Supplementary Materials for:
Pointing out Human Answer Mistakes in a Goal-Oriented Visual Dialogue

Ryosuke Oshima[1] Seitaro Shinagawa[2] Hideki Tsunashima[1] Qi Feng[1] Shigeo Morishima[3]
[1]Waseda University [2]Nara Institute of Science and Technology
[3]Waseda Research Institute for Science and Engineering
{ryosukeoshima@fuji, h.tsunashima@asagi, fengqi@ruri}.waseda.jp, sei.shinagawa@is.naist.jp, shigeo@waseda.jp

1. Question Type

Question type	Example sentence
Spatial	On the right side half?
Object	Is it a car?
Color	Is it white?
Action	Are they wearing jeans?
Size	A small one?
Super-category	Is the object electronic?
Texture	Is it made of metal?
Shape	Is it a round container?
Others	Is it edible?

Table 1. Example sentences of each question type.

Table 1 shows example sentences for each question type. Question types are labeled using the keyword matching method, categorizing question types using keywords like *left* for the Spatial category. Super-category is a higher-level group containing related sub-categories, organizing information hierarchically, such as fruit for banana or vehicle for car.

2. Question Types Analysis

Hypothesis Testing Our hypothesis is that *the rate of human answer mistakes varies by question type*. Therefore, if we can establish that any question type is significantly more likely to be incorrect than another, it would support our hypothesis. Since we are focusing on the difference in the proportion of mistakes in each question type, we conducted hypothesis testing for the difference in the Population proportions.

Moreover, it's important to note that not all answers are provided by the same person; instead, crowd workers contribute answers. As we are dealing with 9 question types, this constitutes an independent test involving more than two groups. As a result, we conducted Fisher's exact rate test. We set the significance level at 1% and performed an upper one-tailed test. The resulting p-value is $0.0004998 (< 0.01)$, which supports our hypothesis that human answer mistakes vary by question type.

Trend of mistaken Question Types Humans are more likely to make mistakes when answering Spatial, Color, Action, and Size questions. One possible reason is that these question types are more difficult than the other question types (Shape, Texture, Object, and Super-category). Most Shape and Texture questions are easy-to-understand, such as "round," "square," "wooden," and "metal." Most Object and Super-category questions are monotonous, such as 'Is it a banana?' Answerer could understand the meaning of the questions by reading a single word without considering it as a question sentence. In contrast, the mistake rates for Action and Spatial questions were higher because Answerer needs to take these questions as sentences, which means they are more difficult to understand.

3. QA Turn Analysis

231 out of 431 samples of answer mistakes occurred in the last turn. This trend is possibly related to a bias on the problem setting. While a mistake in a dialogue would be pointed out instantly, a mistake in the last turn has no chance to be recovered. The former mistake cases would be included *Success*, not *Failure* collection used in our Human Mistake Dataset.

4. Pre-training with Synthetic Dataset

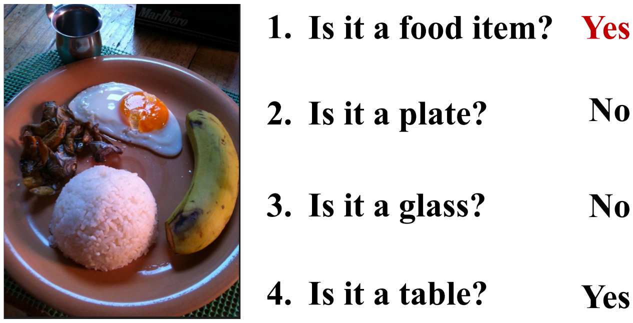
Learning method	Same image	Different image
Human mistake	0.730	0.368
Synthetic + Human mistake	**0.811**	**0.482**

Table 2. The results of each learning method. The score is F-score. Same image and different image mean the results of the same image dataset and the different image dataset, respectively.

We conducted the experiment to investigate the effectiveness of pre-training with Synthetic Dataset in addressing the challenge of limited erroneous human answers in training data. Specifically, we compared two learning methods: (1) training the model solely with Human Mistake Dataset, and (2) pre-training the model with Synthetic Dataset and fine-tuning it using Human Mistake Dataset. Due to fewer positive examples (i.e., human answer mistakes) than negative examples (i.e., correct answers), we oversampled the training data during the fine-tuning process.

Table 2 presents the experiment's results. The pre-training with Synthetic Dataset in this study achieved the highest F-score. This indicates that the pre-training strategy, which employs the Synthetic Dataset even in the absence of human mistakes, is effective.

5. Unnatural Dialogue Flow



1. Is it a food item? Yes

2. Is it a plate? No

3. Is it a glass? No

4. Is it a table? Yes

Figure 1. An example of unnatural dialogue flow in Synthetic Dataset, where the first answer is incorrect.

Figure 1 shows an example of unnatural dialogue flow. Although it is confirmed that the topic is food at time 1, the question '*Is it a plate?*' is asked at time 2. In a natural dialogue, Answerer is expected to ask questions about food, such as *Is it a banana?*.

6. Evaluation for Imbalanced Data

Accuracy $\left(= \frac{\text{correct answers}}{\text{sample size}}\right)$ is not an appropriate evaluation metric in imbalanced data. The reason is that if the model learns to predict negative cases with many samples, the model can achieve high accuracy, even though it cannot predict positive cases at all. Instead, we use the F-score defined by Recall and Precision. It serves as one of the evaluation metrics for imbalanced data. The true positive (correctly predicting a positive label as positive), true negative (correctly predicting a negative label as negative), false positive (incorrectly predicting a negative label as positive), and false negative (incorrectly predicting a positive label as negative) can be used to calculate Recall and Precision as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{1}$$

$$\text{Precision} = \frac{\text{True Postives}}{\text{True Postives} + \text{False Postives}} \tag{2}$$

In imbalanced data, if a model simply learns to predict many instances as "positive," the Recall will be high while the Precision will be low. Conversely, if the model learns to predict many instances as "negative," the Precision will be high while the Recall will be low.

The F-score is used to appropriately evaluate such models, as it represents the harmonic mean of Recall and Precision. This metric considers both Recall and Precision, providing a more balanced assessment of the model's performance in an imbalanced dataset.

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

7. Model Details

We describe below the details of the baseline model. We get the embedding of the whole image I_{emb} and that of the target object's cropped image S_{emb} using ResNet [4]. We also get the embedding of the target object's spatial information $x_{spatial}$ from the bounding box, following [3].

$$\begin{aligned} x_{spatial} = [&x_{min}, y_{min}, x_{max}, y_{max}, \\ &x_{center}, y_{center}, w_{box}, h_{box}] \end{aligned} \tag{4}$$

Next, I_{emb}, S_{emb}, $x_{spatial}$, and q_t^{emb}, the embedding of the question q_t at time t encoded by LSTM [5], are concatenated and passed through MLP layers to generate a semantic vector q_{mean} for each question.

$$q_{mean} = \text{MLP}_m \left(\left[I_{emb}; S_{emb}; x; q_t^{emb}\right]\right) \tag{5}$$

$[\cdot;\cdot]$ denotes vector concatenation. We obtain the probability that the human interlocutor's answer is incorrect p_m:

$$p_m = \text{sigmoid}\left(\text{MLP}_c\left(\left[q_{mean}; a_t^{emb}\right]\right)\right) \tag{6}$$

Finally, we determine whether an answer is correct or incorrect by a threshold value.

$$\begin{cases} 1 & \text{(Incorrect answer)} \quad p_m > 0.5 \\ 0 & \text{(Correct answer)} \quad p_m \leq 0.5 \end{cases} \tag{7}$$

We do not add the category label of the target object, such as a dog or banana, because the model relies on the category label and its spatial information instead of the visual features of the image as [7] mentioned.

Figure 2, 3 shows the Question type model and QA turn model, respectively.

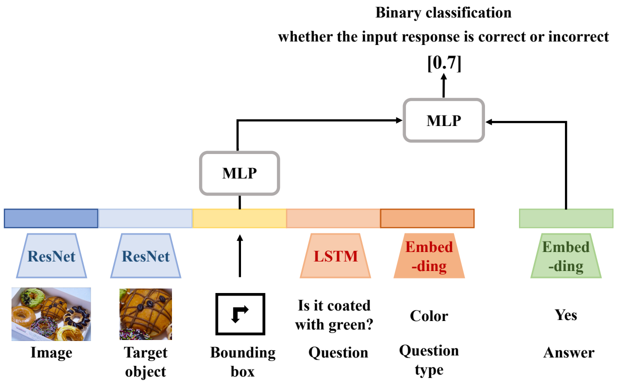


Figure 2. Question type model

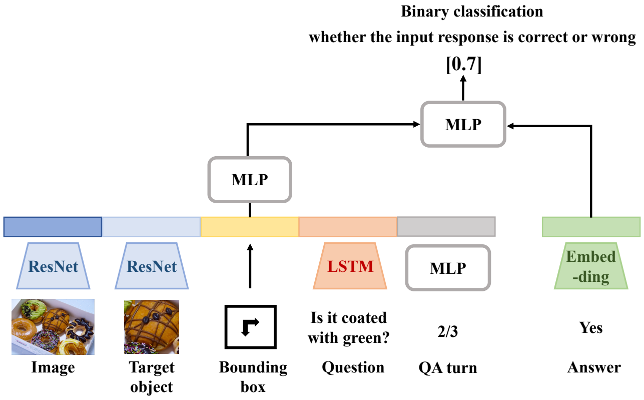


Figure 3. QA turn model

We get the embedding of question type q_{type}^{emb} and QA turn a_{time}^{emb} by the embedding layer and the MLP layers, respectively. We obtain q_{mean} using equation (8) for the Question type model and equation (9) for the QA turn model.

$$q_{mean} = \text{MLP}_m\left(\left[I_{emb}; S_{emb}; x; q_t^{emb}; q_{type}^{emb}\right]\right) \quad (8)$$

$$q_{mean} = \text{MLP}_m\left(\left[I_{emb}; S_{emb}; x; q_t^{emb}; q_{time}^{emb}\right]\right) \quad (9)$$

8. VLM Experiment Details

Evaluation We only evaluated OpenFlamingo [1] using the different image dataset to ensure a fair comparison with the MLP model. This was necessary because OpenFlamingo was not specifically tuned with the exact same image dataset as the MLP model.

Inputs and Prompts We used examples of actual human mistakes for the few-shot prompting, rather than a sample from synthetically created mistakes. In particular, we randomly sampled eight examples from the same image dataset used for fine-tuning the MLP model. We conducted preliminary experiments in an input format similar to MLP, where the object's crop image and the bounding box's position were given in list format, but the F-score was very low. We provided the target object information with the object's position in the image surrounded by a yellow bounding box and the object's name.

OpenFlamingo does not learn the unnatural dialogue flow by including the dialogue history, because it is not pretrained with the Synthetic Dataset. We conducted experiments with both prompts with and without dialogue history. Figure 4 shows an overview of the eight few-shot prompts added as input to OpenFlamingo. Table 8 shows the prompts for each prompt type. We provided the prompts in a structured JSON-like format.

Other VLM Models We also conducted experiments with Instruct BLIP [2] and BLIP2 [6] as well as OpenFlamingo. However, with and without Instruction following, the F-score was very low, about 25%. We think this is because in-context learning did not work well, as BLIP2 was trained with the pre-training dataset, which only contains a single image-text pair per sample, as mentioned in [6]. OpenFlamingo trained with MMC4 [8], which includes documents sourced from web scraping, interleaved images, and text, with multiple image-text combinations in each sequence.

9. How to Use Each Dataset

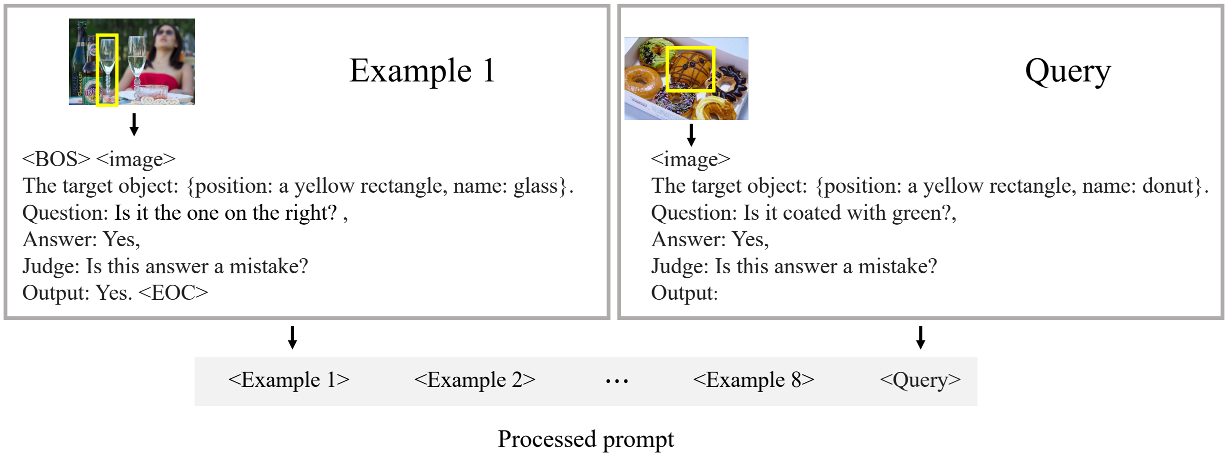
Table 4 shows how we use Synthetic Dataset, the same image dataset, and the different image dataset when we conduct MLP models' experiments.

10. Supplement of Results

Table 5, 6 show F-score, Recall, and Precision in the MLP model's experiment. Table 7 shows the results of OpenFlamingo's experiment.

Type	Prompt
Normal	<BOS> <image> The target object: {position: a yellow rectangle, name: {object category}}, Question: {question text}, Answer: {answer text}, Judge: Is this answer a mistake? Output: {answer}.<EOC>
Qtype	<BOS> <image> The target object: {position: a yellow rectangle, name: {object category}}, Question: {question text}, Answer: {answer text}, This question type: {qtype} Hint: <spatial>, <color>, <action>, and <size> questions are easy to make mistakes on. Judge: Is this answer a mistake? Output: {answer}.<EOC>
Time	<BOS> <image> The target object: {position: a yellow rectangle, name: {object category}}, Question at {answer time} progression of dialogue: {question text}, Answer: {answer text}, Hint: The frequency of answer errors increases as answer time is bigger. Judge: Is this answer a mistake? Output: {answer}.<EOC>
Normal (history)	<BOS> <image> The target object: {position: a yellow rectangle, name: {object category}}, Dialogue history: {history}, Question: {question text}, Answer: {answer text}, Judge: Is this answer a mistake? Output: {answer}.<EOC>
Qtype (history)	<BOS> <image> The target object: {position: a yellow rectangle, name: {object category}}, Dialogue history: {history}, Question: {question text}, Answer: {answer text}, This question type: {qtype} Hint: <spatial>, <color>, <action>, and <size> questions are easy to make mistakes on. Judge: Is this answer a mistake? Output: {answer}.<EOC>
Time (history)	<BOS> <image> The target object: {position: a yellow rectangle, name: {object category}}, Dialogue history: {history}, Question at {answer time} progression of dialogue: {question text}, Answer: {answer text}, Hint: The frequency of answer errors increases as answer time is bigger. Judge: Is this answer a mistake? Output: {answer}.<EOC>

Table 3. Examples of prompts for each type. <image> takes as input the embedding of the image. {object category} is the object category name of each target object (e.g., donut, vase), {question text} and {answer text} is the question and answer (yes or no) of the corresponding part to judge whether it is a mistake or not, and {answer} is the result of judging whether the corresponding response is a mistake or not. {qytpe} contains the question type (e.g., <color>) and {answer time} contains the value of $\frac{\text{current turn}}{\text{total turns}}$.



<BOS> <image>
The target object: {position: a yellow rectangle, name: glass}.
Question: Is it the one on the right? ,
Answer: Yes,
Judge: Is this answer a mistake?
Output: Yes. <EOC>

<image>
The target object: {position: a yellow rectangle, name: donut}.
Question: Is it coated with green?,
Answer: Yes,
Judge: Is this answer a mistake?
Output:

Figure 4. Few-shot prompt overview diagram.

Dataset	Pre-training		Fine-tuning		Test
	Train	Validation	Train	Validation	
Synthetic	75%	15%	-	-	-
Same image	-	-	75% (k-fold cross validation)		25%
Different image	-	-	-	-	100%

Table 4. How to use each dataset in the MLP model experiment. Same image and Different image mean the same image dataset and the different image dataset, respectively. Percentages, such as 75% or 15%, represent how much of each dataset is used.

Learning Method	Same image			Different image		
	F-score	Recall	Precision	F-score	Recall	Precision
Human mistake	0.730	**0.920**	0.605	0.368	0.459	0.308
Synthetic + Human mistake	**0.811**	0.860	**0.768**	**0.482**	**0.541**	**0.434**

Table 5. The detailed results of each learning method in the experiment about pretraining with Synthetic Dataset. The score is F-score, Recall, and Precision.

Model	Same image			Different image		
	F-score	Recall	Precision	F-score	Recall	Precision
Baseline	**0.811**	**0.860**	**0.768**	0.482	0.541	0.434
QA turn	0.718	0.840	0.627	**0.514**	**0.623**	**0.437**
Question type	**0.743**	0.840	**0.667**	**0.527**	**0.639**	**0.448**

Table 6. The detailed results in the MLP model experiment. The score is F-score, Recall, and Precision. The best score is in **black bold**, and the second-best score is in **blue**.

Prompt type	Without history			With history		
	F-score	Recall	Precision	F-score	Recall	Precison
Normal	0.313	0.350	0.283	0.325	0.438	0.259
QA turn hint	**0.374**	**0.463**	**0.314**	**0.377**	**0.538**	**0.291**
Question type hint	**0.366**	**0.438**	**0.315**	**0.372**	**0.500**	**0.296**

Table 7. The detailed results in the OpenFlamingo experiment. The score is F-score, Recall, and Precision. The best score is in **bold black**, and the second-best score is in **blue**.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022.

[2] 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. *arXiv preprint arXiv:2305.06500*, 2023.

[3] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1080–1089. Institute of Electrical and Electronics Engineers, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition*, pages 770–778. Institute of Electrical and Electronics Engineers, 2015.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9:1735–1780, 1997.

[6] 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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023.

[7] Alberto Testoni, Claudio Greco, Tobias Bianchi, Mauricio Mazuecos, Agata Marcante, Luciana Benotti, and Raffaella Bernardi. They are not all alike: Answering different spatial questions requires different grounding strategies. In *Proceedings of the 2020 International Workshop on Spatial Language Understanding*, pages 29–38, Online, 2020. Association for Computational Linguistics.

[8] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.