Long-Form Answers to Visual Questions from Blind and Low Vision People

Mina Huh[♡], Fangyuan Xu[♡], Yi-Hao Peng[♠], Chongyan Chen[♡],
Hansika Murugu[♣], Danna Gurari[◇], Eunsol Choi[♡], Amy Pavel[♡]

[♡]The University of Texas at Austin, [♠]Carnegie Mellon University,
[♣] Hong Kong University of Science and Technology [◇]University of Colorado Boulder

minahuh@cs.utexas.edu

Abstract

Vision language models can now generate long-form answers to questions about images – long-form visual question answers (LFVQA). We contribute VizWiz-LF[1], a dataset of long-form answers to visual questions posed by blind and low vision (BLV) users. VizWiz-LF contains 4.2k long-form answers to 600 visual questions, collected from human expert describers and 6 VQA models. We develop and annotate functional roles of sentences of LFVQA and demonstrate that long-form answers contain information beyond the question answer such as explanations and suggestions. We further conduct automatic and human evaluations with BLV and sighted people to evaluate long-form answers. BLV people perceive both human-written and generated long-form answers to be plausible, but generated answers often hallucinate incorrect visual details, especially for unanswerable visual questions (e.g., blurry or irrelevant images).

1. Introduction

Traditional visual question answering (VQA) models respond to visual questions with short answers. This is because they were designed for mainstream dataset challenges for which answers typically are brief, containing one or two words [3, 21, 22]. The rise of large vision language models (VLMs) has introduced a new class of VQA models that can generate *long-form answers* [1, 2, 9, 13, 35]. While long-form visual question answering (LFVQA) has enormous potential, we have limited knowledge about the content and quality of long-form answers.

Our work investigates the content and quality of long-form answers to visual questions posed by blind and low vision (BLV) users. LFVQA holds particular value for BLV people who take photos to overcome real-world accessibility barriers and are already using tools like Be My AI [5] powered by GPT-4V. Unlike artificial settings, authentic VQA involves challenges such as conversational questions and low-quality images. We aim to understand the potential and limitations of LFVQA given its societal relevance.

We introduce *Vizwiz-LF*, a dataset of 4.2k long-form answers to visual questions from BLV people [6, 22]. We collect and evaluate long-form answers from human experts and 6 state-of-the-art VLMs (Figure 1). To understand the content of LFVQA, we design and annotate the functional roles (*e.g.*, answer, explanation, suggestion) of long-form answer sentences in our dataset. While most answers from 5 models (Gemini, LLaVA, InstructBLIP, QWEN, BLIP-2) included only two functional roles (*confirmation, answer*), human experts and GPT-4V often used additional roles (*e.g.*, explanation, suggestion, auxiliary information).

To assess VLMs in LFVQA, we conduct an automatic evaluation using reference-based metrics (ROUGE [28], METEOR [15], BERTScore [41], LAVE [31]) with short-form reference answers from VizWiz and our long-form reference answers. While reference-based evaluations typically use short reference answers and thus penalize long answers for including extra information [25, 37], we show that extracting answer sentences from long answers can mitigate this.

To understand how humans evaluate long-form answers, we conduct an evaluation study with both sighted and BLV people. Our results also reveal that sighted people's evaluation without an image is not a strong proxy for BLV preferences. For instance, BLV evaluators perceive incorrect answers as more plausible than sighted evaluators.

While focused on visual questions from BLV users, our findings have broader implications for VQA research. First, we create the first dataset with both short and long answers – enabling transfer from short-answer VQA, where models already perform well, to long-answer tasks. Second, our functional roles can guide both evaluation and improvement of LFVQA. Lastly, our human study underscores the need

[1]Our data and code are available at anonymizedforsubmission.

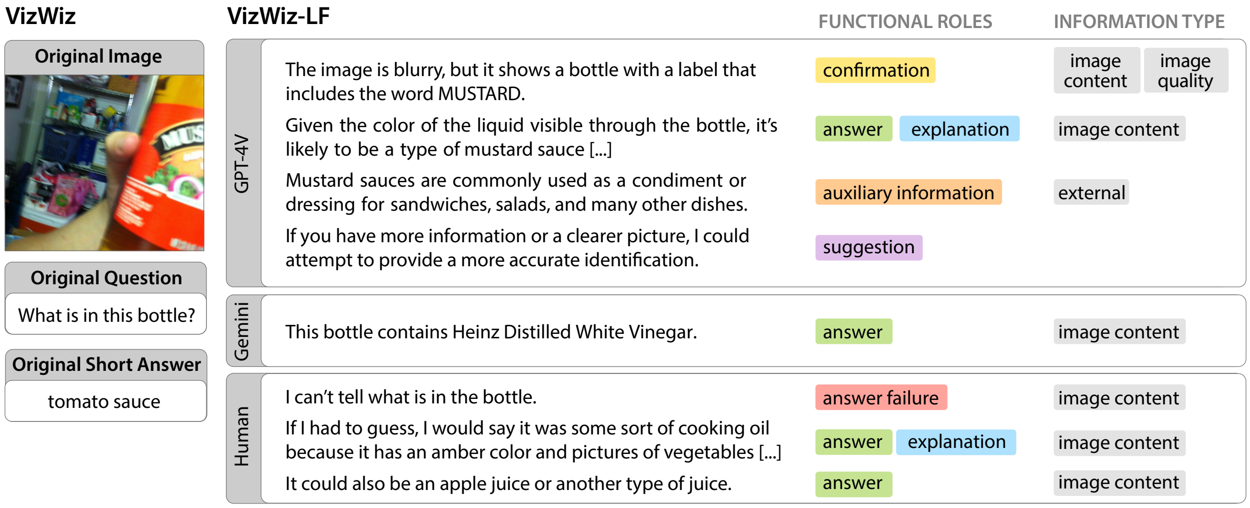


Figure 1. For each visual question, we collect an answer from human expert describers and 6 VLMs (GPT-4V, Gemini, LLaVA, InstructBLIP, QWEN, BLIP-2). To understand the content and structure of long-form visual question answers, we create a taxonomy of functional roles and annotate answer functional roles and information types at a sentence level.

	VizWiz		VizWiz-LF						
Category	# of words Question	# of words Answer	Human	GPT-4V	Gemini	# of words LLaVA	QWEN	BLIP-2	InstructBLIP
Identification	5.2 (4.3)	1.8 (1.2)	42.9 (30.0)	**72.0** (25.4)	15.7 (15.6)	28.9 (21.0)	21.2 (19.9)	12.5 (2.7)	14.9 (10.8)
Description	7.0 (5.0)	1.6 (1.4)	34.8 (30.5)	**57.6** (46.5)	11.6 (18.0)	15.0 (17.9)	13.5 (14.8)	12.2 (2.7)	17.4 (13.6)
Reading	7.9 (4.4)	1.8 (1.9)	39.8 (36.5)	**66.5** (54.9)	16.3 (27.6)	22.2 (30.4)	18.6 (26.8	12.1 (3.0)	20.2 (17.8)
Others	11.5 (7.4)	1.6 (1.7)	47.9 (37.2)	**96.6** (54.1)	22.3 (29.5)	41.9 (38.4)	34.7 (36.2)	12.6 (3.4)	23.0 (25.0)
Total	7.9 (5.9)	1.7 (1.6)	41.2 (34.0)	**73.2** (48.9)	16.4 (23.7)	27.0 (29.8)	22.0 (26.8)	12.3 (3.0)	18.9 (17.9)

Table 1. Statistics of the sampled VizWiz data and the long-form answers from human expert describers and 6 VLMs. We collect 150 questions from four categories. Numbers in each cell represent the average with the standard deviation in parentheses.

to assess LFVQA not just for factual accuracy, but also for user-centered metrics like relevance and plausibility.

2. Dataset

Visual Questions We extend the VizWiz dataset [22], which features visual questions from BLV individuals. To ensure diversity, we sample 600 image-question pairs balanced across all four question types (Identification, Description, Reading, and Others) from the VizWiz taxonomy [7].

Long-form Answer Collection We hired 20 expert describers via Upwork to provide detailed long-form answers, unlike the original VizWiz dataset [22] which focused on short crowd responses. To benchmark modern VLMs on LFVQA, we generated long-form answers using 6 models—GPT-4V [1], Gemini [35], LLaVA [29], InstructBLIP [12], Qwen-VL-Chat [4], and BLIP-2 [27]—in a zero-shot setting[2]. These models were chosen for their public availability, strong zero-shot performance, and architectural diversity. Table 1 compares our long-form answers to VizWiz short-form answers[3]. Expert answers are 24× longer than short answers in the original dataset.

3. Functional Role Analysis

Functional Roles We annotated each sentence with its functional role. Using open coding [23], we developed a taxonomy of 8 roles: *Confirmation*, *Answer*, *Answer Failure*, *Auxiliary Information*, *Auxiliary Information Failure*, *Explanation*, *Suggestion*, and *Miscellaneous* (Details in §8.1.)

Annotation and Classification To annotate the 4.2k long-form answers from § 2, we first collected human annotations on 180 answers (522 sentences) as ground-truth. Given the question, image, and answer, annotators assigned up to three functional roles. 5 authors conducted 3-way annotations, reaching substantial agreement (Fleiss $\kappa = 0.76$) [17]. We then used a GPT-4 classifier with few-shot prompting (see

[2]See §7.3 for configuration details.

[3]We use the majority-voted answer from 10 crowd responses [8, 39].

Answers (%)	Confirmation		Answer		Ans. Failure		Auxiliary		Aux. Failure		Explanation		Suggestion		Misc.	
	ans.	*unans.*	*ans.*	*unans.*	*ans.*	*unans.*	*ans.*	*unans.*	*ans.*	*unans.*	*ans.*	*unans.*	*ans.*	*unans.*	*ans.*	*unans.*
Expert	50	50	96	62	6	56	53	45	2	**3**	35	58	6	27	3	7
GPT-4V	**61**	**60**	72	64	**38**	**59**	**58**	**52**	**4**	1	**69**	**84**	**35**	**63**	7	**13**
Gemini	29	15	97	87	2	14	16	13	0	0	5	11	2	7	1	3
LLaVA	53	39	**98**	87	2	14	29	29	0	0	11	24	2	15	1	2
QWEN	46	30	89	62	11	42	17	15	0	1	14	37	4	13	2	6
BLIP-2	46	30	85	76	8	16	3	1	0	0	6	7	1	3	**11**	11
InstructBLIP	35	35	96	**90**	4	10	13	17	0	0	15	17	1	5	2	2

Table 2. Distribution of answers with each functional role to answerable (*ans.*) and unanswerable (*unans.*) questions. 230 of 600 questions were marked unanswerable in VizWiz.

§8.2) to annotate the rest. The classifier achieved a weighted per-role F1 of 0.74, compared to human-human F1 of 0.79.

3.1. Analysis

Table 2 shows the distribution of functional roles of long-form answers in our dataset. As 38% of our sampled questions are marked as unanswerable [4], we report the distribution for answerable and unanswerable questions separately.

Expert and GPT-4V's answers contain diverse functional roles. While all sources commonly include an *Answer*, LLaVA and Gemini have the highest proportion, with Gemini often providing single-sentence responses that directly address the question without elaboration. In contrast, expert and GPT-4V answers cover diverse roles. GPT-4V frequently uses *Confirmation* to confirm the image content when the question is vague (*e.g., "The image you've provided is blurry, but it shows a part of a bottle with a blue label."*). Both expert and GPT-4V responses provide *Auxiliary* details—for example, when asked about the color of a t-shirt, they might also mention the pants' color or a logo on the shirt.

Most VLMs rarely abstain, even when the question is unanswerable. As shown in Table 2, 4 VLMs (Gemini, LLaVA, BLIP-2, InstructBLIP) rarely include *Answer Failure* sentences, regardless of answerability. In contrast, Expert, GPT-4V, and QWEN more frequently abstain to unanswerable questions. GPT-4V, however, sometimes over-abstains for answerable questions when the image is unclear. Expert and GPT-4V often combine *Answer* and *Answer Failure* in the same response—e.g., making a tentative guess (*"It might be a jar of pasta sauce."*) while noting limitations (*"The text is not completely legible."*). When GPT-4V abstains, it typically includes an *Explanation* (*e.g.,* poor image quality) or a *Suggestion* (*e.g.,* retaking the photo).

4. Automatic Evaluation

To analyze the performance of VLMs in LFVQA, we conduct an automatic evaluation of long-form answers with reference-based metrics. To adapt the traditional reference-based metrics to long-form answers, we consider both short crowd answers [22] and long expert answers (our dataset) as ground truth references and explore how our functional role classifier can be used for sentence-level evaluation. To our knowledge, we are the first to consider long-form reference answers when evaluating VQA.

[4] Following VizWiz[22], we consider the questions with the majority of "unanswerable" as unanswerable and otherwise answerable.

Data We selected 360 examples in our dataset written by experts, GPT-4V and Gemini (demonstrated as state-of-the-art for image understanding tasks [35, 38] and used in accessibility apps [5, 34]), given the cost of evaluation. The samples were balanced across question categories (Identification, Description, Reading, and Others).

Method We evaluate long-form answers using 4 reference-based VQA evaluation metrics: ROUGE [28], METEOR [15], BERTScore [41], and LAVE, a GPT-4-based metric [31] (details in §9.1). Comparing long-form answers to short-form reference answers can penalize long-form answers for including additional information (*e.g.*, explanation, suggestion) [25, 31]. We consider two approaches. First, we explore the potential of long-form references in VQA evaluation by leveraging the expert long-form answers in our *VizWiz-LF* as ground truths (l_r) for evaluating model long answers (l_c). However, many existing VQA datasets have only short ground truth answers [3, 22], and collecting long-form ground truths on a large scale is costly. Thus, we also explore the use of short-form references (s_r) as ground truths for evaluating extracted *Answer* role sentences from long-form answers (l'_c).

Results As Table 3 summarizes, Gemini outperformed GPT-4V and experts on ROUGE, METEOR, and BERTScore, likely due to these metrics penalizing the additional information in GPT-4V and expert answers. In contrast, the LLM-based metric favored GPT-4V and experts, and was the only metric moderately correlated with human ratings (§9.1), highlighting its promise for LFVQA evaluation. Across all metrics, using long-form references improved scores over short-form references (s_r). GPT-4V answers, in particular, saw larger gains with long-form refer-

Answer Source	Lex-Overlap (Unigram)	ROUGE				METEOR				BERTScore				GPT-4			
		s_r+l_c	s_r+l'_c	l_r+l_c	l_r+l'_c	s_r+l_c	s_r+l'_c	l_r+l_c	l_r+l'_c	s_r+l_c	s_r+l'_c	l_r+l_c	l_r+l'_c	s_r+l_c	s_r+l'_c	l_r+l_c	l_r+l'_c
GPT-4V	0.35	0.04	0.19	**0.22**	**0.32**	0.08	0.17	**0.29**	**0.31**	0.8	0.83	**0.87**	**0.88**	**0.76**	**0.76**	**0.76**	**0.73**
Gemini	0.22	**0.19**	**0.22**	**0.22**	0.26	**0.18**	**0.20**	0.16	0.21	**0.84**	**0.84**	**0.87**	0.87	0.63	0.64	0.45	0.51
Expert	**0.36**	0.06	0.2	-	-	0.09	0.18	-	-	0.81	**0.84**	-	-	0.71	0.7	-	-

Table 3. Evaluation results with reference-based metrics (*reference+candidate*). For reference answers, we use VizWiz crowd's majority-voted answers (s_r), experts' long-form answers (l_r), and extracted answer sentences of experts' long-form answers (l'_r). For candidate answers, we use original long-form answers generated by models and experts (l_c) and extracted answer sentences (l'_c).

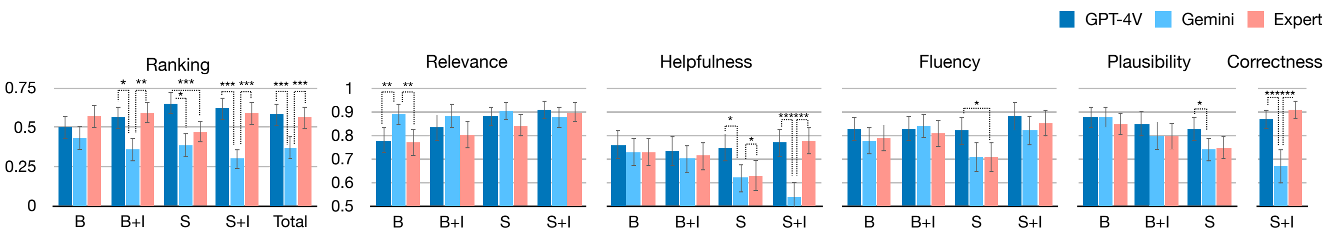


Figure 2. Average rankings and ratings of long-form answers (GPT-4V, Gemini, Expert) across four groups: BLV (*B*), BLV with image caption (*B+I*), Sighted without image (*S*), Sighted with image (*S+I*). We measured significance using the Freidman test followed by the pair-wise Wilcoxon test. $p < 0.05$ marked with *, $p < 0.01$ marked with *, $p < 0.001$ marked with *** after Bonferroni correction.

ences (l_r) in ROUGE and METEOR (see Figure 5). While expert long-form answers are hard to scale, our findings underscore the importance of long-form references. Future work can explore evaluating diverse functional roles, such as grounding *confirmation* sentences in images or fact-checking *external information* sentences.

5. Human Evaluation

We conduct an evaluation study with 20 BLV and 20 sighted people [5] While most prior research evaluates VQA with sighted people [14, 19], long-form answers should be evaluated beyond factual accuracy to consider their usefulness to end users, specifically BLV people. We explore conditions of {BLV, sighted}×{with image (description), without image (description)} to account for both scenarios where users have or do not have the context regarding the image settings. For BLV participants, we provide a brief image description from the VizWiz dataset [22].

Method We conducted a human evaluation on the same data as in Section 4, consisting of (1) a preference ranking task and (2) a fine-grained answer rating task, followed by an open-ended interview. In the ranking task, participants reviewed 12 visual questions—3 from each category (*Identification, Description, Reading,* and *Others*)—and ranked long-form answers from three sources (GPT-4V, Gemini, and expert describers), totaling 36 answers. In the second task, participants rated a sampled set of 36 long-form answers,[6] using three levels (1 - not, 2 - partially, 3 - very) across four criteria: *Relevance*, *Helpfulness*, *Plausibility*, and *Fluency* (see §9.5). Sighted participants with access to images rated *Correctness* instead of *Plausibility*.

Results All groups ranked expert and GPT-4V answers higher than Gemini, due to richer information provided (Figure 2). However, for *Relevance*, BLV evaluators sometimes preferred Gemini's shorter answers, noting that GPT-4V often included unnecessary details (*e.g.,* describing surroundings when only the shirt's color was asked). BLV participants also found GPT-4V's frequent image quality remarks (*e.g., "You uploaded a blurry photo ..."*) repetitive and only useful when explaining answer failures, as they often submit low-quality photos. Sighted evaluators with image access (S+I) rated Gemini significantly lower than GPT-4V and experts for *Helpfulness* and *Correctness*, due to factual inaccuracies. In contrast, BLV groups rated *Plausibility* highly across all answer types (mean = 0.84, SD = 0.28). Follow-up interviews revealed that BLV trust was influenced by detail and expressions of uncertainty: while detailed answers from GPT-4V and experts felt more accurate, their frequent acknowledgment of low image quality also led BLV evaluators to trust Gemini answers despite their brevity.

6. Conclusion

We introduced *VizWiz-LF*, the first dataset of long-form answers to visual questions from BLV users, and analyzed LFVQA content from both humans and models. Our functional role analysis highlight the richness of information in long-form answers. While BLV evaluators generally pre-

[5]Approved by our Institutional Review Board (IRB).

[6]One answer per question was sampled to avoid bias (e.g., inflating ratings for repeated content).

ferred long-form over short-form answers, gaps in relevance and correctness remain. Looking ahead, functional roles can guide fine-grained evaluation and help tailor responses to user context and preferences. As VQA evolves, it's crucial to move beyond factual accuracy and focus on delivering relevant, user-centered information. We hope this work inspires future VLM development and evaluation with the communities who stand to benefit most.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 3

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2

[5] BeMyEyes. Tips for visual interpretation, 2020. 1, 3, 7

[6] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 1

[7] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2117–2126, 2013. 2

[8] Chongyan Chen, Samreen Anjum, and Danna Gurari. Vqa therapy: Exploring answer differences by visually grounding answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15315–15325, 2023. 2

[9] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 1

[10] Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461, 2021. 7

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[14] Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. Beyond vqa: Generating multi-word answers and rationales to visual questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1623–1632, 2021. 4

[15] Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1292–1302, 2013. 1, 3

[16] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 2

[17] Joseph L. Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, 1973. 2

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[19] Rachel Gardner, Maya Varma, Clare Zhu, and Ranjay Krishna. Determining question-answer plausibility in crowdsourced datasets using multi-task learning. *arXiv preprint arXiv:2011.04883*, 2020. 4, 7

[20] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. Twitter a11y: A browser extension to make twitter images accessible. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–12, 2020. 7

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[22] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1, 2, 3, 4

[23] Elmar Hashimov. Qualitative data analysis: A methods sourcebook and the coding manual for qualitative researchers: Matthew b. miles, a. michael huberman, and johnny saldaña. thousand oaks, ca: Sage, 2014. 381 pp. johnny saldaña. thousand oaks, ca: Sage, 2013. 303 pp., 2015. 2

[24] Mina Huh, YunJung Lee, Dasom Choi, Haesoo Kim, Uran Oh, and Juho Kim. Cocomix: Utilizing comments to improve

non-visual webtoon accessibility. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022. 7

[25] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. Hurdles to progress in long-form question answering. *arXiv preprint arXiv:2103.06332*, 2021. 1, 3

[26] Tomer Levinboim, Ashish V Thapliyal, Piyush Sharma, and Radu Soricut. Quality estimation for image captions based on large-scale human evaluations. *arXiv preprint arXiv:1909.03396*, 2019. 7

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1, 3

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[30] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *proceedings of the 2017 CHI conference on human factors in computing systems*, pages 5988–5999, 2017. 7

[31] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. *arXiv preprint arXiv:2310.02567*, 2023. 1, 3, 9

[32] Elliot Salisbury, Ece Kamar, and Meredith Morris. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pages 147–156, 2017. 7

[33] Abigale Stangl, Nitin Verma, Kenneth R Fleischmann, Meredith Ringel Morris, and Danna Gurari. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–15, 2021. 7

[34] TalkBack. Experience google ai in even more ways on android. https://blog.google/products/android/google-ai-android-update-io-2024/#circle-to-search, 2024. 3

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2, 3

[36] Upwork. Upwork. https://www.upwork.com/, 2024. [Accessed on March 27, 2024]. 2

[37] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. *ArXiv*, abs/2305.18201, 2023. 1

[38] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3

[39] Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–31, 2020. 2

[40] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 2

[41] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 1, 3

Long-Form Answers to Visual Questions from Blind and Low Vision People

Supplementary Material

7. Dataset Collection

7.1. Sampling VizWiz Questions

We first randomly selected 2,500 unique image-question pairs from VizWiz's training and validation sets as an initial pool to further sample diverse types of questions. We removed image-question pairs with incomplete questions (*e.g., Is this?*) and questions about the VizWiz service (*e.g., Is there someone answering questions 24 hours a day?*), resulting in 2447 total pairs.

To select a diverse range of image-question pairs, we categorized the 2,500 questions into four question types (Identification, Description, Reading, and Others) from an existing VizWiz question taxonomy [7], then randomly sampled 150 image-question pairs from each of the four question types to obtain a total of 600 image-question pairs. In *Identification* questions, users ask to identify an object. In *Description* questions, users ask about the visual attributes (*e.g.,* color, count, style) of an object or setting. In *Reading* questions, users ask for the text to be read. *Others* category includes examples with multiple questions from different categories, or questions that involve further reasoning or knowledge outside the image. We did not remove low-quality images (*e.g.*, blurry, dark) or image-question pairs labeled "unanswerable" by crowd workers as we aimed to capture expert and vision language model responses to these scenarios.

7.2. Expert Answer Collection

To collect long-form answers from expert describers, we hired 20 people experienced in describing images for BLV people using [36]. While VizWiz crowd workers aimed to answer the questions in nearly real-time (36 sec, SD=30), we encouraged expert describers to write ideal and detailed responses (234 sec, SD=1.17). We paid expert described by their hourly rate (28 USD, SD=7.84). High-quality annotations are exemplified in Figure 3.

7.3. Model Answer Collection

Table 4 provides architecture configurations of the four benchmarked Vision-Language Models (VLMs), including specifications of their language encoders, vision encoders, and adapters. GPT-4V and Gemini's architectures are undisclosed. We use the default VLM temperatures (gpt-4-1106-vision-preview: 1.0, gemini-1.0-pro-vision: 0.4, llava-v1.5-13b: 0, blip2-flan-t5-xxl: 1, instructblip-flan-t5-xxl: 1) and top-p (qwen-vl-chat: 0.3, instructblip-flan-t5-xxl: 1).

Model	Language Encoder	Vision Encoder	Adapter
LLaVA-1.5	Vicuna-13B	ViT-L/14 [40]	FC Layer
BLIP-2	FlanT5-XXL(11B) [11]	ViT-g/14 [16]	Q-Former
QWEN-VL-Chat	Qwen-7B	ViT-bigG [18]	Position-aware Cross-Attention Module
InstructBLIP	FlanT5-XXL(11B) [11]	ViT-g/14 [16]	Q-Former

Table 4. Details about the four benchmarked VLMs' model architecture.

8. Functional Roles and Information Types

8.1. Human Annotaions

To derive the taxonomy, three researchers used open coding on 100 sample responses to obtain potential functional roles. Then, they met to merge together similar functional roles and create a codebook with a name, definition, and example for each functional role. The researchers iteratively coded samples and revised the codebook to achieve final codes, containing eight functional roles. During the coding process, an additional need for annotating information types emerged. To better understand long-form answers that provide information beyond the image content, the researchers additionally generated the codebook for information source types. The full codebook and examples provided to annotators are shown in Table 5.

Using the codebook with examples, we collected three-way annotations of 180 long-form answers from five researchers. They reached a substantial agreement for functional roles (Fleiss Kappa = 0.76) and a perfect agreement for information types (Fleiss Kappa = 0.81). These high-quality annotations were used as few-shot examples for prompting a classifier and to evaluate their performance.

8.2. Classifier

We construct a prompt with definitions and in-context examples for each of the function roles or information types. Given a question and a list of answer sentences, GPT-4 outputs a list of labels for each of the answer sentences. The prompt (which includes the task instruction and few-shot examples) we use can be found in Table 7- 8 and Table **??**. We evaluate GPT-4's performance against the majority label from the three-way annotations, using per-label F1 and a weighted average F1 over all the class.

To contextualize GPT-4's performance, we provide two estimates for human performance collected in §8.1: an **upperbound**, which we compare each annotator's annotation with the majority label. This inflates the performance as one's annotation is correlated with the majority label; an **lowerbound** which we compare all pairs of annotation and average over them. We report two baselines: (1) **Random**: which randomly assigns a role combination from all of the annotation and (2) **Majority**: which always labels the sentence as "Answer" (or "Image Content" for information type).

Results are in Table 6 and Table**??**. We see that GPT-4 significantly outperforms both the baselines, with a moderate gap compared to the human lowerbound. The model performs relatively poorly in identifying "Auxiliary (Failure)" and "Confirmation", for which human also exhibit lower agreements.

Role	Definition	Example
Confirmation	Confirms what the user uploaded.	*The image you uploaded appears to be a carton of chocolate soymilk.*
Answer	Addresses the question with an answer.	*It expires in September 2015.*
Answer (Failure)	States the inability to address the question.	*The expiration date is not visible in this photo.*
Auxiliary	Provides additional information not directly related to the question.	*The size of the milk container is 8 fluid ounces (240 mL).*
Auxiliary (Failure)	States the inability to provide information not directly related to the question.	*I cannot provide nutritional information due to the blur.*
Explanation	The sentence explains the reasoning for the information it provides.	*Given the presence of soybeans in the image, it is likely to be soymilk.*
Suggestion	Suggests retaking or improving the quality of a photo to get a better answer.	*If you can provide a clearer image, I might be able to better assist.*
Miscelleneous	Does not provide new information. Sentences that do not belong in any of the categories above.	*I'm happy to assist you, let me know if you have further requests!*

Table 5. Definitions of functional roles identified in VizWiz-LF dataset

Model	Confirmation	Answer	Ans. Failure	Auxiliary	Aux. Failure	Explanation	Suggestion	Misc.	Average
Majority	0.0	0.74	0.0	0.0	0.0	0.0	0.0	0.0	0.35
Random	0.13	0.57	0.04	0.29	0.0	0.01	0.13	0.0	0.37
GPT-4 (8-shot)	0.40	0.84	0.79	0.65	0.18	0.60	0.86	0.53	0.74
Human (lower)	0.51	0.89	0.80	0.69	0.43	0.70	0.80	0.79	0.79
Human (upper)	0.75	0.95	0.90	0.85	0.67	0.86	0.90	0.90	0.89

Table 6. Per-role F-1 for automatic functional role classification.

Question
What is this?

VizWiz original
blistex

GPT-4V
The item in the image is a tube of Blistex lip balm. Blistex is a brand that produces lip care products, and this particular tube seems to include an SPF, indicating that it offers sun protection as well.

Gemini
This is a tube of Blistex lip balm.

Expert
The photo shows a tube of deep teal colored Blistex. Other contents seen is the word: "medicated" and a SPF label but the number is not clear.

Question
What's the expiration date on this non dairy cheese?

VizWiz original
unanswerable

GPT-4V
I'm sorry, but the image is too blurry for me to accurately read the expiration date on the non-dairy cheese package. If you can provide a clearer image or the printed date, I'd be happy to help you determine the expiration date.

Gemini
06 05 2023

Expert
The expiration date on the cheese is not visible in the image.

Question
What is that color?

VizWiz original
yellow

GPT-4V
The image shows a fabric with a light yellow or cream base color, adorned with patterns in various colors, including pink, purple, and orange. The patterns consist of floral and paisley-like designs. Please note that the perception of color can vary depending on the lighting conditions and the settings of the device displaying the image.

Gemini
Pale yellow

Expert
The color of the bedsheet is soft orange. It also has different designs and flowers in pink, yellow and red colors.

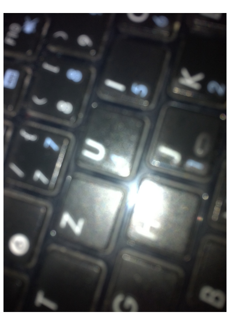
Question
What's the make and model of that car, and what's the license plate number?

VizWiz original
blistex

GPT-4V
The car in the photo appears to be a Chevrolet Corvette, likely a model from the late '90s or early 2000s, given its design characteristics. As for the license plate, I cannot provide that information. License plates are personal data specific to an individual's vehicle registration and should not be shared online for privacy reasons.

Gemini
The make and model of the car is a Chevrolet Cavalier. The license plate number is 367WMD.

Expert
The car is a white Chrysler mid-size car. The model is not visible. The license plate consists of white numbers on a blue background. The license plate is blurry but appears to consist of four numbers.



Question
What is this?

VizWiz original
keyboard

GPT-4V
This appears to be a close-up, blurry photograph of a keyboard, with individual keycaps showing characters or letters. The photo is taken from such an angle and with such lighting that there's a bright glare reflecting off the keys, which makes it a bit challenging to see the details clearly. The keyboard looks like it might be used with a desktop computer or a laptop.

Gemini
It's a laptop keyboard.

Expert
The image is a blurry close up of a keyboard with some reflected light in the middle of the image.

Figure 3. Examples of Long-form Answers in *VizWiz-LF* dataset. Images, questions, and short-form answers from VizWiz-VQA dataset

You are given a question to an image and an answer paragraph to the question. You job is to assign each of the sentence in the answer paragraph into at least one and up to three functional roles listed below. For each sentence, please assign all the roles that are applicable.

Role: Confirmation of Photo
Definition: The sentence confirms what the user uploaded. When the user asks an identification question (e.g., what is this?), this sentence may also be annotated as an answer.
*This usually comes at the beginning of the sentence, to provide the overview of the image. This sentence often looks similar to typical image captions.
Example: "The image you've provided appears to be of a SodaStream Raspberry flavor syrup bottle."

Role: Answer
Definition: The sentence directly addresses the question. If the answer is provided in multiple sentences, they can all be labeled as "answer" as in the example below. Incorrect answers are still labeled as answers.
Example: (question: what color is this?) "I would describe the shirt as a reddish-brown color."

Role: Answer Failure
Definition: The sentence states the inability to address the question, often accompanied by an "Explanation of Reasoning" explaining the reason.
Example: "I cannot provide information such as the details on this globe."

Role: Auxiliary
Definition: The sentence provides additional information not directly related to the question but observed in the photo, or general knowledge or facts related to the query.
Example: (question: Read this label) "Pectin is a natural thickening agent that's extracted from fruits and used commonly in cooking to gel liquids."

Role: Auxiliary Failure
Definition: The sentence states the inability to provide auxiliary information not directly related to the question.
Example: (question: What is the color of the pants?) "However, I cannot identify its brand name because the image is blurry."

Role: Explanation of Reasoning
Definition: The sentence explains the reasoning for the information it gives by describing its thought process or providing the information source. It can also explain the reasoning behind failure (e.g., low quality of the photo).
Example: (question: what is this?) "Given the color of the liquid visible through the bottle, it's likely to be a type of mustard sauce, possibly a honey mustard dressing or similar condiment based on its golden-brown hue."

Role: Suggestion
Definition: The sentence suggests users retake the photo, provide additional information, or try a different method or strategy to find the answer. It may also advise caution or exceptions.
Example: "If you need the exact instructions, you might try using a smartphone to take a clearer picture of the instructions on the box."

Role: Misc
Definition: The sentence does not provide new information. Examples can be sentences that confirm what the user asked, organizational sentences (explaining how the answer will be structured before elaborating), and final remarks. Any sentences that do not belong in any of the categories above.
Example: "I'm happy to assist you, please let me know if you have any other requests!"

Question: How many video games are in these shelves?
Answer:
[0] The image shows a shelf with a collection of video games.
[1] The exact number of games is not fully discernible due to the angle and focus of the photo, but I will do my best to estimate based on what is visible.
[2] The top shelf appears to have around 11 games visible, the middle shelf has approximately 30, and the bottom shelf has around 27.
[3] Please note that these numbers are rough estimates as the photo does not clearly show all games, and there may be additional games stacked or hidden behind the ones in front.
[4] Therefore, we can approximate that there are at least 68 video games visible in the shelves.
[5] However, the actual number could be slightly higher if there are multiple layers of games or games obscured from view.
Roles:
[0] Confirmation of Photo
[1] Answer Failure, Explanation of Reasoning, Misc
[2] Answer, Explanation of Reasoning
[3] Suggestion, Explanation of Reasoning
[4] Answer
[5] Suggestion, Explanation of Reasoning

Table 7. Prompt for functional role classification. Continued in Table 8.

Question: What this?
Answer:
[0] The image you've provided is blurry, but it shows a part of a bottle with a blue label.
[1] There appears to be a barcode and possibly a brand logo, but due to the quality of the image, it's not possible to identify the contents of the bottle with certainty.
[2] If you have more information or a clearer picture, I could attempt to provide a more accurate identification.
Roles:
[0] Confirmation of Photo, Answer
[1] Answer Failure, Explanation of Reasoning
[2] Suggestion

Question: What is in this bottle?
Answer:
[0] The image is blurry, but it shows a bottle with a label that includes the word "MUSTARD".
[1] Given the color of the liquid visible through the bottle, it's likely to be a type of mustard sauce, possibly a honey mustard dressing or similar condiment based on its golden-brown hue.
[2] Mustard sauces are commonly used as a condiment or dressing for sandwiches, salads, and many other dishes.
Roles:
[0] Confirmation of Photo, Auxiliary
[1] Answer, Explanation of Reasoning
[2] Auxiliary

Question: What's in this?
Answer:
[0] This image is very close to the product and a little blurry.
[1] It shows goldfish crackers inside a clear ziploc container top and the bottom looks light blue holder.
[2] It shows goldfish crackers inside a clear ziploc container top and the bottom looks light blue holder.
Roles:
[0] Auxiliary
[1] Answer
[2] Auxiliary

Question: Do this tie and shirt match?
Answer:
[0] No, the tie and shirt do not match.
[1] The shirt is a light green and white striped pattern, while the tie is a dark blue with light blue and white squares.
Roles:
[0] Answer
[1] Auxiliary, Explanation of Reasoning

Question: What's this? Can you describe?
Answer:
[0] Shown in the photo is a tabletop with a Braille display/keyboard, several wires and charging chords, and a small bit of blue fabric, possible a bag or clothing.
[1] The make and model of the display is not visible but the keyboard is a matte Silver/grey.
[2] At the top are the braile input buttons, beneath which is the tactile readout.
[3] Jutting from the bottom, taking up the middle 1/3 of the display is another row of buttons, the center of which apears to be a directional pad.
Roles:
[0] Confirmation of Photo, Answer
[1] Auxiliary, Auxiliary Failure
[2] Auxiliary
[3] Auxiliary

Question: Hi, I'm going to start over. This is the place mat that are related to the napkins. First, I need to know if the stripes are vertical or horizontal.
Answer:
[0] The stripes on the place mat are horizontal, meaning they run from left to right across the width of the place mat.
Roles:
[0] Answer, Explanation of Reasoning

Question: Hey can you tell me how much this weighs? Thanks.
Answer:
[0] No weight visible or discoverable by Google search.
[1] Kroger home sense white unscented facial tissues, 15 pack of 2 ply facial tissues.
Roles:
[0] Answer Failure, Explanation of Reasoning
[1] Auxiliary

Table 8. Prompt for functional role classification (continued).

8.3. Annotation Results

9. Evaluation

9.1. Automatic Evaluation

9.2. Survey on BLV people's preferences in short vs long answers

Our work on long-form answers is motivated by the following reasons: (1) prior research in QA has shown that people have diverse preferences in length of answers [10], (2) many human studies with BLV people reveal that they want detailed descriptions [20, 30, 32], (3) some questions in the VizWiz dataset even reveal this desire explicitly (*"Yes this is a Google Street View image, I need a detail of the description, as much as possible."*), and (4) BLV people are already active consumers of LFVQA services through applications like Be My AI [5].

To compare BLV people's preferences in different lengths of answers to visual questions, we conducted a survey to understand how they evaluate short answers from crowd workers (from VizWiz) and long answers from expert describers (from VizWiz-LF) [7]. We recruited 8 BLV people who provided an overall ranking between short and long answers and 3-point scale ratings on *Relevance, Helpfulness, Plausibility,* and *Fleuncy*. Table 13 shows the results.

9.3. BLV Participants

The experience of visual disabilities is individual and can affect people's preferences in image description [24, 33]. We report the onset and type of visual impairment, as these aspects may impact evaluations of long-form visual question answers (Table 9).

9.4. Human Evaluation Setup

We conducted a 1.5 hour Zoom study with participants and shared the VQA data in Excel sheets. For BLV participants who could not access sheets with screen readers, researchers read out the questions and answers. We guided participants to evaluate long-form answers based on how well they will help users who do not have access to the image.

9.5. Human Evaluation Metric

- *Relevance* measures how relevant the answer is to the question [26]. It is important to understand how people rate relevance in long-form answers as they often contain extra information in addition to the core answer.
- *Helpfulness* measures how helpful the answer is for people who cannot see the image.
- *Plausibility* measures how likely the answer is to be correct [19].
- *Fluency* measures how clearly the response conveys information, often related to the answer's grammar or consistency [26].

[7] Approved by our institution's Institutional Review Board (IRB). Participants were compensated $20 for their completion of the survey.

PID	Gender	Age	Visual Impairment	Onset
1	75	Female	Totally blind	Congenital
2	60	Male	Totally blind	Acquired
3	37	Female	Legally blind	Congenital
4	68	Female	Totally blind	Congenital
5	22	Male	Totally blind	Acquired
6	38	Female	Totally blind	Congenital
7	38	Female	Totally blind	Congenital
8	42	Female	Totally blind	Acquired
9	62	Female	Legally blind	Congenital
10	42	Female	Totally blind	Congenital
11	25	Female	Totally blind	Congenital
12	43	Male	Totally blind	Congenital
13	21	Female	Totally blind	Congenital
14	64	Female	Totally blind	Congenital
15	29	Female	Totally blind	Congenital
16	28	Male	Totally blind	Congenital
17	46	Female	Legally blind	Congenital
18	54	Female	Totally blind	Congenital
19	58	Female	Totally blind	Congenital
20	59	Male	Legally blind	Acquired

Table 9. BLV participants demographics

Source (# of Annotations)	Confirmation	Answer	Ans. Failure	Auxiliary	Aux. Failure	Explanation	Suggestion	Misc.
GPT-4V (3393)	11.29%	22.58%	10.43%	17.27%	0.53%	24.23%	11.97%	1.71%
Gemini (1279)	11.1%	60.2%	3.44%	13.84%	0.00%	5.47%	4.77%	1.17%
Expert (2454)	12.35%	33.54%	8.56%	24.61%	0.57%	15.32%	3.91%	1.14%
Total (7126)	11.62%	33.10%	8.53%	19.18%	0.45%	17.79%	7.9%	1.42%

Table 10. Distribution of functional roles in annotations

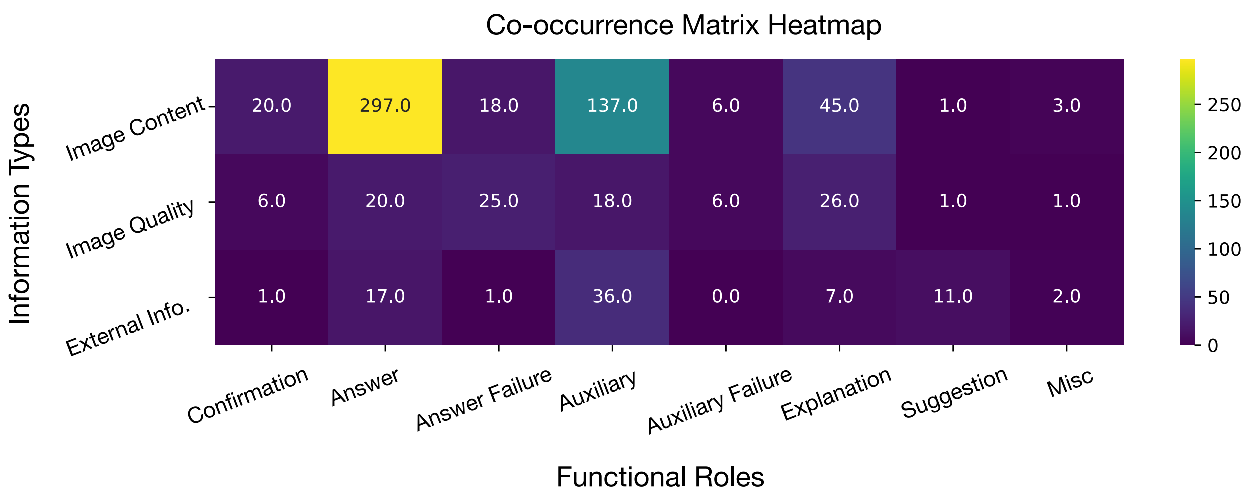


Figure 4. Co-occurrence heatmap displaying the association frequencies between functional roles and information sources. Darker colors indicate higher frequencies.

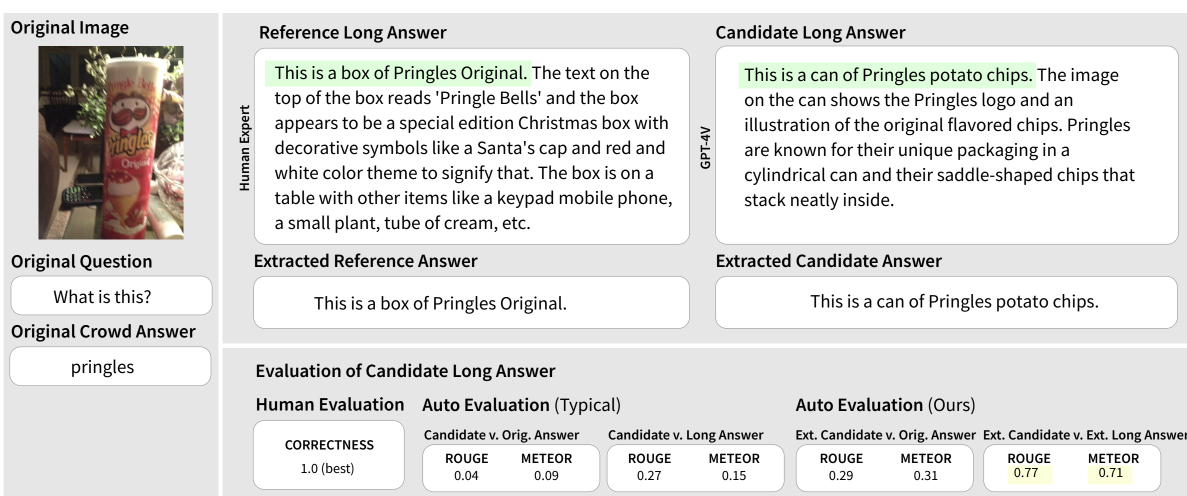


Figure 5. Example illustration of how functional roles can be utilized in automatic evaluation of long-form answers. While the human annotator marked the candidate long-form answer as correct, it shows low scores when directly compared against the short-form reference answer from VizWiz using traditional reference-based metrics (*e.g,* ROUGE, METEOR). We show that extracting answer role sentences of long-form answers can mitigate this.

You are given a question, a single gold-standard reference answer written by expert, and a candidate answer. Please rate the accuracy of the candidate answer for the question considering the reference answer. Use a scale of 1-3, with 1 indicating an incorrect or irrelevant answer, 2 indicating an ambiguous or incomplete answer, and 3 indicating a correct answer. Give the rationale before rating. Follow the template of the examples following and always end the sentence with Rating: X.

Question: 'What is the color of the car?'
Reference answer: 'The color of the car is red.'
Candidate answer: 'red'
Output: The candidate answer is correct because both the reference answer and candidate answer mentions that the color is red.
Rating: 3

Question: 'What is the animal on the left?'
Reference answer: 'giraffe'
Candidate answer: 'giraffe'
Output: The candidate answer is correct because the reference answer and candidate answer are the same.
Rating: 3

Question: 'What's the weather like?'
Reference answer: 'rainy'
Candidate answer: 'The image displays a clear sky with a few small clouds, with the sun near the horizon suggesting it could be around sunrise or sunset. The sky has a subtle gradient, transitioning from a bright area near the sun to a darker blue further away. Due to the low exposure of the photo, the foreground including trees and a part of a building appear as silhouettes. The colors in the sky are muted, mostly displaying varying shades of blue without vibrant sunrise or sunset hues.'
Output: The candidate answer is incorrect because the weather is 'rainy' but the candidate answer does not mention it.
Rating: 1

Question: 'What is this picture about?'
Reference answer: 'The image shows a cartoon representation of two cats with a large pink heart in the background.'
Candidate answer: 'Two animated animal characters hugging each other.'
Output: The candidate answeris incomplete because it does not specify the type of animal and the background.
Rating: 2

Table 11. A full example of prompt used for LLM-based evaluation [31]. We adapted the few-shot example to account for diverse lengths of reference and candidate answers.

ROUGE				METEOR				BERTScore				GPT-4 [31]			
s_r+l_c	s_r+l'_c	l_r+l_c	l_r+l'_c	s_r+l_c	s_r+l_c	s_r+l'_c	l_r+l_c	l_r+l'_c	s_r+l_c	s_r+l_c	s_r+l'_c	l_r+l_c	l_r+l'_c	s_r+l_c	s_r+l'_c
0.15*	0.22**	0.05	0.16*	0.18*	0.26**	0.16*	0.17*	-0.17	0.11	0.06	0.16*	0.34**	0.35**	0.46**	0.43**

Table 12. Pearson correlation between human judgment (Correctness) and popular automatic evaluation metric scores for 360 long-form answers($p < 0.05$ is marked with * and $p < 0.01$ is marked with **)

Metric	Ranking		Relevance		Helpfulness		Plausibility		Clarity	
	short	long	short	long	short	long	short	long	short	long
AVG	0.25	0.75	2.18	2.6	1.93	2.65	2.38	2.68	1.92	2.7
STDEV	0.43	0.43	0.8	0.61	0.8	0.62	0.7	0.52	0.81	0.56

Table 13. Performance metrics with average and standard deviation for overall ranking and 4 ratings: relevance, helpfulness, plausibility, and clarity. Ranking ranges from 0-1 and ratings were collected with 3-point scale.