Supplementary Material

Bram Willemsen and **Livia Qian** and **Gabriel Skantze**
Division of Speech, Music and Hearing
KTH Royal Institute of Technology
Stockholm, Sweden
{bramw, liviaq, skantze}@kth.se

Supplementary material for the paper *"Resolving References in Visually-Grounded Dialogue via Text Generation"*. For code and annotations, we refer the reader to our repository[1].

1 Hyperparameters

Model	davinci (base)
Epochs	10
Batch size	4
Learning rate multiplier	0.1
Prompt loss weight	0.01

Table 1: Hyperparameters for fine-tuning GPT-3 for context windows **3** and **7**.

Max tokens	256
Temperature	0
Top p	1
Presence penalty	0
Frequency penalty	0

Table 2: Hyperparameters for inference GPT-3 for context windows **3** and **7**.

	3	**7**
Model	gpt2-xl	gpt2-xl
Epochs	1	2
Batch size	1	1
Learning rate	5e-6	1e-6

Table 3: Hyperparameters for fine-tuning GPT-2 for context windows **3** and **7**.

Max tokens	128
Temperature	0
Top p	0.9
Repetition penalty	1

Table 4: Hyperparameters for inference GPT-2 for context windows **3** and **7**.

2 Human Evaluation

Additional details of the human evaluation experimental setup.

Independent To ensure independence of observations, we randomly sample one label per image set. This means that each participant is asked to select the corresponding image from the set of candidate images for five labels, each label for a different dialogue associated with a different image set, presented in a random order. We randomly add a sixth item that serves as an attention check. Each participant provides data for only one randomly assigned set of six items. We limit this evaluation to single-image referents. For an example of the task (with instructions) as shown to participants, see Figure 4. Participants were recruited via Amazon Mechanical Turk[2]. Eligible workers were those that had completed at least 1000 previous HITs and a historic completion percentage of at least 97%.

Holistic Detailed task instructions as presented to participants are shown in Figure 2 and Figure 3. An example of the task itself is shown in Figure 4. The order in which the images are presented is randomized. We add an attention check after every 25 mentions. We do not limit this evaluation to single-image referents. Participants are allowed to select more than one image as well as an image that has already been ranked if they believe this to be necessary. Participants were recruited via Prolific[3]. Eligible workers were those that had indicated that they are fluent in English, had a minimum approval rate of 99 and had a minimum of 100 previously completed submissions. The expected time-on-task was adjusted for the number of mentions in the dialogue.

[1] https://github.com/willemsenbram/reference-resolution-via-text-generation, doi:10.5281/zenodo.8176114

[2] https://www.mturk.com/

[3] https://www.prolific.co/

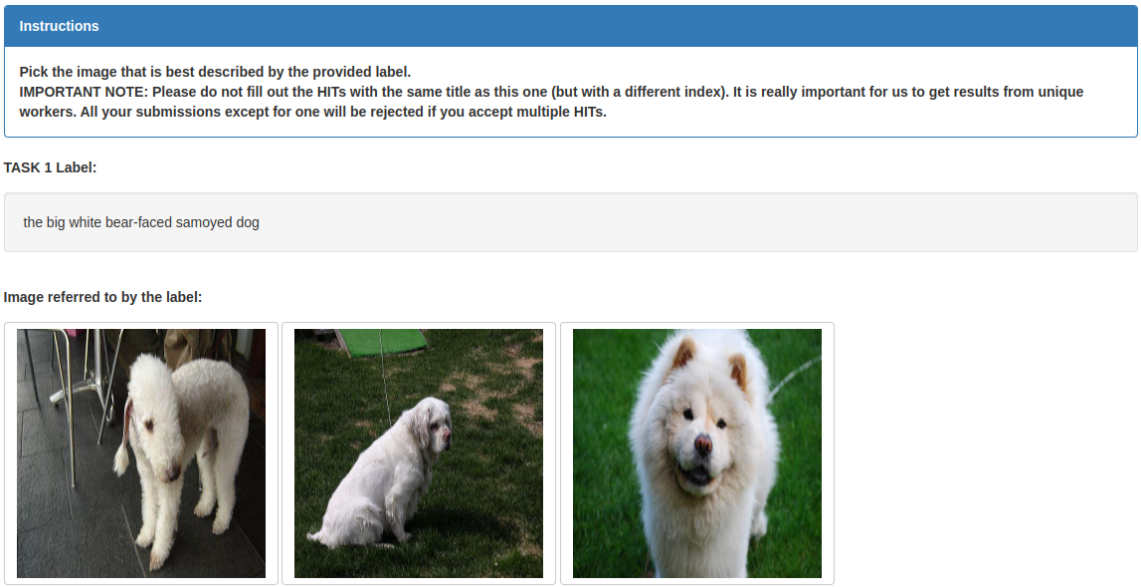

Figure 1: Example of the independent human evaluation task as presented to the participant. Shown are the task instructions and one of the six items the participant was asked to provide an answer for.

Instructions

You will read a conversation between two people that are discussing images. The two people that are having this conversation are playing a game in which they have to rank 9 images. This game is played over 4 rounds. Each round the players are given a different scenario (indicated by "**Task**", in bold letters) by which they have to rank the images (for example: "which of these animals is the cutest?"). Each time you see a new scenario a new round has started.

We ask you to find which image or images the players are referring to each time they mention one or more of the images shown below the conversation. To help you see when and where a player mentioned an image (or images), the expression that we need you to find the image (or images) for will be marked with two red arrows (➡ and ⬅). As you progress, you will be able to see increasingly more of the dialogue. Use that to your advantage, but note that you cannot undo any of your previous selections.

Note that this is a relatively long task! We estimate it to take **50 minutes on average**. We will need you to pay close attention all the way through to the end. But you will be compensated accordingly for your efforts! **Be aware that there will occasionally be an attention check**.

To illustrate the task, please have a look at the following example:

A: I think ➡ the dog ⬅ is cute. What do you think?

What image or images did **A** refer to?
(between the ➡ and ⬅ arrows)

Answer: Here, "the dog" refers to the image of the dog, thus you would select the image in the middle

A: I think the dog is cute. What do you think?
B: I also think ➡ it ⬅ is cute.

What image or images did **B** refer to?
(between the ➡ and ⬅ arrows)

Answer: Here, "it" refers back to "the dog", which refers to the image of the dog, thus you would again select the image in the middle

A: I think the dog is cute. What do you think?
B: I also think it is cute.
B: I do like ➡ the cat ⬅ as well. It looks a bit grumpy, but it's cute. Honestly, I think all these animals are adorable in their own way

What image or images did **B** refer to?
(between the ➡ and ⬅ arrows)

Answer: Here, "the cat" refers to the image of the cat, thus you would select the leftmost image

A: I think the dog is cute. What do you think?
B: I also think it is cute.
B: I do like the cat as well. ➡ It ⬅ looks a bit grumpy, but it's cute. Honestly, I think all these animals are adorable in their own way

What image or images did **B** refer to?
(between the ➡ and ⬅ arrows)

Answer: Here, "It" refers back to "the cat", which refers to the image of the cat, thus you would again select the leftmost image

Figure 2: Instructions for the holistic human evaluation task as presented to the participant (part 1).

A: I think the dog is cute. What do you think?
B: I also think it is cute.
B: I do like the cat as well. It looks a bit grumpy, but ➡ it ⬅'s cute. Honestly, I think all these animals are adorable in their own way

What image or images did **B** refer to?
(between the ➡ and ⬅ arrows)

Answer: Here, "it" again refers back to "the cat", which refers to the image of the cat, thus you would again select the leftmost image

A: I think the dog is cute. What do you think?
B: I also think it is cute.
B: I do like the cat as well. It looks a bit grumpy, but it's cute. Honestly, I think ➡ all these animals ⬅ are adorable in their own way

What image or images did **B** refer to?
(between the ➡ and ⬅ arrows)

Answer: Here, "all these animals" refers to all three images, thus you would select all three images

A: I think the dog is cute. What do you think?
B: I also think it is cute.
B: I do like the cat as well. It looks a bit grumpy, but it's cute. Honestly, I think all these animals are adorable in ➡ their ⬅ own way

What image or images did **B** refer to?
(between the ➡ and ⬅ arrows)

Answer: Here, "their" refers back to "all these animals", which refers to all three images, thus you would again select all three images

On a final note, be aware that as players progressed through their game, they would rank images along the way. Although it is possible for the players to refer back to an image that they have already ranked, it is relatively uncommon. We visually indicate for each question which images were ranked and which were not (a ranked image will look greyed out). An example of an unranked (left) and a ranked (right) image:

*By clicking "Yes", you indicate that you have read the instructions and are willing to participate in this study.
The main purpose of this data collection is to understand to what extent people are capable of grounding referring expressions in conversation.
You agree that any data we collect from you through your participation in this survey may be used for research purposes and in publications.
Any such data will be anonymized prior to publication.
We will manage your data in accordance with the General Data Protection Regulation (GDPR).
This means that you have the right to withdraw your consent, request your data, and request that your data be deleted, at any time.
Your participation in this study is voluntary and you may decide to stop at any point.
Note that not completing the survey will affect your compensation.
In case you have any questions or concerns you can contact us by sending a message on Prolific or by sending an email to email address

✔ Yes ⊘ No

Figure 3: Instructions for the holistic human evaluation task as presented to the participant (continued, part 2).

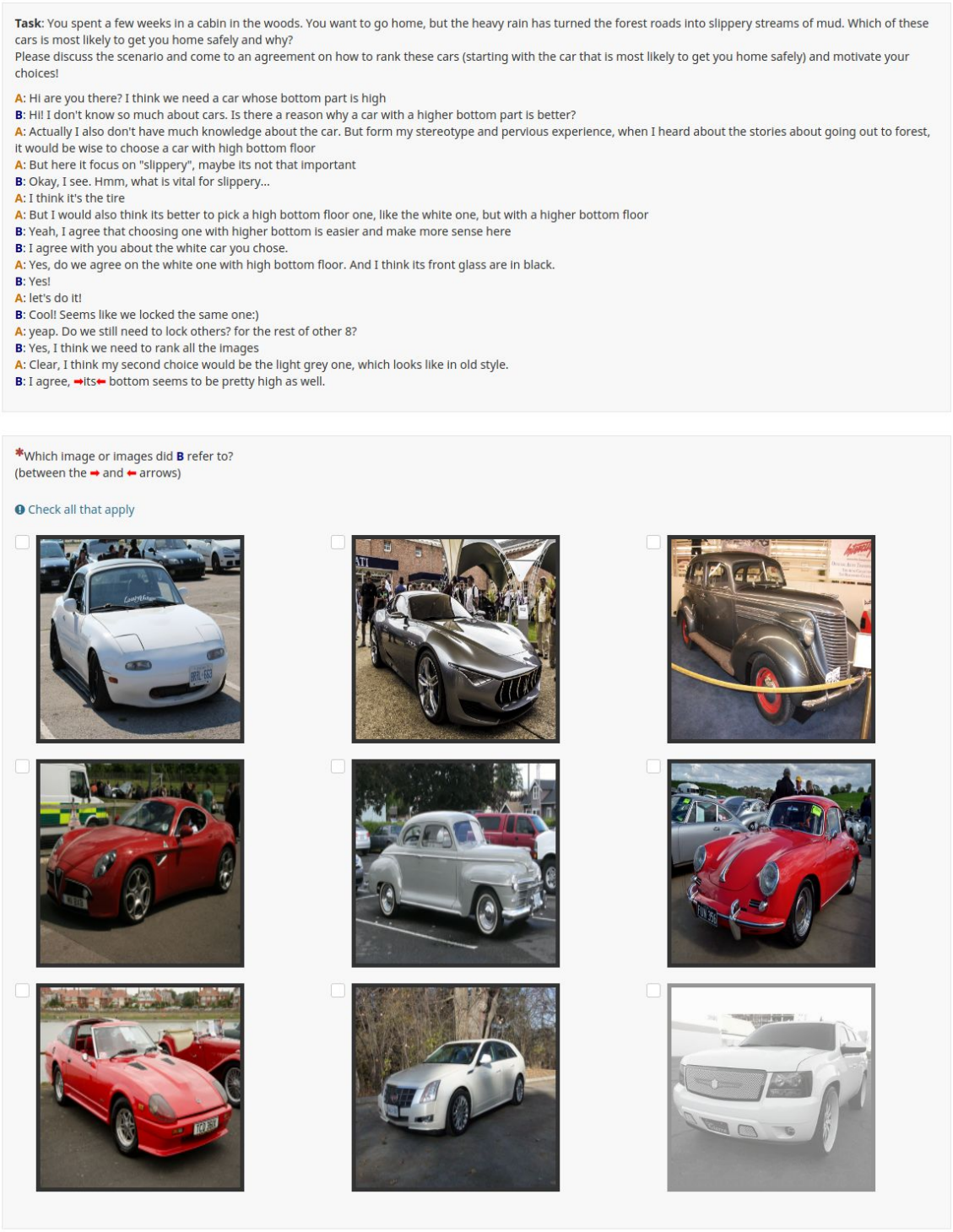

Figure 4: Example of the holistic human evaluation task as presented to the participant. Images with a faded appearance have been successfully ranked by the players at that point in the interaction.