15.03.2024

Human Evaluation (HumanEval) Codebook

Introduction

Purpose of the codebook: To provide detailed information on each variable in the HumanEval dataset collected during the AligNet project. In the AligNet project we study Human alignment of machine learning models, which is becoming central to representation learning (e.g., Muttenthaler et al. 2023, Sucholutsky et al., 2023). Our goal is to develop models that align closely with human perception and intentions. To accomplish this objective, with this study, we aim to collect a set of **human odd-one-out similarity judgments** to evaluate whether (synthetically generated) similarity judgments of existing state-of-the-art human-aligned models (Muttenthaler et al., 2023) correspond to ground-truth human judgments.

This codebook provides information on all variables used in the **Odd-One-Out Online Task.**

General Information

Dataset Name: HumanEval Dataset
Creator: Frieda Born
Date of creation: 03-2024
Version 1.0
Acess Information: The dataset is openly available here: (fill in detailed information here)

Variable descripltion

Note: The variables that are indicated in bold are the important variables for data anylsis:

0) consent_status
1) rt
2) image1Path
3) image2Path
4) image3Path
5) selected_image
6) exp_trial_type
7) response (demographic information)

Name	Type	Description	Coding	Related variables
rt	numeric	reaction time	Number in millisecond	/
url	string	URL of external used html elebemt that accommodates consent form	/	consent_status, consent_button_press 1, **consent_button_pre ss2**
trial_type	string	definition of experiment plugin	external-html, survey, preload, instructions, html-keyboard-response, ...	/
trial_index	numeric	ascending number that strechtes throughout the exp. and increase with every new trial (not only exp. also instructions, training, etc.)	0,1,2 - N	/
internal_node_id	string	unique identifier assigned to each trial or timeline node	X.X-X.X-X.X	/
start	numeric	start of experiment	Date / Time	end_time

subject_id	string	unique participant identifier	participant_X	/
study_id	string	unique study ID assigned from prolific		/
OS	string	operating system of participant	e.g., Win32, MacOS	/
prolific_id	string	unique prolific ID. Note: this is removed for anonymized data	e.g., string of numbers and letters	/
session_id	string	unique session ID assigned to each participant by prolific	string of numbers and letters	/
fullscreenexit	categorical	variable set to FALSE when participant enters fullscreen	FALSE = participant is in fullscreen TRUE = participant exited fullscreen	interaction_data
consent_status	string	indicates if participant has given consent (logs summary of consent status)	1) Participant has read consent form and consents to study participation. 2) Participant does not consent and is redirected	consent_button_press 1 consent_button_press 2
consent_button_press2	string	logs button press in consent form (logs consent status)	1) I consent 2) I do not consent	consent_button_press 1
consent_button_press1	string	logs button press in consent form (logs that consent form was read)	1) I have read the above consent form	consent_button_press 2
experiment_complete	categorical	is set to TRUE at the end of the experiment when participant finished the whole task	TRUE = participant finished exp. NA = exp. was not finished; data probably incomplete	/
end_time	numerical	end time of experiment	Date / Time	start
success	categorical	can indicate different things depending on the plugin it is refering to (e.g., participant enters fullscreen as expected then this variable is set to TRUE).	TRUE FALSE NA	Plugins used in experiment
response	string	reponse to demographic survey	{"P0_Q0":null,"Gender":"Female/Male/Divers/I prefer not to say","Age":"18-X"}	/
timeout	categorical	timeout set for a specific part of the task (e.g., max break time 2 minutes)	TRUE = timeout was reached FALSE = reponse within time NA	/
failed_images		contains the paths or identifiers of any images that could not be successfully loaded or displayed during the experiment		
failed_audio		contains the paths or identifiers of any audios that could not be successfully loaded or displayed during the experiment		
failed_video		contains the paths or identifiers of any video that could not be successfully loaded or displayed during the experiment		
view_history		relevant for instruction only where it documents how long each instruction page was viewed	e.g.,{"page_index":0,"viewing_time"Xms},{"page_index":1,"viewing_time":X},{"page_index":2,"viewing_time":X}	
stimulus	string	stimulus array that document the set of stimuli (e.g., triplet of		image1Path, image2Path, image3Path

		images) presented in a trial		
image1Path	string	separate stimulus name of the triplet image on the left	stimuli/imagenet2012_human_eval_v1.1_images/image_name.JPEG	image2Path, image3Path
image2Path	string	separate stimulus name of the triplet image in the middle	stimuli/imagenet2012_human_eval_v1.1_images/image_name.JPEG	Image1Path, Image3Path
image3Path	string	separate stimulus name of the triplet image on the right	stimuli/imagenet2012_human_eval_v1.1_images/image_name.JPEG	Image1Path, image2Path
exp_trial_type	categorical	definition of trial type, do not confuse with "trial_type", which refers to the plugin type	1) training_trial 2) exp_trial 3) break 4) catch_trial → should be used to filter for variables of interest in analysis	
catch_trial_correct	categorical	checks if in catch trial predefined most obvious odd-one-out was selected	TRUE = correct FALSE = incorrect Null = we are not in a catch trial	
keypress	categorical	logs the keypress during odd-one-out choice	1)ArrowLeft 2)ArrowRight 3) ArrowDown 4)NoResponse (when no selection was made -> timeout	
selected_image	string	logs the image that was selected as the odd-one-out	stimuli/imagenet2012_human_eval_v1.1_images/**image_name_selected.JPEG**	Image1Path, image2Path, Image3Path
break_ending	string	logs the time after which particiant ended break (if < 2 minutes of max)	ended by participant's action after XXX ms	
interaction_data	string	refers to data collected about the participant's interactions with the browser during the experiment. This data is part of the automatic data collection features of jsPsych and is aimed at recording events that might influence the participant's performance or the integrity of the data collected during the experiment	It is added to response collection at the end of the experiment. It can involve event like: 1) blur 2) focus 3) fullscreenenter 4) fullscreenexit	fullscreenexit
test_part	string	documents timeout message when participant fails to respond in a triplet trial	timeout_message = means participant failed to respond within 15s	

Data Collection Methods

Data was collected through an online experiment hosted on the Servers of the Max Planck Institute for Human Development, Berlin. Participants were recruited via Prolific (https://www.prolific.com/).
The experiment uses jsPsych v7.3.3 (https://www.jspsych.org/7.3/), which is a JavaScript library for creating and running behavioral experiments in a web browser. Some of the base scripts for the jsPsych plugins where customized to accommodate additonal functionality. Changes are made transparent in comments within the plugin codes (see GitHub repository).

Data Processing and Cleaning

Data processing involved removing prolific IDs and any other direct identifiers to anonymize data collected from the experiment.

Data Quality Assurance

Quality Checks: Conducted data entry verification and consistency checks of reaction times and selected odd-one-out images, catch trial checks, etc.

Change Log

Version History: V1.0 - Initial release. No subsequent changes.

Contact Information

Contact Details: For further information, contact Frieda Born (born@mpib-berlin.mpgd.de)

Appendices

Additional Documentation: Detailed methodology descriptions and experiment code can be found on GitHub here: https://github.com/Frieda-Josefine/HumanEval