Towards VQA Models That Can Read

Amanpreet Singh[1], Vivek Natarajan, Meet Shah[1], Yu Jiang[1], Xinlei Chen[1],
Dhruv Batra[1,2], Devi Parikh[1,2], and Marcus Rohrbach[1]

[1]Facebook AI Research, [2]Georgia Institute of Technology

1. OCR and Answer Space Analysis

We perform the following analysis on TextVQA's validation set of size 5,734. We find that 44.9% (2575) of LoRRA's predicted answers are from OCR tokens (i.e., using the copy mechanism). The remaining 55.1% of predicted answers are from the SA. This shows that our approach does in fact rely heavily on what it reads in the image, and relies on its copy mechanism to generalize and produce answers that have never been seen before or are rare in the training data. While predicting answers from OCR tokens, the model gets the entire answer string correct 27% (696) of the time, and partially correct (i.e., matches one word in answer) 11% (284) of the time. The percentage of partially correct answers shows the possibilities of getting better results by using n-grams of OCR tokens or spelling correction for fixing incorrect OCR predictions. Similarly, while predicting from the answer space, the model gets the answer correct 22.4% (709/3,159).

We find that 30.6% (1759) of questions have their answers in OCR tokens. For these questions, LoRRA chooses to predict from OCR tokens, 68% (1,208/1,759) of the times and correctly answers 57.5% (697/1,208) of these. Similarly, 48% (2794) of questions have their answers in SA. For these questions, LoRRA chooses to predict from LA 66.75% (1,865/2,794) of the times and gets 38% (710/1,865) of these correct.

81% questions in TextVQA's validation set have more than or equal to 2 OCR tokens. Among these 4645 questions, LoRRA chooses to copy from OCR tokens 49.7% (2,309) and gets 24.3% (560/2,309) of these correct. This suggests that LoRRA doesn't randomly copy OCR token from list of available tokens.

2. TextVQA Examples and LoRRA Predictions

In Fig. 1, we show representative examples from our TextVQA dataset along with the predictions of Pythia+LoRRA. Each example shows the ground truth answer, the predictions from LoRRA, whether the answer prediction was from OCR tokens or the pre-determined answer space, and attention weights for each of the OCR tokens. The examples indicate the following points:

- The model is able to successfully answer questions about times, dates, brands, cities and places, and is often able to correctly spell them even if the OCR tokens had them misspelled (as it is able to predict the correct answer from answer space). See Fig. 1k (short hand's hour), Fig. 1g (birthday date), Fig. 1r (no turn on red sign), Fig. 1s (city "london" correct from the large text), Fig. 1o (samsung).

- The model is able to successfully answer questions involving colors and spatial reasoning. See Fig. 1e (player on the right), Fig. 1f (location of 25 coin), Fig. 1c (location of banner).

- Model is able to distinguish objects based on spatial positioning, colors and relative positioning well. See Fig. 1q where the model needs to identify the correct sign based on multiple colors, or Fig. 1r where the model needs to identify the correct sign in the red circle and shows that the model is not biased toward "stop" as other VQA models are, or Fig. 1a where model needs to predict the correct number based on spatial reasoning from the choice of 7 or 14, or Fig. 1f where the model needs to predict the correct coin's value based on the position mentioned.

- The model is also able to reason about basic sizes (less, greater, smallest) and shapes (circle). See Fig. 1f where the model needs to figure out which one is top and silver coin, or Fig. 1k where the model needs to figure out which one is the shorter hand, or Fig. 1q where the model needs to figure out which one is lowest measurement among four.

- The model often predicts an answer from the answer space as informed by OCR tokens. See Fig. 1k where the Pythia model which doesn't use OCR predicts 3 in this case, but our approach predicts 4 which is the correct answer.

- The model often answers questions about cities with "new york". See Fig. 1j where the model predicts New York instead of San Francisco. We have observed this bias in other city related questions as well.

- For yes/no questions, even though “yes” is the more common answer, the model does predict “no” frequently. See Fig. 1m, Fig. 1l.

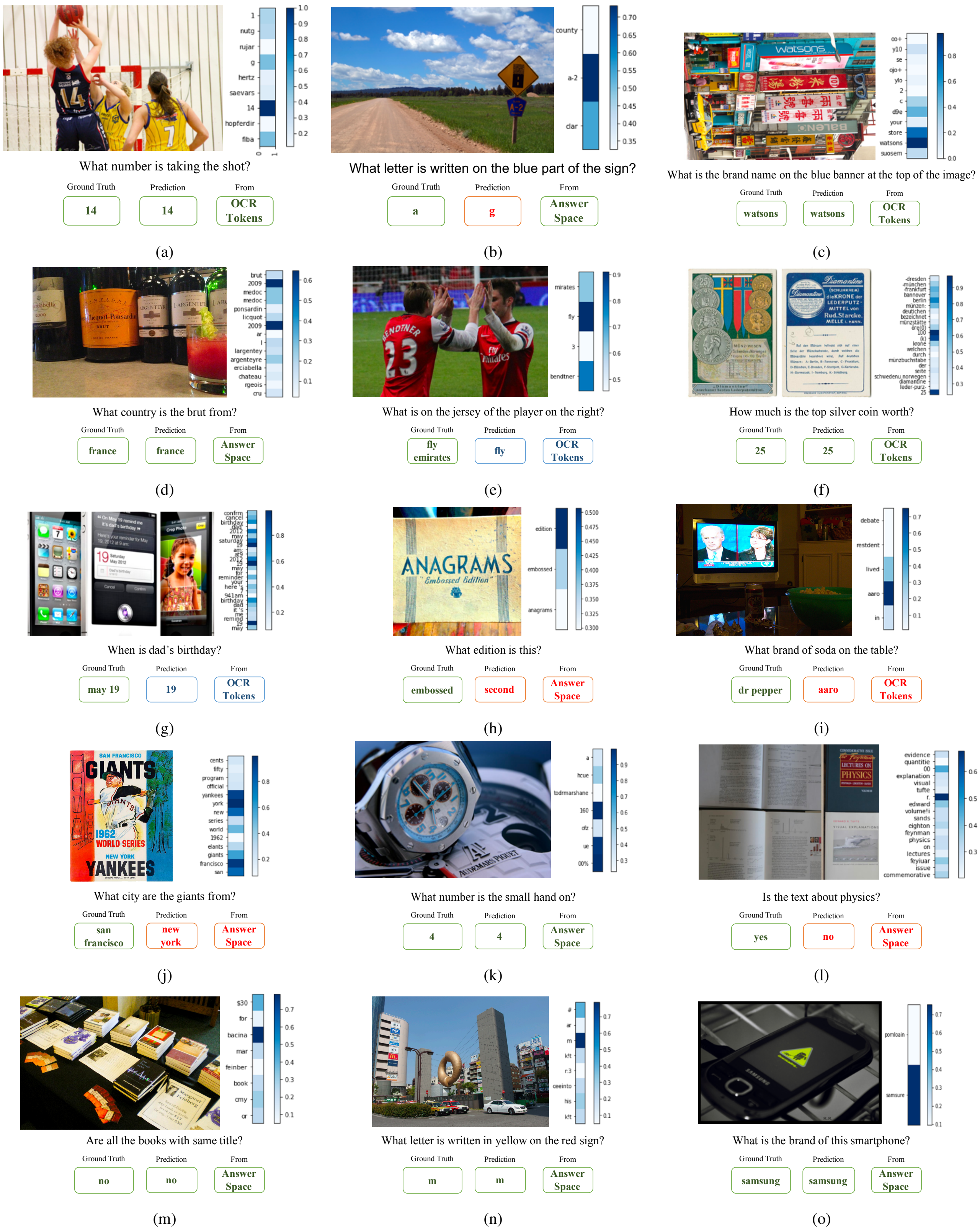
- Sometimes when the answer is not in the answer space, but the partial answer is in OCR tokens, the model predicts the partial answer which is closest to the actual answer. See Fig. 1e where the model predicts “fly” instead of “fly emirates”, or Fig. 1g where the model predicts only the birthday date “19”, instead of “may 19”. By construction our model can only copy a single OCR token, but our TextVQA dataset contains Q/A pairs which require copying multiple OCR tokens in the right order. Explore this is an interesting direction for future work.

- The model sometimes gets seemingly simple questions wrong by predicting generic answers. See Fig. 1h where the model can’t predict “embossed” even though it is in OCR tokens, or see Fig. 1b where the model predicts most common letter “g” in the answer space instead of predicting based on “a-2” in the OCR tokens.

- The model has a strong dependency on the quality of OCR tokens produced. If the OCR module missed some text in the image, the model’s output can be wrong. See Fig. 1i or Fig. 1p where the OCR tokens do not contain the ground truth answer or see Fig. 1u where the OCR system is unable to correctly read “irig” the second time.

3. Interface Screenshots

We show three stages of data collection pipeline in Fig. 2, Fig. 3, Fig. 4 and Fig. 5. Fig. 2 and Fig. 3 shows the entrance and first stage of our task which is used to identify and remove images without text in them. Fig 4 shows the second stage of our task which is used to collect questions on images with text. Finally, the third stage interface is shown in 5 which is used to collect the answer for a question about an image.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

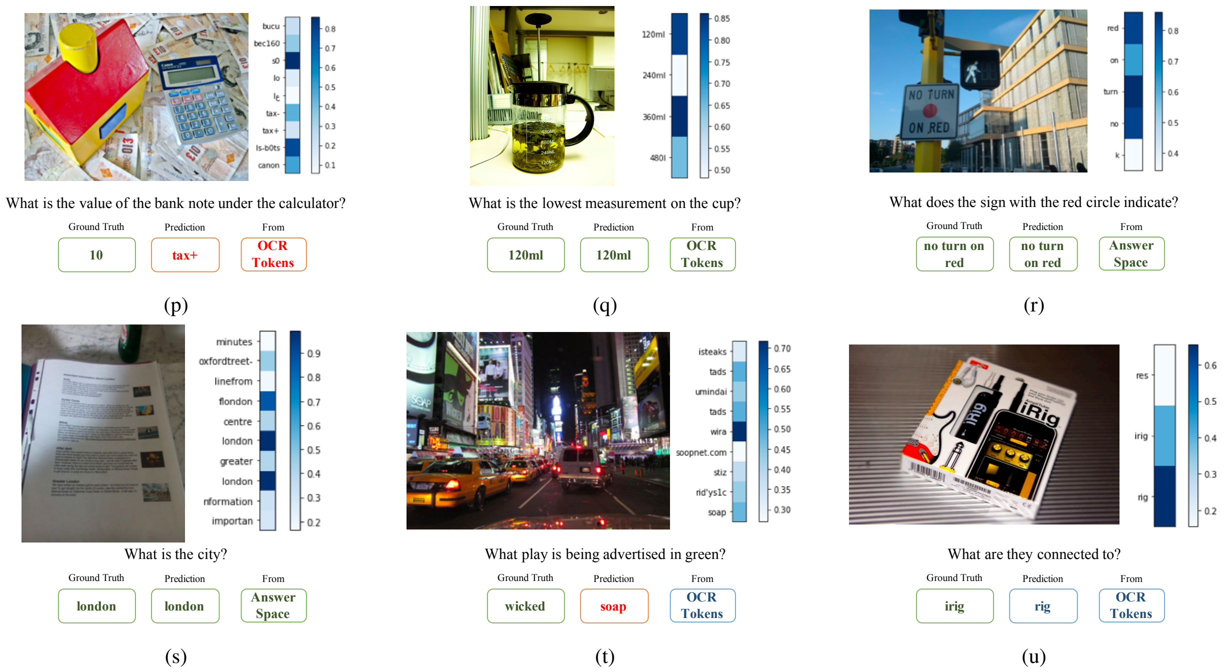
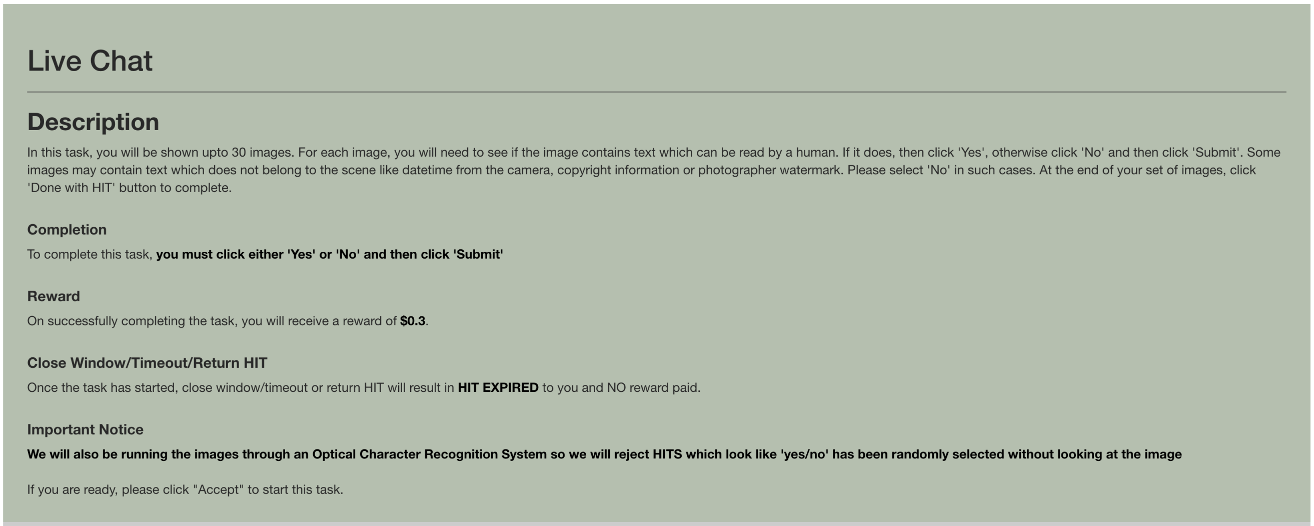


Figure 1: **TextVQA Examples and LoRRA's predictions on them.** We show multiple examples from TextVQA, ground truth answers, along with predictions from LoRRA, attention maps on OCR tokens and where LoRRA predicted the answer from (OCR tokens or pre-determined answer space. Green, red, and blue boxes correspond to correct, incorrect, and partially correct answers, respectively. On the right side of each image, we show attention bars which depict attention weights (0-1) for each of the OCR tokens.



Live Chat

Description

In this task, you will be shown upto 30 images. For each image, you will need to see if the image contains text which can be read by a human. If it does, then click 'Yes', otherwise click 'No' and then click 'Submit'. Some images may contain text which does not belong to the scene like datetime from the camera, copyright information or photographer watermark. Please select 'No' in such cases. At the end of your set of images, click 'Done with HIT' button to complete.

Completion

To complete this task, **you must click either 'Yes' or 'No' and then click 'Submit'**

Reward

On successfully completing the task, you will receive a reward of **$0.3**.

Close Window/Timeout/Return HIT

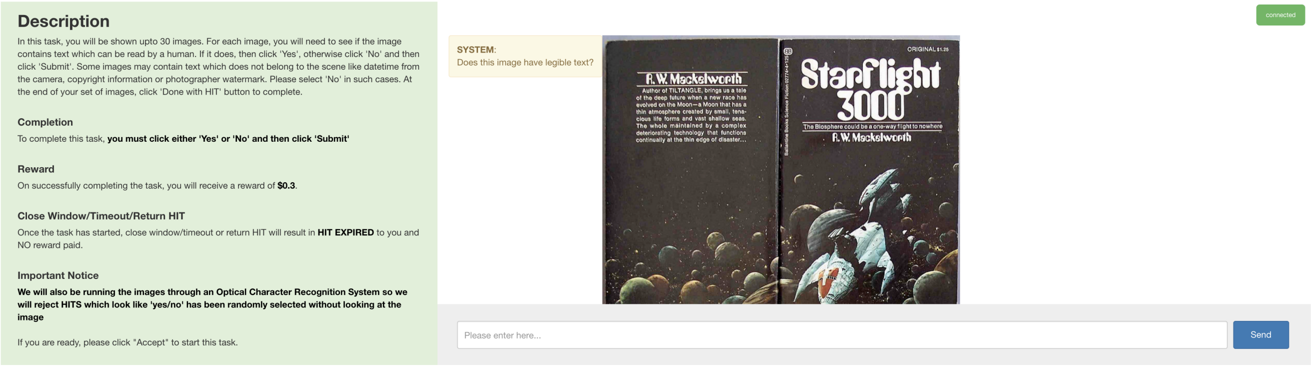
Once the task has started, close window/timeout or return HIT will result in **HIT EXPIRED** to you and NO reward paid.

Important Notice

We will also be running the images through an Optical Character Recognition System so we will reject HITS which look like 'yes/no' has been randomly selected without looking at the image

If you are ready, please click "Accept" to start this task.

Figure 2: **Introduction page for our task.**



Description

In this task, you will be shown upto 30 images. For each image, you will need to see if the image contains text which can be read by a human. If it does, then click 'Yes', otherwise click 'No' and then click 'Submit'. Some images may contain text which does not belong to the scene like datetime from the camera, copyright information or photographer watermark. Please select 'No' in such cases. At the end of your set of images, click 'Done with HIT' button to complete.

Completion

To complete this task, **you must click either 'Yes' or 'No' and then click 'Submit'**

Reward

On successfully completing the task, you will receive a reward of **$0.3**.

Close Window/Timeout/Return HIT

Once the task has started, close window/timeout or return HIT will result in **HIT EXPIRED** to you and NO reward paid.

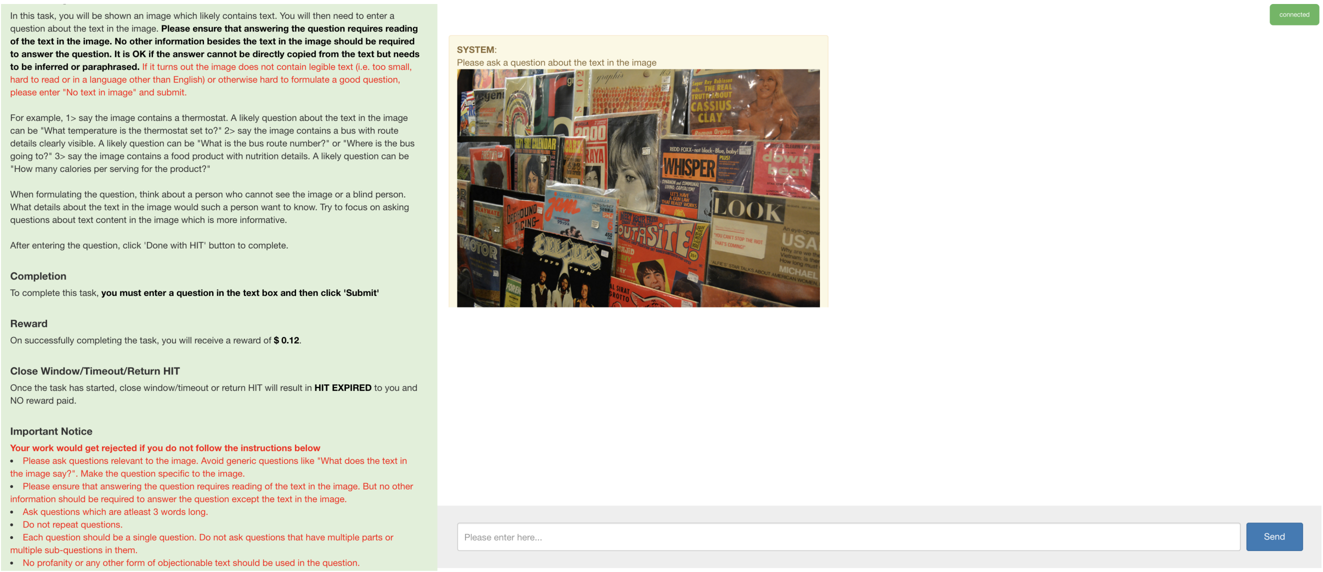
Important Notice

We will also be running the images through an Optical Character Recognition System so we will reject HITS which look like 'yes/no' has been randomly selected without looking at the image

If you are ready, please click "Accept" to start this task.

SYSTEM:
Does this image have legible text?

Figure 3: **Text detection main task.** First stage of our tasks is used to identify and remove images without text.



In this task, you will be shown an image which likely contains text. You will then need to enter a question about the text in the image. **Please ensure that answering the question requires reading of the text in the image. No other information besides the text in the image should be required to answer the question. It is OK if the answer cannot be directly copied from the text but needs to be inferred or paraphrased.** If it turns out the image does not contain legible text (i.e. too small, hard to read or in a language other than English) or otherwise hard to formulate a good question, please enter "No text in image" and submit.

For example, 1> say the image contains a thermostat. A likely question about the text in the image can be "What temperature is the thermostat set to?" 2> say the image contains a bus with route details clearly visible. A likely question can be "What is the bus route number?" or "Where is the bus going to?" 3> say the image contains a food product with nutrition details. A likely question can be "How many calories per serving for the product?"

When formulating the question, think about a person who cannot see the image or a blind person. What details about the text in the image would such a person want to know. Try to focus on asking questions about text content in the image which is more informative.

After entering the question, click 'Done with HIT' button to complete.

Completion

To complete this task, **you must enter a question in the text box and then click 'Submit'**

Reward

On successfully completing the task, you will receive a reward of **$ 0.12**.

Close Window/Timeout/Return HIT

Once the task has started, close window/timeout or return HIT will result in **HIT EXPIRED** to you and NO reward paid.

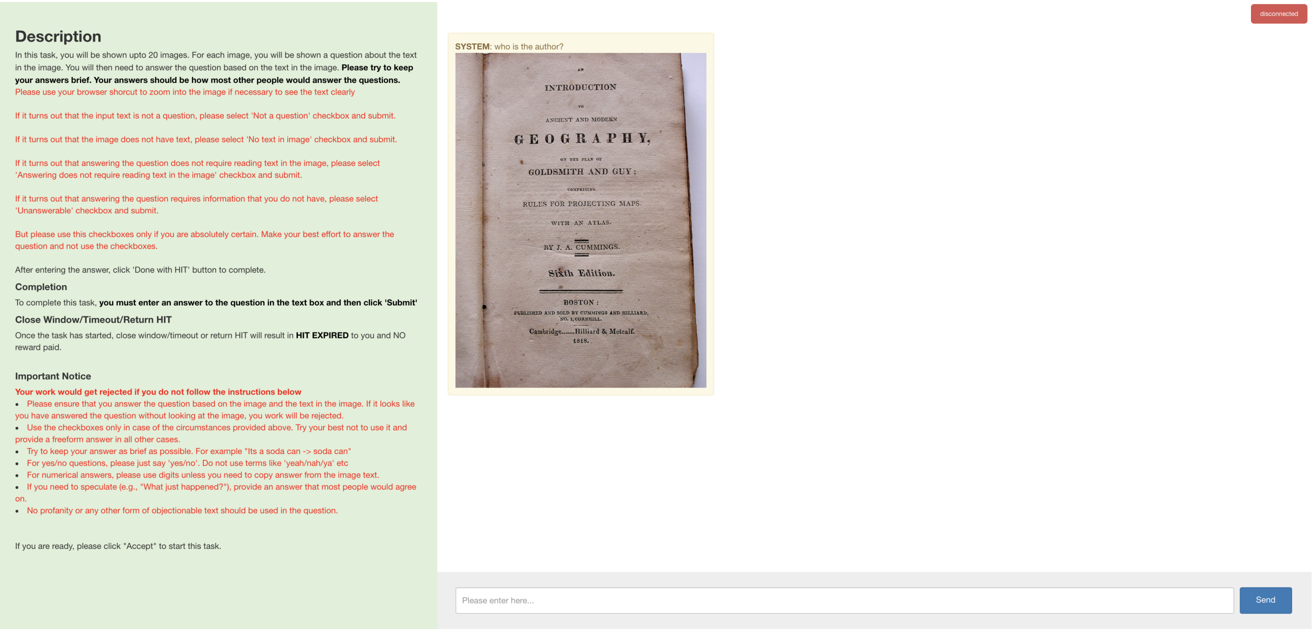
Important Notice

Your work would get rejected if you do not follow the instructions below

- Please ask questions relevant to the image. Avoid generic questions like "What does the text in the image say?". Make the question specific to the image.
- Please ensure that answering the question requires reading of the text in the image. But no other information should be required to answer the question except the text in the image.
- Ask questions which are atleast 3 words long.
- Do not repeat questions.
- Each question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.
- No profanity or any other form of objectionable text should be used in the question.

SYSTEM:
Please ask a question about the text in the image

Figure 4: **Question task.** In the second stage, we ask workers to ask a question about an image whose answer requires reading text in the image. We provide instructions and rules to ensure that we get high quality questions.



Description

In this task, you will be shown upto 20 images. For each image, you will be shown a question about the text in the image. You will then need to answer the question based on the text in the image. **Please try to keep your answers brief. Your answers should be how most other people would answer the questions.**
Please use your browser shorcut to zoom into the image if necessary to see the text clearly

If it turns out that the input text is not a question, please select 'Not a question' checkbox and submit.

If it turns out that the image does not have text, please select 'No text in image' checkbox and submit.

If it turns out that answering the question does not require reading text in the image, please select 'Answering does not require reading text in the image' checkbox and submit.

If it turns out that answering the question requires information that you do not have, please select 'Unanswerable' checkbox and submit.

But please use this checkboxes only if you are absolutely certain. Make your best effort to answer the question and not use the checkboxes.

After entering the answer, click 'Done with HIT' button to complete.

Completion

To complete this task, **you must enter an answer to the question in the text box and then click 'Submit'**

Close Window/Timeout/Return HIT

Once the task has started, close window/timeout or return HIT will result in **HIT EXPIRED** to you and NO reward paid.

Important Notice

Your work would get rejected if you do not follow the instructions below

- Please ensure that you answer the question based on the image and the text in the image. If it looks like you have answered the question without looking at the image, you work will be rejected.
- Use the checkboxes only in case of the circumstances provided above. Try your best not to use it and provide a freeform answer in all other cases.
- Try to keep your answer as brief as possible. For example "Its a soda can -> soda can"
- For yes/no questions, please just say 'yes/no'. Do not use terms like 'yeah/nah/ya' etc
- For numerical answers, please use digits unless you need to copy answer from the image text.
- If you need to speculate (e.g., "What just happened?"), provide an answer that most people would agree on.
- No profanity or any other form of objectionable text should be used in the question.

If you are ready, please click "Accept" to start this task.

SYSTEM: who is the author?

AN
INTRODUCTION
TO
ANCIENT AND MODERN
GEOGRAPHY,
ON THE PLAN OF
GOLDSMITH AND GUY;
COMPRISING
RULES FOR PROJECTING MAPS.
WITH AN ATLAS.
BY J. A. CUMMINGS.
Sixth Edition.
BOSTON:
PUBLISHED AND SOLD BY CUMMINGS AND HILLIARD,
NO. 1, CORNHILL.
Cambridge.......Hilliard & Metcalf.
1818.

Figure 5: **Answer task.** In the third stage, we ask workers to answer a question about the image.