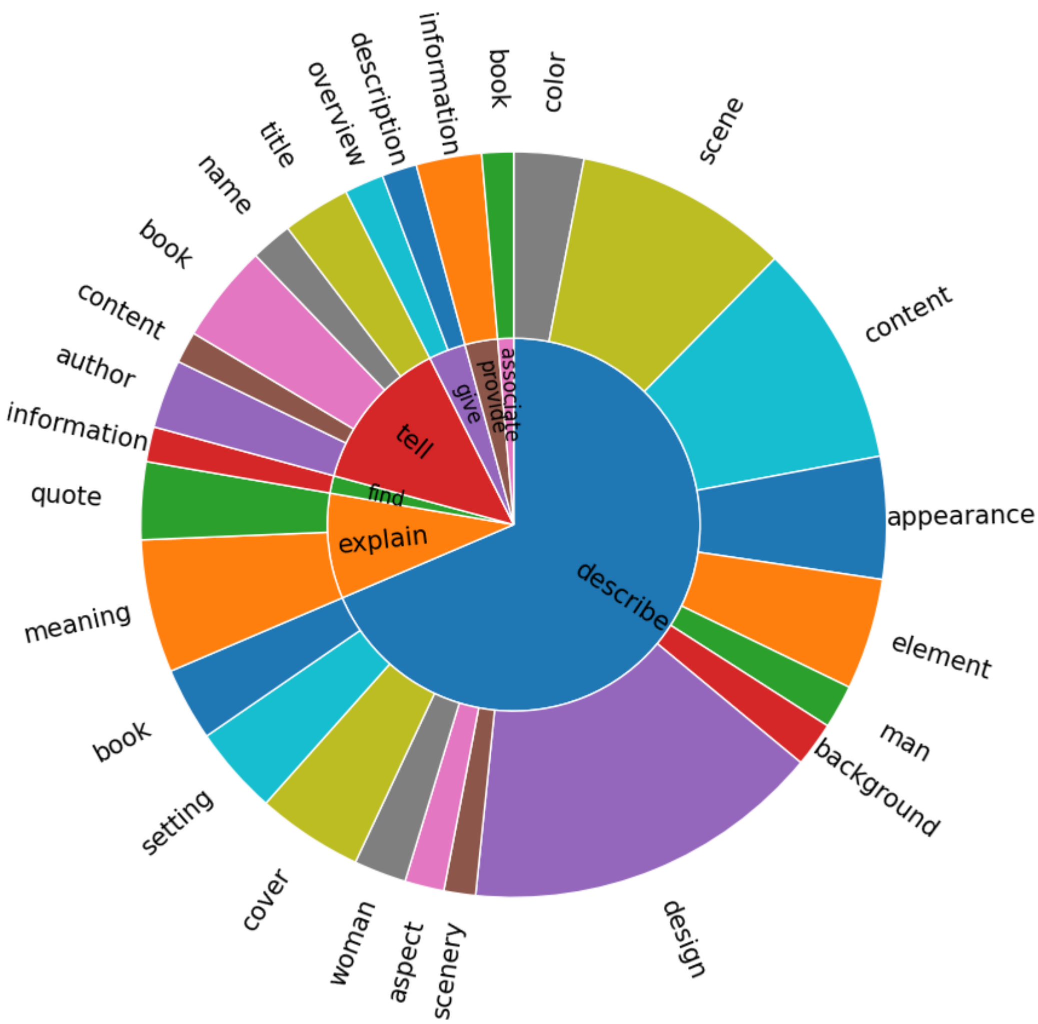


Figure 9: Visualization of collected instructions.

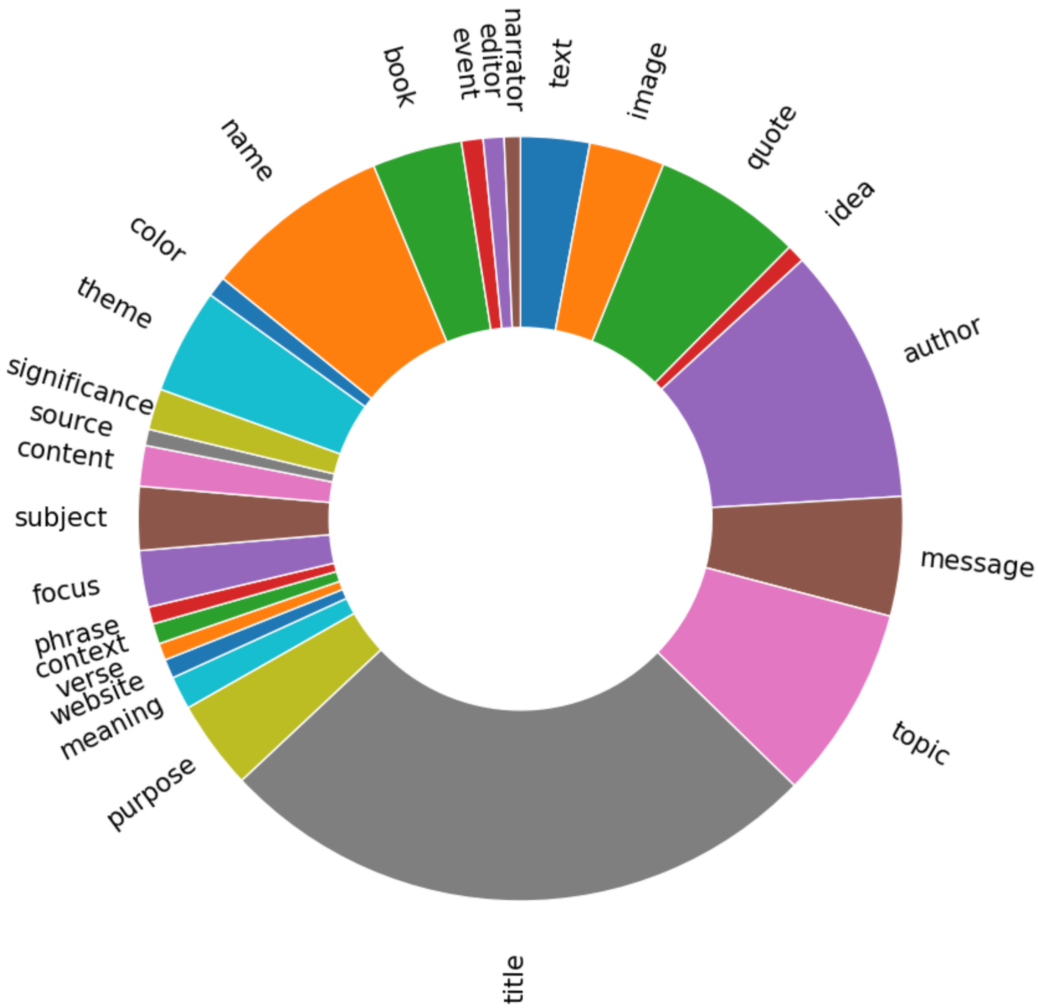


Figure 10: Visualization of collected instructions.

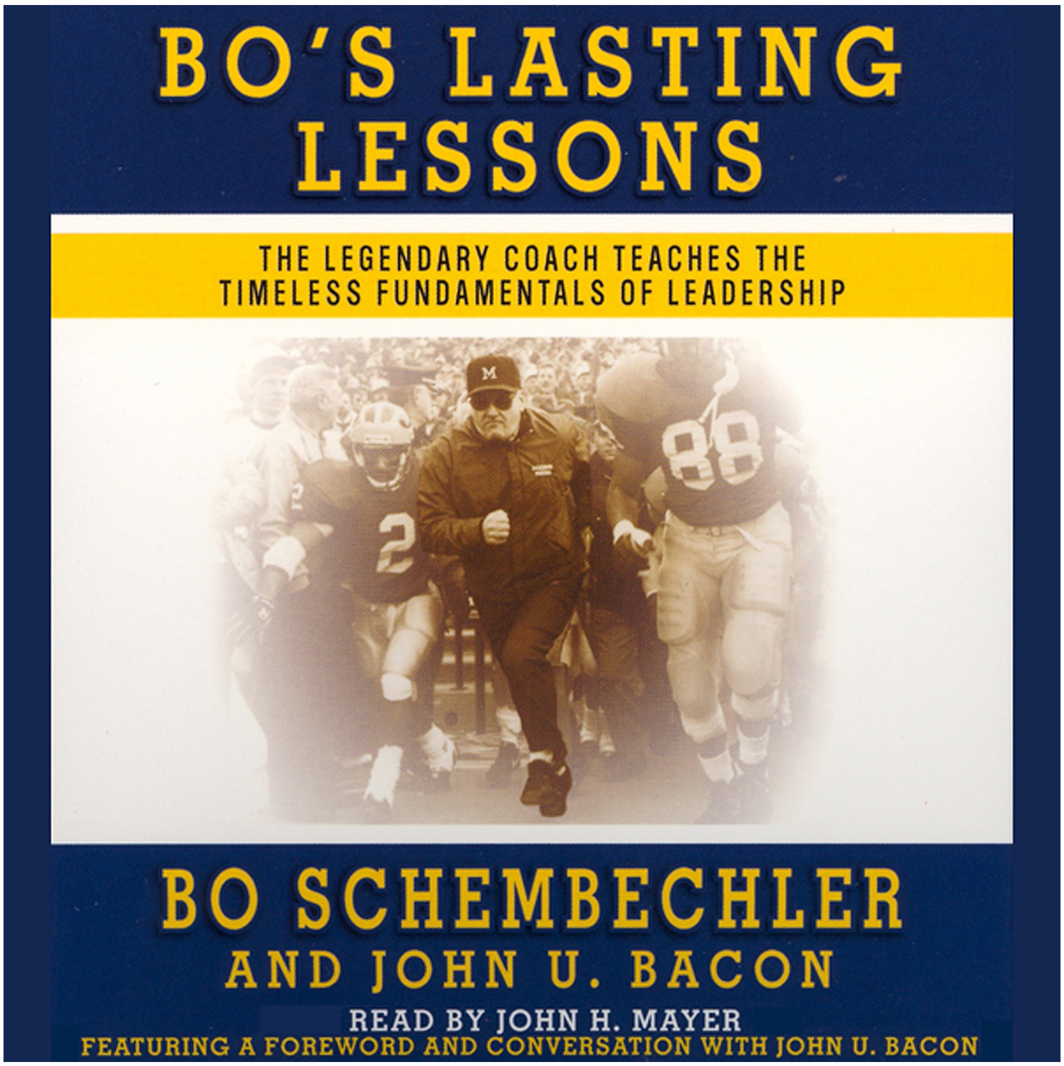


Figure 11: An example for the Read dimension of GPT-4-based instruction-following evaluation.