

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



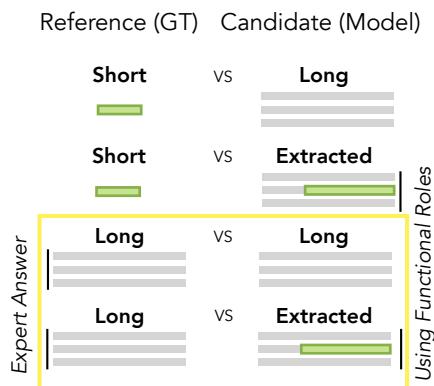
Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, Amy Pavel

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**.

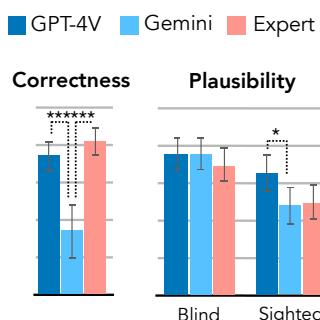
| Annotation | Long-Form Answers | Functional Roles | Information Source |
|---|--|--|--|
| Visual Question What is in this bottle? | <p>GPT-4V</p> <p>The image is blurry, but it shows a bottle with a label that includes the word MUSTARD.</p> <p>Given the color of the liquid visible through the bottle, it's likely to be a type of mustard sauce [...]</p> <p>Mustard sauces are commonly used as a condiment or dressing for sandwiches, salads, and many other dishes.</p> <p>If you have more information or a clearer picture, I could attempt to provide a more accurate identification.</p> | confirmation answer explanation auxiliary information suggestion | image content image quality image content external image content |
| Short Answer Tomato sauce | <p>Gemini</p> <p>This bottle contains Heinz Distilled White Vinegar.</p> | answer | image content |
| | <p>Human</p> <p>I can't tell what is in the bottle.</p> <p>If I had to guess, I would say it was some sort of cooking oil because it has an amber color and pictures of vegetables [...]</p> <p>It could also be an apple juice or another type of juice.</p> | answer failure answer explanation answer | image content image content image content |

Automatic Evaluation



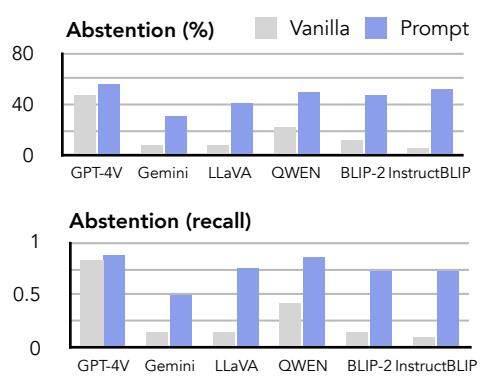
- * Evaluation using **long reference answers** shows a better correlation with human
- * **LLM metric** > ROUGE/METEOR/BERTScore

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.