ChatGPT outperforms crowd-workers for text-annotation tasks

Description of Replication Material, Datasets, and Scripts

Fabrizio Gilardi Meysam Alizadeh Maël Kubli

This file describes all datasets and R scripts required to rerun and reproduce the analyses reported in the paper and Supporting Information. To have the data, scripts, and outputs saved in the correct folders, which is required to run the scripts successfully, we recommend downloading all files as a single ZIP file (click the "Access Dataset" button in the right-hand corner and select "Original Format ZIP"). If you only need selected files, make sure to download data in the original file format. Please do not hesitate to contact the authors if you have any questions.

Datasets

The repository for replicating the empirical analyses contains the following datasets. The following files (in alphabetic order) are used in the R and Python scripts listed below. All other files in the repository, which are not listed here, are intermediate products of the analysis and are there to save time if only some steps of the analysis should be reproduced.

- *analysis_data_ChatGPT_results.csv* Main Data for Analysis
- *annotation_data_raw.xlsx*
- *annotation_data_raw_newspapers_raw.xlsx*
- *annotation_data_ra_completed.xlsx*
- *annotation_data_tweetscongress_raw.xlsx*
- *annotation_data_tweets23_raw.csv*
- *annotation_data_tweets23_ra_completed.xlsx*
- *batch_results_frames_final.csv*
- *batch_results_newsarticles_problem_solution_final.csv*
- *batch_results_newsarticles_relevance_final.csv*
- *batch_results_problem_solution_final.csv*
- *batch_results_relevance_final.csv*
- *batch_results_stance_final.csv*
- *batch_results_topics_final.csv*
- *batch_results_tweets23_problem_solution_final.csv*
- *batch_results_tweets23_relevance_final.csv*
- *batch_results_tweetscongress_frames_final.csv*
- *batch_results_tweetscongress_relevance_final.csv*
- *section230_training_data_completed.xlsx*
- *training_data_problem_solution.csv*
- *training_data_relevance.csv*

- *training_data_tweets23_problem_solution.csv*
- *training_data_tweets23_relevance.csv*
- *training_frames.csv*
- *training_data_frames_final.csv*
- *training_data_newsarticles_problem_solution_final.csv*
- *training_data_newsarticles_relevance.csv*
- *training_data_problem_solution.csv*
- *training_data_problem_solution_final.csv*
- *training_data_relevance_final.csv*
- *training_data_stance_final.csv*
- *training_data_topics_final.csv*
- *training_data_tweets23_problem_solution_final.csv*
- *training_data_tweets23_relevance.csv*
- *training_data_tweetscongress_frames_final.csv*
- *training_data_tweetscongress_relevance.csv*
- *training_data_stance_final.xlsx*
- *training_data_tweetscongress_frames.csv*
- *tweets_annotation_data_ra_completed.xlsx*

Scripts

The repository contains the following scripts. If you want to reproduce the complete analysis:

- 01-01-mTurk_prep_data.R
- 01-02-mTurk_prep_data_newsarticles.R
- 01-03-mTurk_prep_data_tweets_2023.R
- 01-04-mTurk_prep_data_tweets_congress.R
- 02-01-mTurk_eval_data.R
- 02-02-mTurk_eval_data_newsarticles.R
- 02-03-mTurk_eval_data_tweets2023.R
- 02-04-mTurk_eval_data_tweets_congress.R
- 03-01-chatgpt-Zeroshot-Task-template.py
- 04-01-figures_annotations.R
- emoji_regex_vector.txt