Details of Data Collection and Crowd Management for GLUCOSE (Generalized and Contextualized Story Explanations)

Lori Moon
Lauren Berkowitz, Jennifer Chu-Carroll, Nasrin Mostafazadeh

Abstract

The GLUCOSE dataset is the result of an effective multi-stage design and implementation process for collecting common sense mini-theories about the world from crowd workers at scale. Leveraging Amazon Mechanical Turk (AMT) for the crowdsourcing task required breaking it down into Human Intelligence Tasks (HIT), which has an expected amount of work and pay rate by crowd workers. Breaking a cognitively complex task such as GLUCOSE into appropriate tasks required numerous iterations on all aspects of the task. Having a truly intuitive UI, adaptive to workers' requests and comments, played the most crucial role in the success of the task. Inevitably, accomplishing the task involved a lot of training of workers. In this document, we provide further details on GLUCOSE's data collection pipeline and crowd quality management processes. Given the complexity of crowdsourcing such a dataset at scale, we hope that sharing the details of our practice can help future AI research.

1 Introduction to GLUCOSE

GLUCOSE (GeneraLized and COntextualized Story Explanations) is a large-scale dataset of such commonsense causal knowledge. Given a short story and a sentence X from the story, the GLUCOSE task asks crowd workers to compose specific statements about X's causes and effects and to generalize each specific statement into a general causal rule. For each sentence, ten pairs of specific statements and general rules are solicited, corresponding to ten dimensions of commonsense causal explanation. GLUCOSE is described in our EMNLP 2020 paper[1], which introduces the dataset, rationalizes its knowledge model, and presents the results from various models trained on the dataset. The full dataset, along with the trained models, can be accessed on GitHub [2].

The rest of this document describes GLUCOSE's knowledge model, data collection pipeline includign the GLUCOSE task, our procedures for data quality management, and some analyses of the collected data.

2 The GLUCOSE Knowledge Model

GLUCOSE has a unique take on explaining story events. As illustrated in the main paper, each story is explained through ten causal dimensions. The semi-structured explanation for each dimension includes both a specific statement and a general rule.

For each sentence X in a story,[3] the annotations are split into two categories: **causes** of X—i.e., events **before** X that caused or enabled it—and **effects** of X—i.e., events **after** X that resulted from it. Within each category, five types of causes or effects are described:

[1] Main GLUCOSE paper: `https://arxiv.org/abs/2009.07758`

[2] The GLUCOSE repository: `https://github.com/ElementalCognition/glucose`

[3] The stories used in the task were from the ROCStories corpus (`https://www.cs.rochester.edu/nlp/rocstories/`). We filtered the corpus to stories that target 5–8 year old children.

Dimension	**Content**
Dimension 1	An event that directly causes or enables the sentence X
Question 1	Consider the events that happen before X (or are likely to happen). Does any of them directly cause X, or simply make X possible (i.e., enable X)?
Dimension 2	An emotion or basic human drive that motivates X
Question 2	Consider the likely emotions and basic human drives of the participants in X. Does any of these states of mind/feelings motivate the participant to do X?
Dimension 3	A location state that enables X
Question 3	Consider the likely locations of the story participants (people, things, etc.) before X. Does one of these location states make X possible?
Dimension 4	A possession state that enables X
Question 4	Consider which things each story participant possesses (or is likely to possess) at different times. Does any of these possession states make X possible? (This often happens when X is an event that involves physical transfer, change of ownership, or destroying something.)
Dimension 5	Other property (besides location, emotional state, or possession) that enables X
Question 5	Consider everything else about the participants (people, things, etc.) in the story. Does some likely state of a participant besides their location, emotional state, and possessions make X possible?
Dimension 6	An event that is directly caused or enabled by X
Question 6	Consider the events that happen after X (or are likely to happen). Does X directly cause any of them, or simply make it possible (i.e., enable it)?
Dimension 7	An emotion that is caused by X
Question 7	Consider the likely emotions of the participants in X and those affected by it. Is any of these emotions caused by X?
Dimension 8	A change of location that X results in
Question 8	Consider the likely locations of the story participants (people, things, etc.) after X. Does X directly result in any of these location states? (This often happens when X is an event that involves change of location or movement.
Dimension 9	A change of possession that X results in
Question 9	Consider which things each story participant possesses (or is likely to possess) at different times. Does X directly result in any of these possession states? (This often happens when X is an event that involves physical transfer, change of ownership, or creating something.
Dimension 10	Other change in property (besides location, emotional state, or possession) that X results in
Question 10	Consider everything else about the participants (people, things, etc.) in the story. Does X directly result in some participant being in some state? For this question, ignore locations, emotional states, and possessions.

Table 1: The prompts for GLUCOSE's 10 dimensions of explanation for each sentence X.

1. Events
2. Basic human drives and emotions
3. Location states
4. Possession states
5. Any other attributes (states) not captured by the above

Given the 'before' and 'after' categories for the list of five types of knowledge above, GLUCOSE captures ten **dimensions of explanation**. The dimensions within the two categories mirror each other: dimension 1 is events or states that caused/enabled X, while dimension 6 is events or states that X likely caused/enabled; and likewise for dimensions 2 and 7, 3 and 8, and so on. The questions used to elicit an explanation along each dimension are listed in Table 1.

For each dimension, workers were expected first to provide a story-specific causal statement (henceforth referred to as just a **specific statement**), and then to generalize that statement into a broader causal rule (henceforth referred to as a **general rule**). For example, consider the story below:

> *My daughter was born last week. I have a ton of pictures to show you. She is so cute. She has a ton of hair. I am in love.*

Say X is the second sentence: *I have a ton of pictures to show you.* For dimension 1—events before X that caused or enabled it—the worker's specific statement might indicate that the author's daughter having been born enabled him to have pictures of her to show. Based on this statement, the worker would then state a general rule about the world—e.g., that someone being born enables her to be photographed by her parents. The same process was repeated for each dimension listed in Table 1. For example, for question 6, the specific statement might be that the speaker showing pictures to people causes the people to see pictures. The corresponding general rule would be that someone showing someone else something results in them seeing that thing.

Not all dimensions are relevant for every sentence, so workers were permitted to leave up to eight of the ten dimensions blank (this was particularly common for dimensions 5 and 10—the "other attributes" dimensions). However, if a worker did consider a dimension to be relevant, then they had to fill out both the specific statement and the general rule about it.

3 Data Collection and Quality Management Pipeline

This section describes how we translated the GLUCOSE data schema into a user friendly interface that worked within the constraints of the Mturk crowd-sourcing platform.

Mturk requires that tasks be broken down into units called HITs (Human Intelligence Tasks). These are the units of work that workers are paid for. Because crowd workers accept and work on HITs from many different requesters, it is important to follow similar practices in terms of the amount of work per HIT and the rate of compensation. One of the challenges in our task design was in determining how to break down the GLUCOSE task into portions that cohered as individual HITs.

The GLUCOSE dataset was collected through our main user interface (UI) that was developed in-house and hosted on Amazon Web Services (AWS)[4]. Crowd workers were recruited exclusively through MTurk, but they were vetted with several layers of testing and training designed in-house. All the data reviewing and analysis was done through our in-house tooling that periodically retrieved the crowd worker responses, as shown in figure 1.

[4] GLUCOSE main data acquisition UI: `https://bit.ly/2R8XcTt`

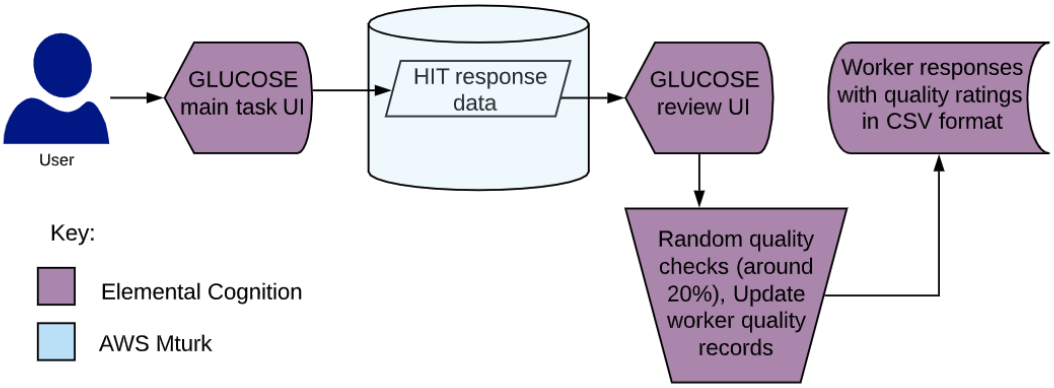

User
GLUCOSE main task UI
HIT response data
GLUCOSE review UI
Worker responses with quality ratings in CSV format
Random quality checks (around 20%), Update worker quality records
Key:
Elemental Cognition
AWS Mturk

Figure 1: Our in-house user interfaces and quality control (purple) used AWS Mturk worker pools and storage. The diagram is simplified in that, although crowd workers entered our main task UI to see the questions and give responses, our UI was accessed through the 'HITs' panel on their AWS Mturk worker account page. Our review UI also interacted with AWS Mturk, retrieving data then sending bonuses and giving approvals with calls to Mturk.

3.1 Qualification Task

Although Mturk provides access to a large worker pool, the level of worker quality management that it supports is limited. Therefore, in addition to task design, we designed and maintained data quality management strategies to ensure high quality responses to the task. For a task like GLUCOSE, where annotators have a high degree of liberty in terms of what constitutes a correct response, a lot of hands-on quality checking and training is necessary. Making training and quality control feasible for such a large-scale task involved a variety of mechanisms including testing, random sample checking, and intermittent statistical analysis.

In order to qualify workers for the main GLUCOSE task, crowd workers had to take a qualification test. AWS Mturk allows for only very limited question types in their qualification framework, so we developed another HIT with an in-house UI[5] to thoroughly test workers for entry to the main task[6].

3.1.1 Participants

The GLUCOSE qualification task used three of these Mturk-internal ratings:

1. The worker had successuflly completed more than 100 HITs on Mturk,

2. The worker had an acceptance rate of at least 95%, meaning at least 95% of their submitted HITs had been approved by requesters, as opposed to rejected,

3. The worker had an IP address matching a list of country codes.[7]

[5] GLUCOSE qualification UI: `https://bit.ly/34Pej0N`

[6] We also piloted the qualification task multiple times before using it to recruit workers for the main task.

[7] Country codes were originally restricted to a standard list Mturk provides, containing the US, Great Britain, Ireland, New Zealand, Australia, South Africa, and Canada. Half way into the project, we opened it up to many other countries per worker requests, including India, China, Germany, Jamaica, the Dominican Republic, Belize, and Nigeria. Many countries that are not in the standard list are already English speaking, and, furthermore, the qualification test itself would be challenging to pass if someone did not already have a good command of English. Since we were not limited to a particular dialect of English, we attempted to add the country code of nearly 50 countries which are English-speaking, but the Mturk interface limited us to 35 countries.

We did not limit how many times workers could take the test, however, they could only take it once per launch. Each launch of the qualification task was a one-HIT launch with 1,500 workers able to do the HIT. The task was launched nine times over the course of running the main task. A pool of 1039 workers qualified for our main task via the Qualifying Exam.

We wrote 35 qualification questions and each HIT used a random 10 to test workers. Workers who got 90-100% correct were automatically qualified for the main task. Workers who got 70-80% correct were automatically added to the warm-up task. Workers with a score below 70% were encouraged to study the guidelines and take the test on the next batch release. We did not provide answers to workers in order to ensure an answer key was not leaked.

3.1.2 Qualification Content

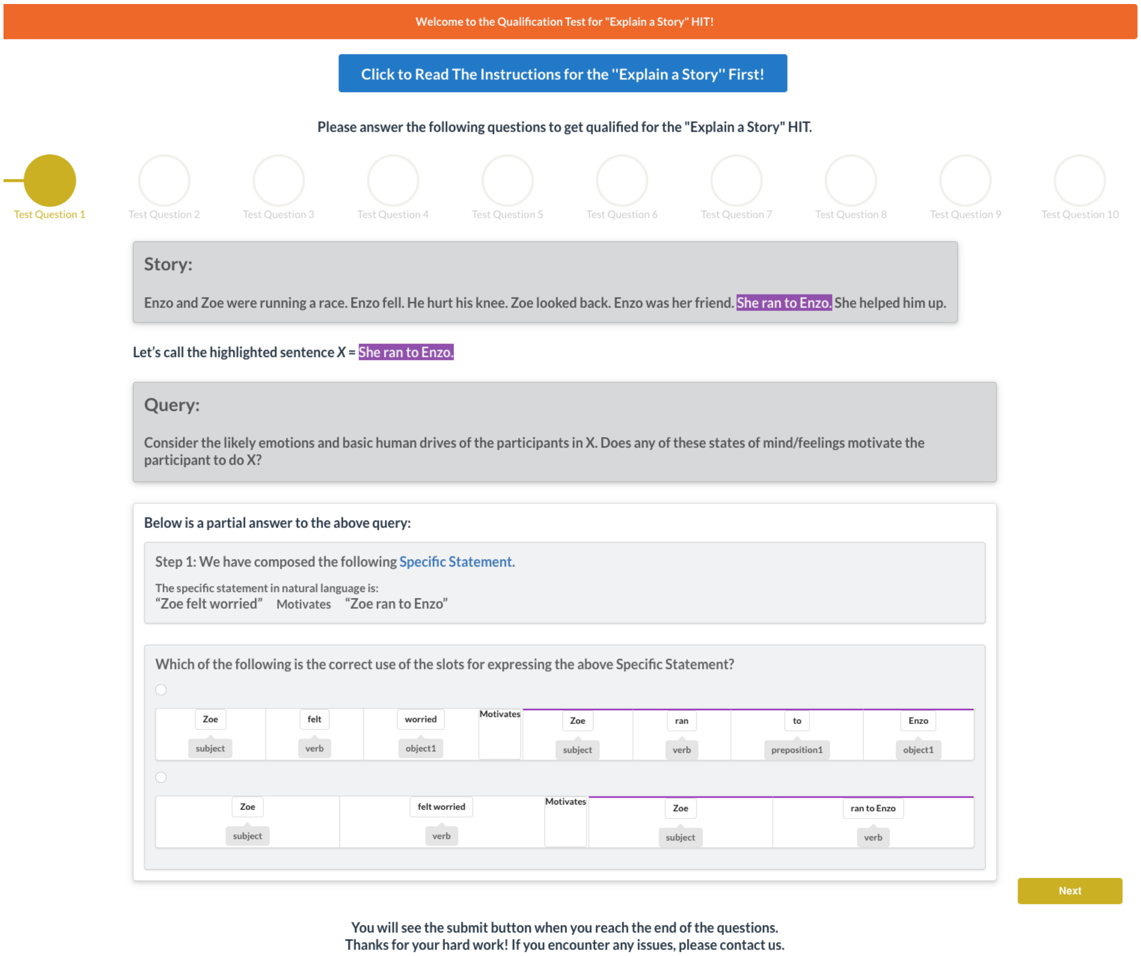
The qualification content consisted of the qualification HIT and training materials for doing the qualification HIT. Users were encouraged to carefully read the instructions in the qualification HIT UI, which included stories with examples of GLUCOSE style specific statements and general rules about the stories. The qualification HIT UI also provided a link to a text document that contained general guidelines for the task.[8] This document was revised and expanded in the course of the task to improve worker training. The general guidelines also contained a link to a FAQ page that was also continuously revised and updated as workers asked about the task.

The qualification test contained questions testing expertise in three areas. The first area involved identifying correct use of the UI slots for language expressions. The main task would use slots to guide annotators in selecting the essential expressions to use in a rule. These slots had labels like 'subject' and 'verb' and sometimes had drop-down lists of selections. For example, 'preposition' had a drop-down list of prepositions that users could choose. In anticipation of this format, we put multiple-choice questions on the qualification test HIT that showed correct and incorrect uses of the slots. An example is in Figure 2). In this example, the crowd-worker had to choose which slot use was correct for the sentence, 'Zoe felt worried'. One choice was 'Zoe (subject) felt (verb) worried (object1)' and the other was 'Zoe (subject) felt worried (verb)'. The correct answer was the former, in which 'worried' takes the object1 slot. The slots were simplified from part of speech tags to allow for easier understanding and to encourage simplification of the language of the stories.

The second area covered in the test involved recognizing the right level of generalization (see Figure 3). For generalization understanding, prospective workers were presented with a specific statement based on the target story. Then they had to choose the best general rule to derive from that statement. General statements in the main task would have drop-down menus with variables, like 'Someone_A', which we introduced to avoid pronoun reference and co-reference issues. The questions about generalization tested their understanding of the variable format, their understanding of when to replace referents with variables (e.g., changing proper names to variables), and their understanding that generalization must not be so abstract that it does not make sense. In the example in figure 3, they have to generalize the specific statement 'Fernando puts his plant in the sun **causes or enables** Fernando's plant looks healthy.' The choices are 'Someone_A puts a plant Somewhere_A **causes or enables** the plant becomes healthy' (too specific), 'Someone_A puts Something_A that is a plant in the sun **causes or enables** Something_A becomes healthy' (correct response), and 'Someone_A puts Something_A in Somewhere_A **causes or enables** Something_A becomes healthy' (too general).

The third area tested involved identifying causes and effects. One of the major issues people had on pilots was with representing a proper temporal understanding of the stories (see Figure 4). For understanding causes and effects, users were presented with sentences in the cause and effect structure, one of which was

[8] The shared document is at: `https://tinyurl.com/ql2vhaj`



Please answer the following questions to get qualified for the "Explain a Story" HIT.

Story:

Enzo and Zoe were running a race. Enzo fell. He hurt his knee. Zoe looked back. Enzo was her friend. She ran to Enzo. She helped him up.

Let's call the highlighted sentence *X* = She ran to Enzo.

Query:

Consider the likely emotions and basic human drives of the participants in X. Does any of these states of mind/feelings motivate the participant to do X?

Below is a partial answer to the above query:

Step 1: We have composed the following Specific Statement.

The specific statement in natural language is:
"Zoe felt worried" Motivates "Zoe ran to Enzo"

Which of the following is the correct use of the slots for expressing the above Specific Statement?

○

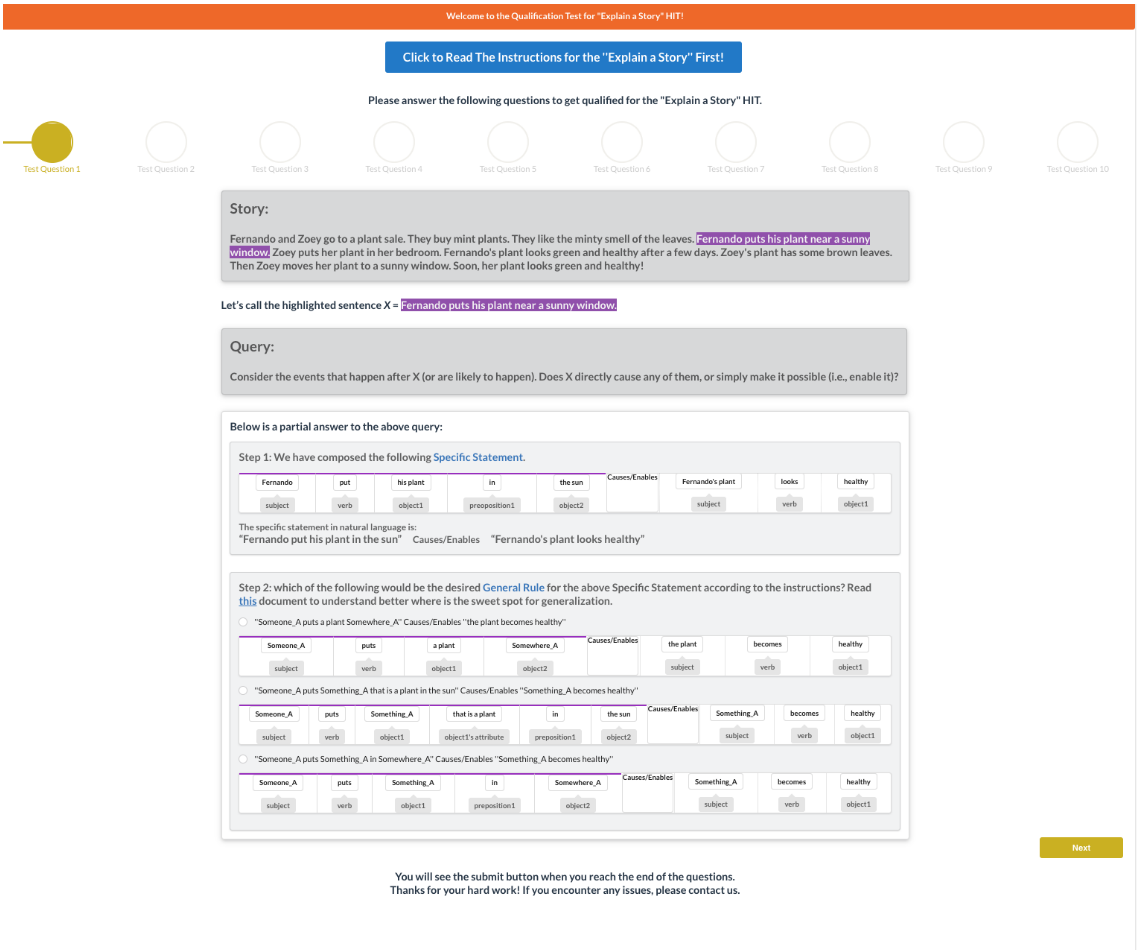
Zoe	felt	worried	Motivates	Zoe	ran	to	Enzo
subject	verb	object1		subject	verb	preposition1	object1

○

Zoe	felt worried	Motivates	Zoe	ran to Enzo
subject	verb		subject	verb

You will see the submit button when you reach the end of the questions.
Thanks for your hard work! If you encounter any issues, please contact us.

Figure 2: Qualification question about slot use: Workers had to choose which slot use was correct for the sentence, 'Zoe felt worried'. One choice correct answer was the former, in which 'worried' takes the object1 slot.



Please answer the following questions to get qualified for the "Explain a Story" HIT.

Story:

Fernando and Zoey go to a plant sale. They buy mint plants. They like the minty smell of the leaves. Fernando puts his plant near a sunny window. Zoey puts her plant in her bedroom. Fernando's plant looks green and healthy after a few days. Zoey's plant has some brown leaves. Then Zoey moves her plant to a sunny window. Soon, her plant looks green and healthy!

Let's call the highlighted sentence *X* = Fernando puts his plant near a sunny window.

Query:

Consider the events that happen after X (or are likely to happen). Does X directly cause any of them, or simply make it possible (i.e., enable it)?

Below is a partial answer to the above query:

Step 1: We have composed the following Specific Statement.

Fernando	put	his plant	in	the sun	Causes/Enables	Fernando's plant	looks	healthy
subject	verb	object1	preoposition1	object2		subject	verb	object1

The specific statement in natural language is:
"Fernando put his plant in the sun" Causes/Enables "Fernando's plant looks healthy"

Step 2: which of the following would be the desired General Rule for the above Specific Statement according to the instructions? Read this document to understand better where is the sweet spot for generalization.

○ "Someone_A puts a plant Somewhere_A" Causes/Enables "the plant becomes healthy"

Someone_A	puts	a plant	Somewhere_A	Causes/Enables	the plant	becomes	healthy
subject	verb	object1	object2		subject	verb	object1

○ "Someone_A puts Something_A that is a plant in the sun" Causes/Enables "Something_A becomes healthy"

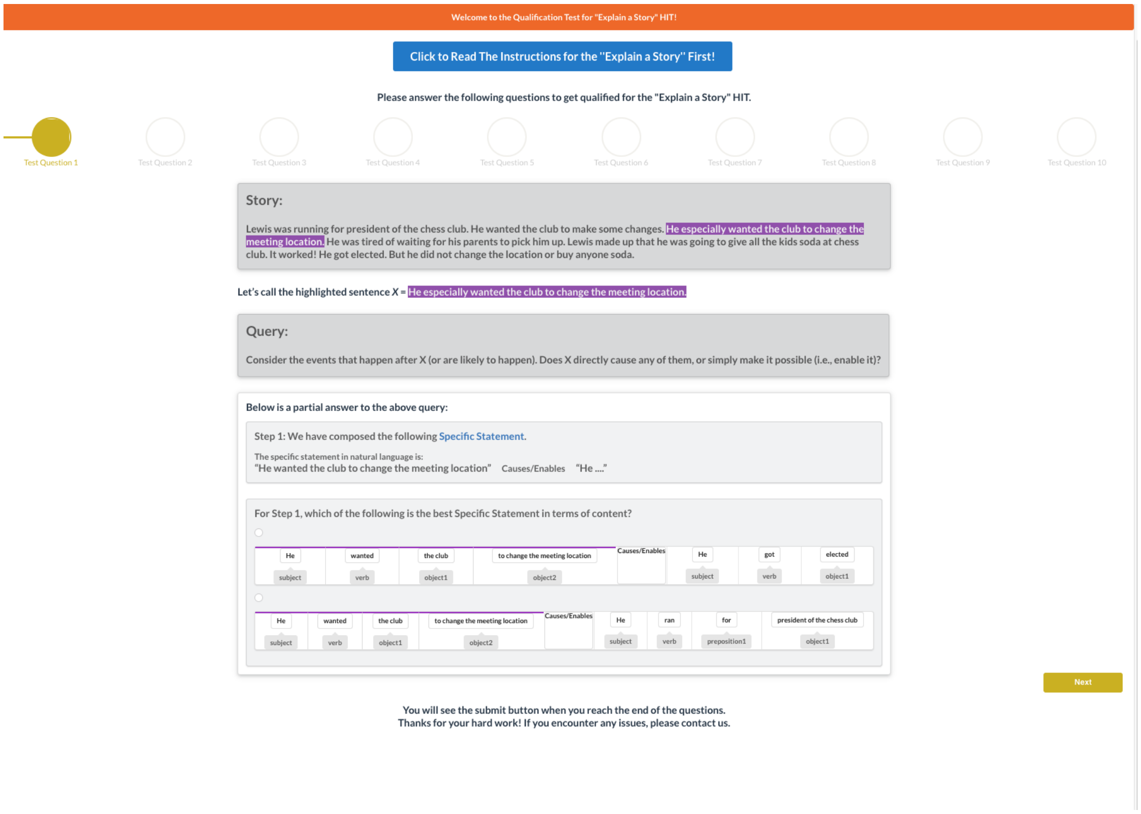
Someone_A	puts	Something_A	that is a plant	in	the sun	Causes/Enables	Something_A	becomes	healthy
subject	verb	object1	object1's attribute	preposition1	object2		subject	verb	object1

○ "Someone_A puts Something_A in Somewhere_A" Causes/Enables "Something_A becomes healthy"

Someone_A	puts	Something_A	in	Somewhere_A	Causes/Enables	Something_A	becomes	healthy
subject	verb	object1	preposition1	object2		subject	verb	object1

You will see the submit button when you reach the end of the questions.
Thanks for your hard work! If you encounter any issues, please contact us.

Figure 3: Sample qualification question about the right level of generalization: Workers have to generalize the specific statement 'Fernando puts his plant in the sun **causes or enables** Fernando's plant looks healthy.' The choices are 'Someone_A puts a plant Somewhere_A **causes or enables** the plant becomes healthy' (too specific), 'Someone_A puts Something_A that is a plant in the sun **causes or enables** Something_A becomes healthy' (correct response), and 'Someone_A puts Something_A in Somewhere_A **causes or enables** Something_A becomes healthy' (too general).



Please answer the following questions to get qualified for the "Explain a Story" HIT.

Story:

Lewis was running for president of the chess club. He wanted the club to make some changes. He especially wanted the club to change the meeting location. He was tired of waiting for his parents to pick him up. Lewis made up that he was going to give all the kids soda at chess club. It worked! He got elected. But he did not change the location or buy anyone soda.

Let's call the highlighted sentence *X* = He especially wanted the club to change the meeting location.

Query:

Consider the events that happen after X (or are likely to happen). Does X directly cause any of them, or simply make it possible (i.e., enable it)?

Below is a partial answer to the above query:

Step 1: We have composed the following Specific Statement.

The specific statement in natural language is:
"He wanted the club to change the meeting location"   Causes/Enables   "He ...."

For Step 1, which of the following is the best Specific Statement in terms of content?

○

He	wanted	the club	to change the meeting location	Causes/Enables	He	got	elected
subject	verb	object1	object2		subject	verb	object1

○

He	wanted	the club	to change the meeting location	Causes/Enables	He	ran	for	president of the chess club
subject	verb	object1	object2		subject	verb	preposition1	object1

You will see the submit button when you reach the end of the questions.
Thanks for your hard work! If you encounter any issues, please contact us.

Figure 4: Sample qualification question about what events follow: Workers are asked which is a correct cause and effect, given the story: 'He wanted the club to change meeting times **causes or enables** he got elected' or 'He wanted the club to change meeting times **causes or enables** he ran for president of the chess club'. The second answer is correct. In the story, the boy's desire to have the time of the club changed convinces him to run for president of the club. Although getting elected president was an eventual outcome of his desire to change the meeting time, it is not as directly caused as his decision to run.

a valid cause-effect in the story and the others which were not. The test cases used specific statements to avoid conflate these questions with questions about how to write general rules. In figure 4, workers are asked which is a correct cause and effect, given the story: 'He wanted the club to change meeting times **causes or enables** he got elected' or 'He wanted the club to change meeting times **causes or enables** he ran for president of the chess club'. The second answer is correct. In the story, the boy's desire to have the time of the club changes convinces him to run for president of the club. Although getting elected president was an eventual outcome of his desire to change the meeting time, it is not as directly caused as his decision to run.[9]

3.1.3 Results

We found that the qualification task was sufficient to populate the main task with qualified workers. We found that workers were willing to engage with lengthy training documents in order to do well on the qualification HIT. One big motivating factor for workers was the size of the main task batches. Because

[9] The training did not exclude multi-hop reasoning, but it encouraged using more direct causes and effects over ones that would be separated in chains of causes and effects.

passing the qualification test HIT allowed them to do an unlimited amount of the thousands of GLUCOSE main task HITs that were available, they were very motivated to pass.

3.2 Warm-up Task

In our mid-rounds, we started an additional level of testing for entrance called 'the warm-up task' as a way to ease workers into the large-scale task, if they did got between 70% and 80% correct on the qualification task. The warm-up was a sample of HITs like those seen in the main task (described in Section 3.3 below), but it had simpler sentences and only seven HITs. We requested that workers submit three of the seven HITs, so that we had an ample sample-size to give feedback and see if they were ready for the large-scale task.

HITs were approved by an expert who also provided feedback on how to improve responses. The expert rated workers on their work quality on the sample to see if they qualified for the main task. If they got a score of '3', then they showed a very good understanding of the task, and they were added to the main task. If they got a score of '2', they were often writing rules that were too specific or too general. They were given feedback and asked to do a few more warm-up HITs to show improvement. If they improved, then they were added to the main task. For workers getting a score of '1', it was considered unlikely that they would be able to improve with the amount of expert feedback offered and were told to study the guidelines and re-take the qualification test HIT.

3.3 Main GLUCOSE Task

The main GLUCOSE task used the GLUCOSE data schema to collect specific statements and general rules. Qualifying workers were able to access large batches of data with no limit on how many HITs they could complete. Each HIT was a story/sentence pair with questions about each of the ten dimensions. Each annotator could only do a given story/sentence pair once. The following GLUCOSE main knowledge acquisition UI shows the exact task template that was launched as a HIT `https://bit.ly/2R8XcTt`.

3.3.1 Participants

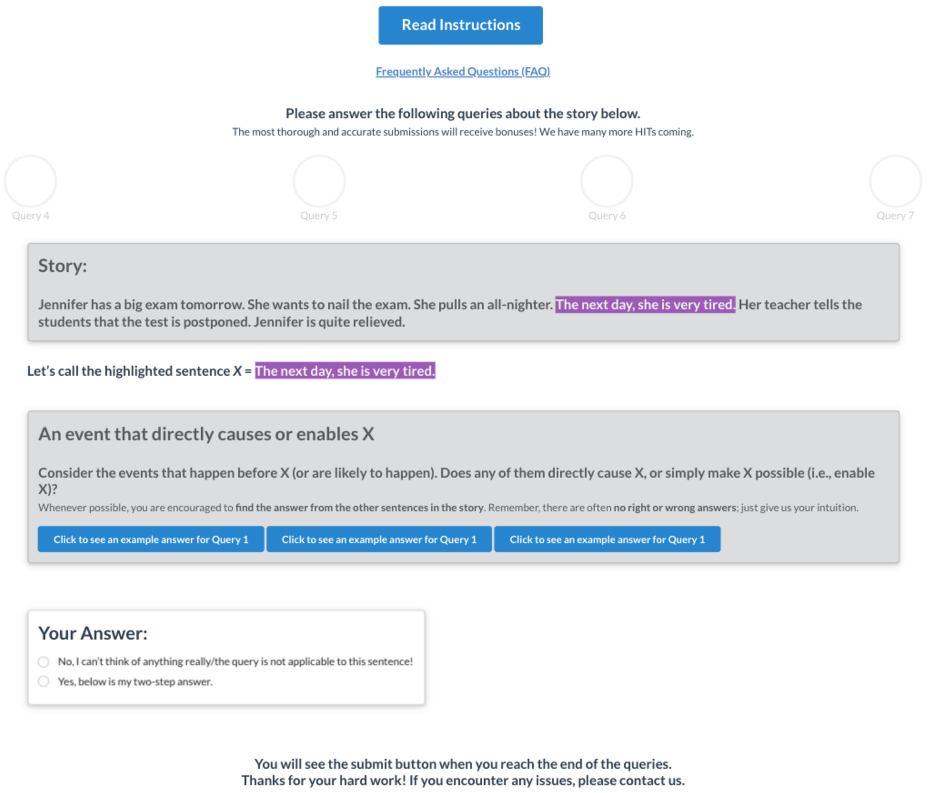
The general qualifications for workers on the main task was the same as for the qualification task, but with the added qualification of '10', indicating that they had been admitted directly through getting a high score on the qualifying test or by an expert after contributing to the warm-up task.

3.3.2 Main UI Content

Each time a worker accepted a main UI HIT, it displayed a page with the story, the target sentence, and the question about the story/sentence pair for each of the ten GLUCOSE dimensions (see figure 5 for an example). Each page also included 3-4 links to examples of correct answers for the same dimension in other sentence/story pairs (see the three buttons in blue in figure 5). If workers said that the dimension was relevant for the selected sentence (by clicking 'Yes..' in the "Your Answer:" box in figure 5), the task progressed to a screen where they could input answers, as shown in figure 6. This took place for each dimension. Workers were required to respond 'Yes' and fill in at lest 2 dimensions per HIT.

We choose not to randomize the dimension order because the task was already long, and the predictable order of questions was something that helped workers form a strategy for doing the task in a manageable amount of time.[10]

[10] The average time it took a worker to complete one HIT was 8 minutes and 2 second, and the median was 5 minutes and 8 seconds.


Read Instructions
Frequently Asked Questions (FAQ)
Query 4
Query 5
Query 6
Query 7
Click to see an example answer for Query 1
Click to see an example answer for Query 1
Click to see an example answer for Query 1

Please answer the following queries about the story below.

The most thorough and accurate submissions will receive bonuses! We have many more HITs coming.

Story:

Jennifer has a big exam tomorrow. She wants to nail the exam. She pulls an all-nighter. The next day, she is very tired. Her teacher tells the students that the test is postponed. Jennifer is quite relieved.

Let's call the highlighted sentence *X* = The next day, she is very tired.

An event that directly causes or enables X

Consider the events that happen before X (or are likely to happen). Does any of them directly cause X, or simply make X possible (i.e., enable X)?

Whenever possible, you are encouraged to **find the answer from the other sentences in the story**. Remember, there are often **no right or wrong answers**; just give us your intuition.

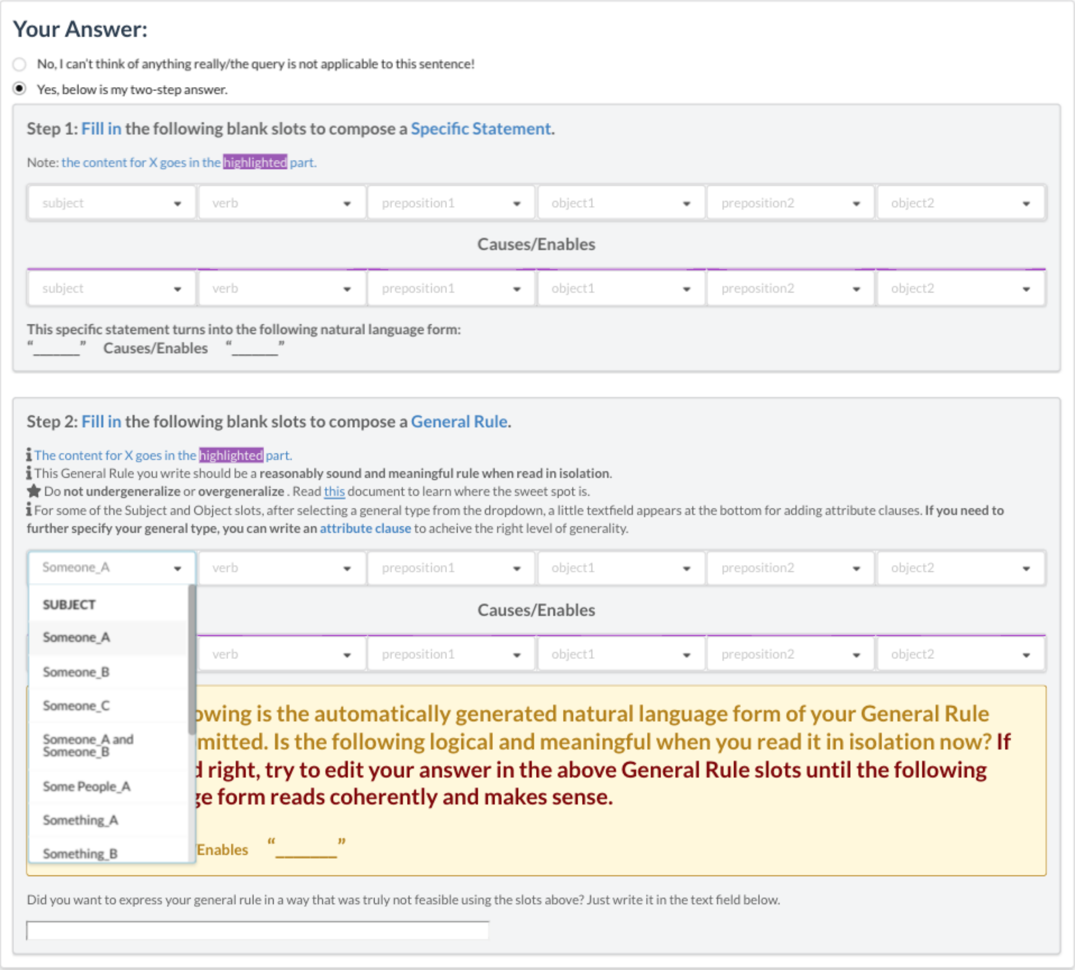
Your Answer:

○ No, I can't think of anything really/the query is not applicable to this sentence!

○ Yes, below is my two-step answer.

You will see the submit button when you reach the end of the queries.
Thanks for your hard work! If you encounter any issues, please contact us.

Figure 5: The Main UI features a question about the story/sentence pair for each dimension.



Your Answer:

○ No, I can't think of anything really/the query is not applicable to this sentence!

● Yes, below is my two-step answer.

Step 1: Fill in the following blank slots to compose a Specific Statement.

Note: the content for X goes in the highlighted part.

subject | verb | preposition1 | object1 | preposition2 | object2

Causes/Enables

subject | verb | preposition1 | object1 | preposition2 | object2

This specific statement turns into the following natural language form:
"______" Causes/Enables "______"

Step 2: Fill in the following blank slots to compose a General Rule.

ℹ The content for X goes in the highlighted part.
ℹ This General Rule you write should be a **reasonably sound and meaningful rule when read in isolation**.
★ Do **not undergeneralize** or **overgeneralize** . Read this document to learn where the sweet spot is.
ℹ For some of the Subject and Object slots, after selecting a general type from the dropdown, a little textfield appears at the bottom for adding attribute clauses. **If you need to further specify your general type, you can write an** attribute clause to acheive the right level of generality.

Someone_A | verb | preposition1 | object1 | preposition2 | object2

Causes/Enables

verb | preposition1 | object1 | preposition2 | object2

owing is the automatically generated natural language form of your General Rule mitted. Is the following logical and meaningful when you read it in isolation now? **If d right, try to edit your answer in the above General Rule slots until the following ge form reads coherently and makes sense.**

Enables "______"

Did you want to express your general rule in a way that was truly not feasible using the slots above? Just write it in the text field below.

Figure 6: When "Yes" is selected for "Your Answer:" on the main UI in figure 5, workers can input answers. Answers are guided by slots and drop-down selections.

For the specific statements, users freely entered text within the constraints of some part-of-speech guidance for some dimensions, but, for others, there were constraints on verbs or emotions. A list of the contents of the drop-down slots with suggestions is shown in table 2. All preposition slots contained a drop-down list of English prepositions. For general rules, the subject position had a drop-down menu of variables for people, places, and things, such as ‘Someone_A’. The verb slot was constrained in the same way as with the corresponding specific statement for that dimension.

3.3.3 Results

We were able to collect about 335K pairs of specific statements and general rules using the main task UI.

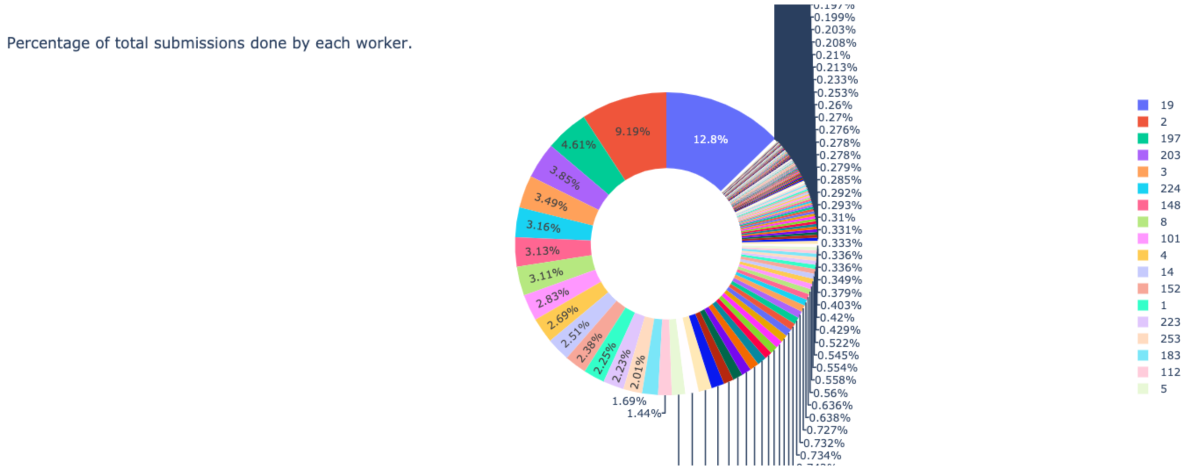


Figure 7: The data was annotated by a diverse set of workers with no one worker doing more than 12.8% of the data

Of the 1039 workers who qualified for the main GLUCOSE task via the qualification task and warm-up, 373 contributed to the main task. Our worker pool was diverse with no particular worker contributing more that 12.8% of the data (see figure 7). A smaller number of workers contributed the majority of the HITs, which was something we encouraged by giving high bonus rates and bulk incentive bonuses to our best workers. A large number of workers contributed fewer than 70 HITs.

Many workers who qualified found the main task to be too time consuming to complete. We reached out to them several times. Their responses were that either the task was too complicated for the pay rate or that they felt uncertain of how to go from multiple-choice questions to constructing specific statements and general rules. In order to help, we gave some of them entrance to the warm-up task and provided feedback. Even with the additional recruitment measures, 666 workers did not choose to do a single HIT on the main task after receiving qualification. Although we would like to have seen more workers qualified workers contribute, we were still able to reach our goal and even annotate more HITs than planned. Qualifying the majority of the workers with an automatically scored multiple-choice qualification test was overall beneficial for recruiting qualified workers for the main task.

3.4 Data Quality Control Pipeline

For work contributed through the main UI, data quality was controlled through daily monitoring of a percentage of incoming submissions. Additional monitoring was done by checking daily statistics, such as the

Dimension	Task Connective	Slot Constraints
dim 1 An event that directly causes or enables X	Causes/Enables	none
dim 2 An emotion or basic human drive that motivates X	Motivates	verb [feels, wants, likes] object [curiosity, independence, competition, honor, approval, power, status, romance, success, friendship, belonging, health, safety, livelihood, happy, stressed, angered, disgusted, sad, surprised, fearful, trusting, love, obedient, amazed, disappointment, regret, worthless, aggression, optimistic]
dim 3 A location state that enables X	Enables	verb [am is are] preposition [above, across from, at, below, far from, in, in front of, inside of,near, next to, on top of, outside of]
dim 4 A possession state that enables X	Enables	verb [possesses]
dim 5 Other property besides location, emotional state, and possessions make X possible	Enables	verb [am, is, are, has, have, want, wants, need, needs]
dim 6 An event that is directly caused or enabled by X	Causes/Enables	none
dim 7 An emotion that is caused by X	Motivates	verb [feels, wants, likes] object [curiosity, independence, competition, honor, approval, power, status, romance, success, friendship, belonging, health, safety, livelihood, happy, stressed, angered, disgusted, sad, surprised, fearful, trusting, love, obedient, amazed, disappointment, regret, worthless, aggression, optimistic]
dim 8 A change of location that X results in	Enables	verb [am is are] preposition [above, across from, at, below, far from, in, in front of, inside of,near, next to, on top of, outside of]
dim 9 A change of possession that X results in	Enables	verb [possesses]
dim 10 Other change in property (besides location, emotional state, or possession) that X results in	Enables	verb [am, is, are, has, have, want, wants, need, needs]

Table 2: The content of each drop-down list workers could choose from per dimension, as described in section 3.3

average number of dimensions filled out and the average time workers took to complete HITs. The samples were used by experts to modify worker quality ratings and provide feedback to workers on how to improve.

Figure 8 shows the strategic flow of worker ratings. Workers enter the tasked with a score of “-1” then advance to “2” as they become more proficient, getting a pay increase in the form of a bonus. The top numeric rating is “3”, which has an additional increase in bonus size. Select workers with a “3” rating were also moved into “top rated” batches that paid more per HIT and included higher bonuses and incentives. If work quality dropped, workers’ ratings were reduced. If their work was at a risk of degrading corpus quality, they were given a “0” and disqualified from the task. When a worker’s status was set to “0”, all incoming submissions were approved but the worker did not receive a bonus. Several workers were disqualified and then, over time, worked to re-qualify and became top rated workers. Referring workers to the General Guidelines and giving individual feedback on answers often provided sufficient remedial learning to get good responses, if work quality slipped. Most data quality issues were due to workers trying to rush through the task and not reading the General Guidelines carefully. Figure 9 shows a state of the percentage of contributions to the data by worker score.

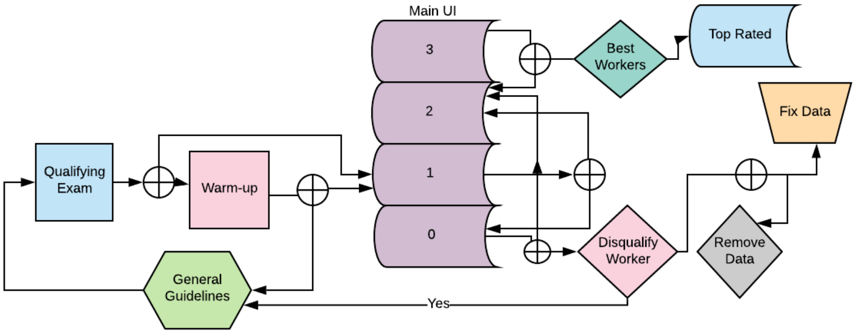


Figure 8: Data collection pipeline. Numeric values in the Main-UI (lavender) correspond to worker ratings. Workers enter the pipeline through the Qualifying Exam. The then proceed either to the warm-up or main UI task. In the main UI tasks, workers are given ratings based on work quality. A variety of remedial interventions are used when quality levels fall.

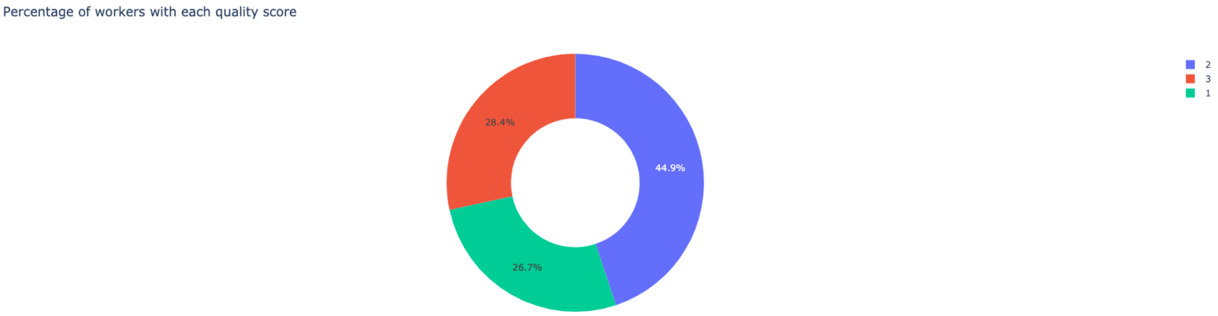
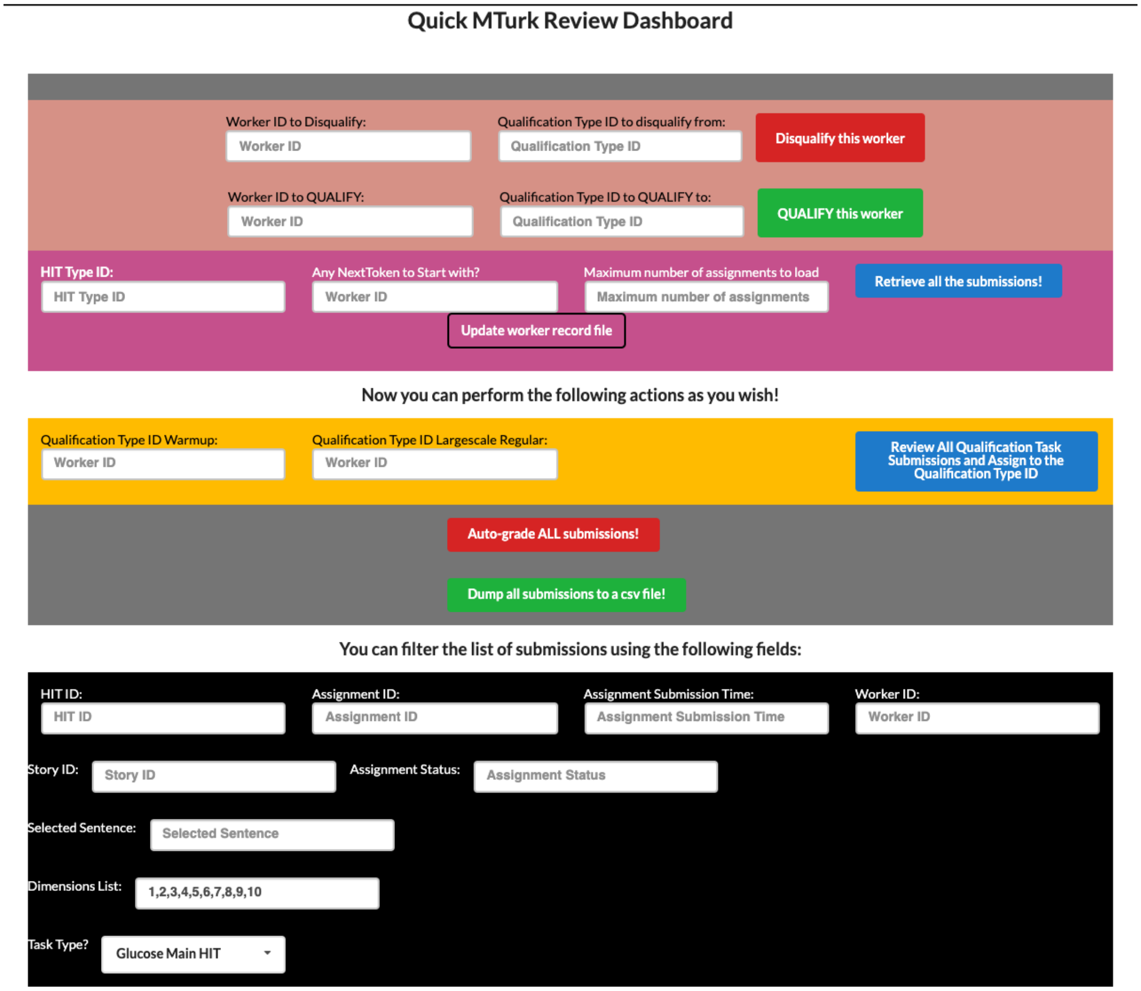


Figure 9: Workers by quality score rating, Purple = ‘2’(mid), Red = ‘3’(high), , Green = ‘1’(lower)

3.4.1 Review Dashboard

Incoming submissions to the large batches were monitored daily through the in-house review dashboard for quality control. The dashboard is shown in Figure 10.



Quick MTurk Review Dashboard

Worker ID to Disqualify: Worker ID
Qualification Type ID to disqualify from: Qualification Type ID

Worker ID to QUALIFY: Worker ID
Qualification Type ID to QUALIFY to: Qualification Type ID

HIT Type ID: HIT Type ID
Any NextToken to Start with? Worker ID
Maximum number of assignments to load: Maximum number of assignments

Now you can perform the following actions as you wish!

Qualification Type ID Warmup: Worker ID
Qualification Type ID Largescale Regular: Worker ID

You can filter the list of submissions using the following fields:

HIT ID: HIT ID
Assignment ID: Assignment ID
Assignment Submission Time: Assignment Submission Time
Worker ID: Worker ID
Story ID: Story ID
Assignment Status: Assignment Status
Selected Sentence: Selected Sentence
Dimensions List: 1,2,3,4,5,6,7,8,9,10
Task Type? Glucose Main HIT

Disqualify this worker
QUALIFY this worker
Retrieve all the submissions!
Update worker record file
Review All Qualification Task Submissions and Assign to the Qualification Type ID
Auto-grade ALL submissions!
Dump all submissions to a csv file!

Figure 10: Data review dashboard allows reviewers to see user responses, sorted by desired parameters, and give feedback to workers

Submissions displayed through the review UI could be filtered by batch, submission status (e.g., Submitted or Approved), worker identification number, date of submission, story identification number, dimension, and sentence. When work in need of approval was displayed, it showed the worker's current quality rating and any notes on the worker's history. These ratings were maintained in a spreadsheet that was uploaded to the UI as changes were made.

4 Preparation and Analysis of the Final Data Set

After we completed the collection task, we did additional post-processing of the data. This section discusses that process and the results of the final data.

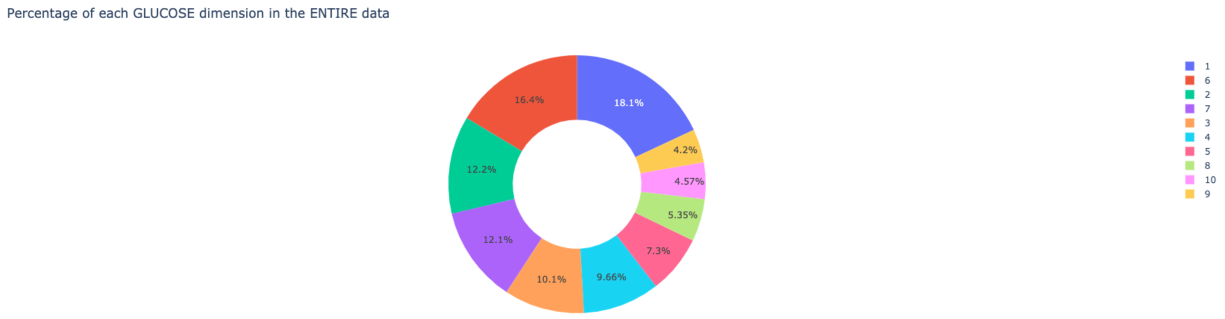


Figure 11: Data collected by dimension in the total data

4.1 Data Post-Processing

In order to get more granularity in the final worker quality ratings, we coordinated the commit date of worker quality rating changes with the submission date of the HITs. The increased granularity of ratings improved our quality level such that we saw 80% very high quality answers when filtering a random 50 samples from workers rated as 3.

To clean the data even more, we got rid of many ratings used in the initial classification. Of the ratings '3', '2', '1.5', '1', '0', '-1', and '-2', we kept only '3', '2', and '1'. Ratings were changed with the following policies: For workers who contributed more than 200 HITs to the data ranked as '3' (20 workers total), we reviewed the quality of their work by rating and moved the ratings, as appropriate. For instance, if a review of their '3' rated work showed it to be of lower quality than expected, we included a policy in a cleaning script for moving all of their '3' rated work to '2' rated work.

For the ratings we were not using, '1.5', '0', '-1', and '-2', we reviewed the set of work and moved it by scripted policy to the appropriate rating. For workers with 100-200 HITs in the '3' rating, we checked only their '3' rated work and demoted it, if necessary. For the remainder of workers, we created a policy for moving their '-2' ratings to '2'. This choice was because the '-2' rating was used when workers were going too fast or not filling out more than 1 or 2 dimensions. The content of their mini-theories was not bad, but rather they weren't contributing as many mini-theories per HIT. For the remaining scores, '1.5', '0', and '-1', the scores were moved to '1'. After running the policy script, we reviewed the data again and found much better consistency across ratings. A random sample of 100 answers ranked as '3' showed 92% of the data to be up to the highest quality standard.

4.2 Data Statistics

Throughout the dataset, workers have filled in an average of 4.7 dimensions, with the median being 4 dimensions.

Figure 11 shows the distribution of dimensions filled out by type. We expected dimensions 5 and 10 to be less frequent (Is there anything else that causes or enables the sentence?), due to the fact that they refer to rare changes of state, such as becoming wet. We saw some bias towards answering more of the first questions than the later ones, offset by an overall preference for dimensions 1 and 6, the causality dimensions.

Dimensions 8 and 9 were less frequent than their counterparts 3 and 4. The main reason is because a location state is often part of the scene setting or background of a story, whereas a location state that is the result of the highlighted sentence only occurs when a change of location occurs. Similarly, there are things

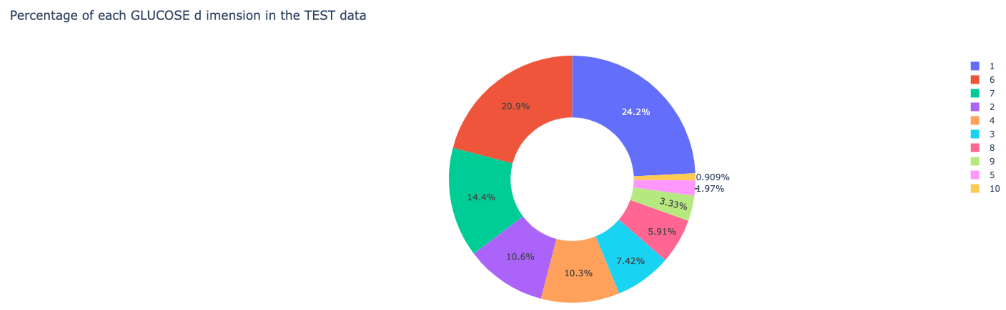


Figure 12: Data collected by dimension in the test data

possessed at the outset of a story, as part of the scene it occurs in, that enable things that are talked about in the story, however, for a sentence to lead to a possession state, it needs to be the case that something is acquired.

Of the final data, we used a portion of the data ranked as quality level '3' to construct a test set. This test set was judged by a group of human annotators and determined to be 'correct' or 'nearly correct' for all dimensions[11]. We consider the test data to be representative of the larger data set in terms of which dimensions are relevant. Figure 12 shows the percentage of each dimension filled out in the test data.

5 Acknowledgements

This work could not have been completed without the group of Mturk workers who contributed their time and effort. We are thankful for their patience as we iterated the task instructions and quality control.

[11] Please refer to the main published paper for the details `https://arxiv.org/abs/2009.07758`.