

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

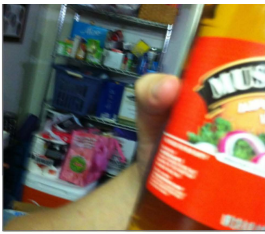
VizWiz-LF Dataset

Our dataset *VizWiz-LF* contains **4.2k** long-form answers (avg. **30 words**) to visual questions, collected from **human expert** describers and **6 VQA models**.

Annotation

Visual Question

What is in this bottle?



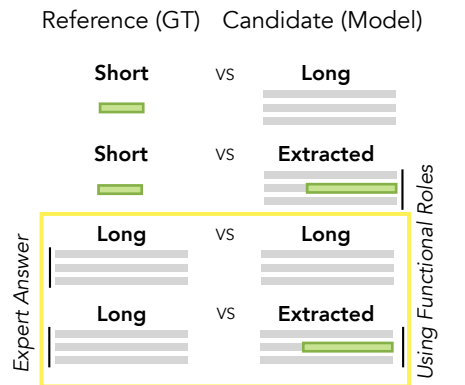
Short Answer

Tomato sauce

Long-Form Answers

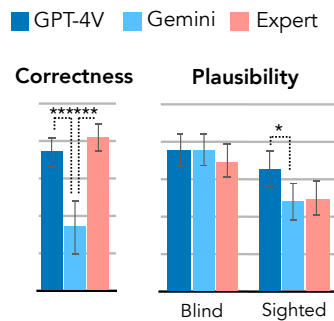
	Functional Roles	Information Source
GPT-4V	confirmation	image content, image quality
	answer, explanation	image content
	auxiliary information	external
	suggestion	
Gemini	answer	image content
Human	answer failure	image content
	answer, explanation	image content
	answer	image content

Automatic Evaluation



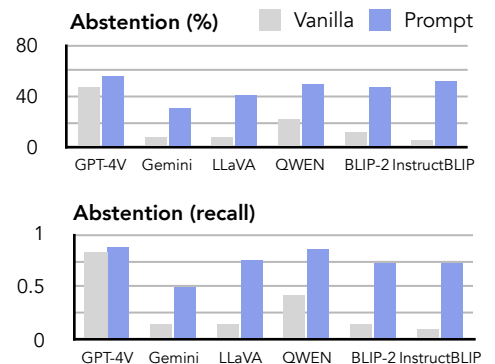
- * Evaluation using **long reference answers** shows a better correlation with human
- * **LLM metric** > ROUGE/METEOR/BERTScore

Human Evaluation



- * Model answers hallucinate with incorrect visual details, but **blind people often perceive them as plausible**
- * Sighted people's evaluation is not a strong proxy for blind people

Abstention Experiment



- * We can use **functional roles** to check if a model's long-form answer abstained.
- * **Prompting models with abstain instructions** can reduce hallucinations.