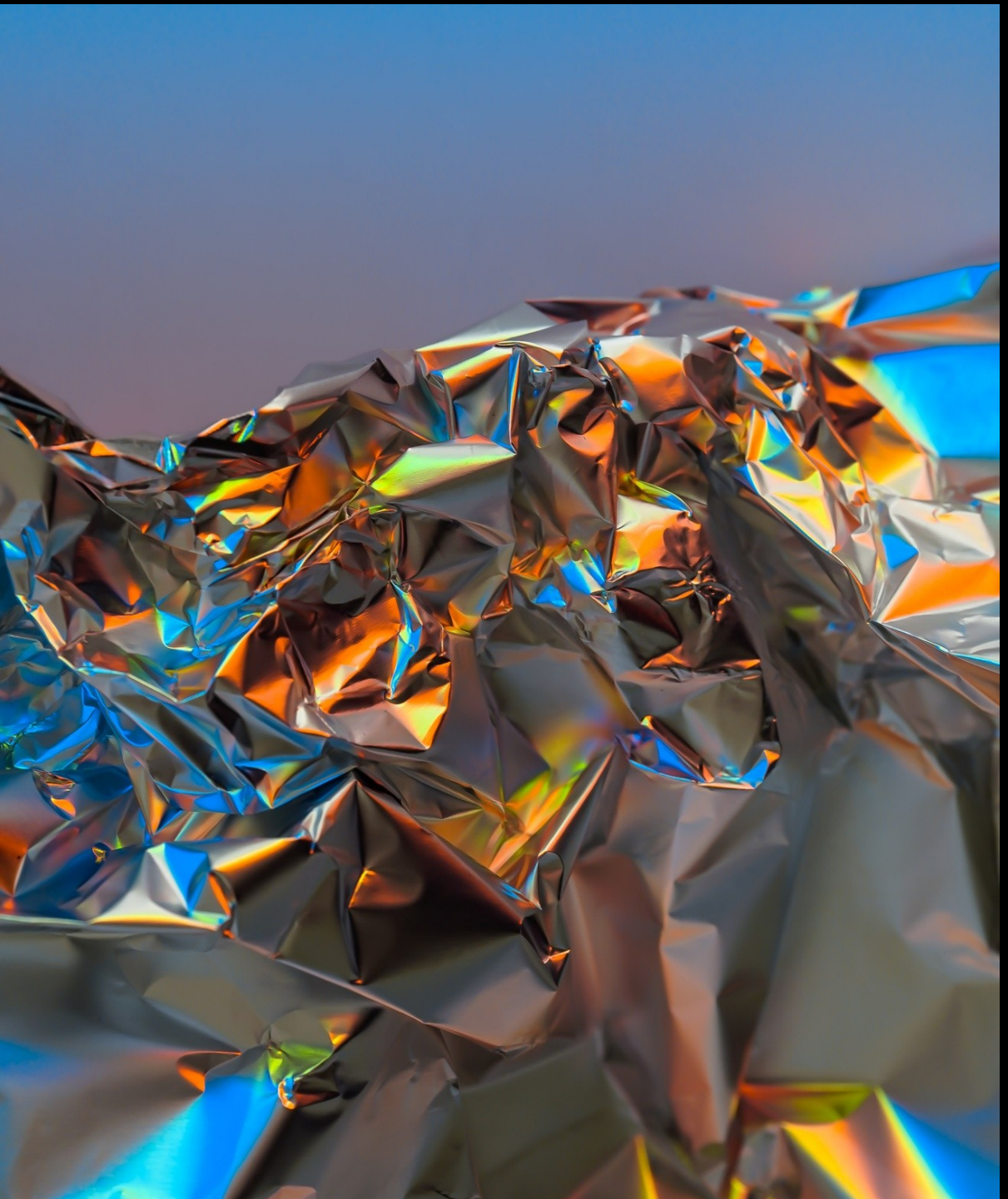


ChatEval: A Tool for Chatbot Evaluation

Authors: Joa˜o Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, Chris Callison-Burch

University of Pennsylvania

Introduction

Evaluating open-domain dialog systems (chatbots) is challenging due to the reliance on **human judgments** and lack of **standardized procedures**.

Limited transparency: Model parameters and code are often unpublished, hindering systematic comparison.

ChatEval is a web-based platform that:

- **Standardizes** the human evaluation process for chatbots.
- Provides a **hub for researchers** to share and compare their models.
- Ensures **transparency** with open-source evaluation code, baseline models, and datasets.

High-level Summary

- ## Challenges in Chatbot Evaluation

  Lack of standardized evaluation metr procedures make reproducibility anc assessment difficult for open-domair systems

- ## The ChatEval Framework

  Introduces a scientific framework for evalu chatbots, including an open-source codebase and web portal for sharing resource

- ## Human Evaluation in ChatEval

  Uses two-choice comparison tests wi response theory to statistically evaluate cha responses

- ## Automatic Evaluation in ChatEval

  Includes metrics like lexical diversity, co similarity, BLEU-2 score, and perplexity

- ## Baseline Models and Datasets

  Provides baseline for Seq2Seq models layer LSTM, bidirectional encc unidirectional decoder) trained on T SubTle and OpenSubtitles datasets. Dialogue Breakdown Detection Chal (DBDC) datasets a benchmark for ch evaluation

The ChatEval Web Interface

- ## Model Submission Form

  Researchers submit their models for evalua
  by uploading the model's responses on at
  one of the evaluation datasets. They also provi
  a description of the model, including o
  links to a paper, project page, code repos
  and pre-trained model

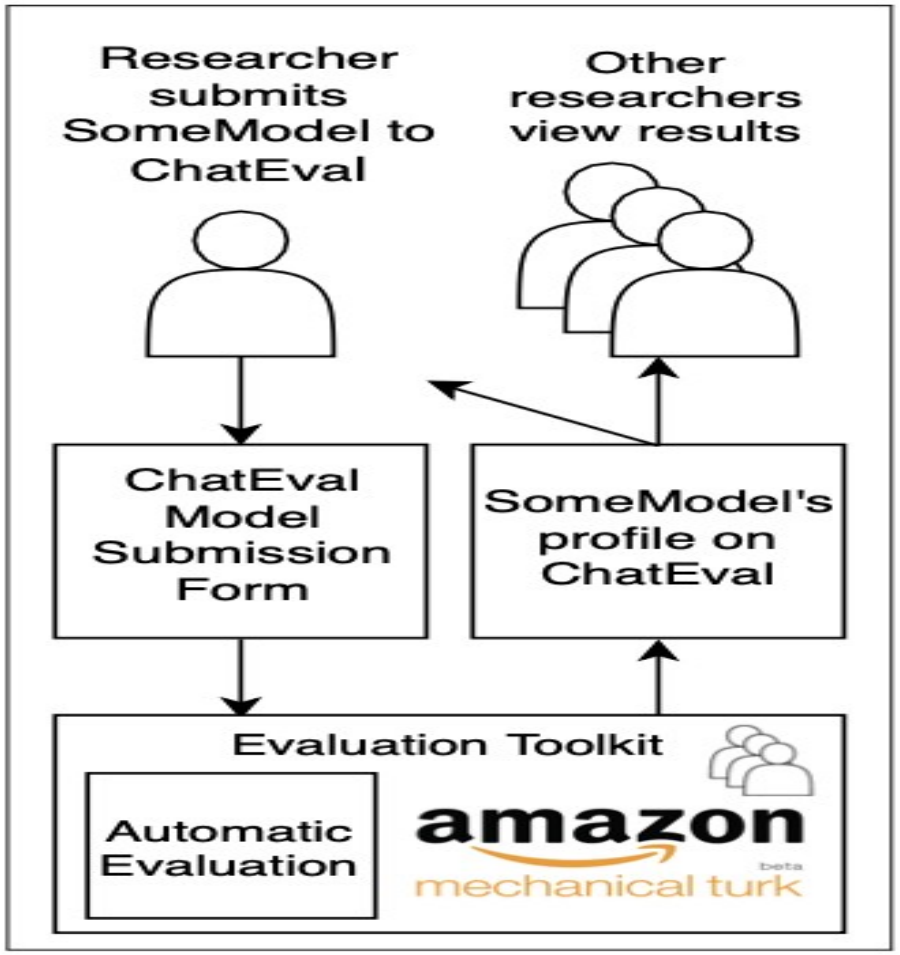
- ## Model Profile

  Each submitted model, as well as
  models, have a profile page that displa
  model's description, responses to eva
  prompts, and visualizations of hur
  automatic evaluation results

- ## Response Comparison

  Users can view and compare the r
  generated by different models for each promp
  the evaluation datasets

Flow of information in ChatEval



Evaluation Toolkit

Automatic Evaluation Metrics

- Lexical Diversity: Measures the diversity of the mode vocabulary
- Average Cosine Similarity: Compares the gen responses to the ground-truth
- Sentence BLEU-2 Score: Evaluates the quality c generated responses
- Response Perplexity: Measures the likelihood model's predicted response

Human Evaluation

The human evaluation component of ChatEval uses comparison tests, where evaluators (Mechanical Turk compare the model's resp prompts and select the better one or indicate a tie evaluations are done in blocks of 10 prompts, evaluators per prompt

Consistency and Generalization

To ensure the evaluation is consistent and the model not overfit, the ChatEval team plans to performing models to additional datasets beyond the init evaluation sets, including datasets from prior work such Baheti et al. (2018) and Li et al. (2019

Open-Source and Transparency

The ChatEval toolkit is available on GitHub, researchers to run the human and automatic evaluati on their own models before submitting them to the pul ChatEval website. The raw evaluation data, including A assessments, is also publicly available in database and JSON format

Strengths

Standardized Evaluation Framework

The paper introduces ChatEval, framework for conducting automatic and hur evaluations of chatbots in a consistent manne

Publicly Available Evaluation Datasets

ChatEval provides access to a evaluation datasets, including the DBDC data and subsets of Twitter and Oper enabling researchers to test their moc relevant benchmarks

Open-Source Baseline Models

The toolkit includes several baseline models trained on standard datasets, allowing f easy comparisons between new and e models.

Transparent Evaluation Process

The paper highlights the importance of publishi model code and parameters, as well as the data from human evaluation experimer promote transparency and reproducib chatbot research

The ChatEval framework offers a standardized, open, and transparent approach to evaluating chatbots, enabling researchers to compare models and drive progress in the field of open-domain dialog systems.

Weaknesses

Low Inter-Annotator Agreement

The paper notes that the overall in
annotator agreement (IAA) f
human evaluations conducted throu
ChatEval varies and is oft
ranging from 0.2 to 0.54 when includir
tie choices

Limited Evaluation Criteria

ChatEval emphasizes task com
fluency, and engagement but lacks focus c
conversational context,
intelligence, and nuanced user satisfactic
This makes it less comprehensive
holistic models like cor
intelligence challenge or newer evaluatic
integrating empathy and

Expensive Human Evaluations

While crowd-sourcing can be
effective approach, the paper acknowledge
that human evaluations can
expensive to obtain, espe
systematic comparisons across m
chatbot models

The paper highlights several weaknesses in the ChatEval framework, including low inter-annotator agreement, limited evaluation criteria, and the overall expense of conducting human evaluations, which are crucial for the systematic comparison of chatbot models.

Relation to Interactive Fiction and Story Generation

Storytelling Needs Similar Qualities to Chatbots

In interactive fiction and storytelling, immersive, relevant, and coherent character interactions are essential for maintaining player engagement and suspension of disbelief.

Using ChatEval for Story Evaluation

Relevance and Coherence: ChatEval's metrics for relevance and coherence can be used to ensure character interactions are logical and consistent with the story's context and prior events.

Fluency: It can evaluate the narrative flow, dialogue quality, and ensure responses feel polished and natural.

Automatic Metrics for Scaling IF Testing

ChatEval's automatic metrics enable the scaling of interactive fiction (IF) testing by quickly evaluating if a narrative responds accurately to a wide range of player actions, reducing the reliance on human play testers.

Expanding ChatEval’s Potential in IF and Storytelling

Potential New Metrics for IF

Adapting ChatEval to IF could involve creating new metrics that measure story-specific qualities, like *plot progression* (does each interaction move the story forward?) and *character consistency* (are characters acting in ways that align with their established personalities?).

Game-Like Evaluations

ChatEval could test the quality of player-NPC (non-player character) interactions, helping evaluate branching dialogues for responsiveness and variety. This would be especially useful in complex interactive stories where player choices impact narrative progression.

Expanding on Creativity Metrics

ChatEval could be extended to measure creativity in generated text, evaluating originality in responses, a key factor for engaging storytelling and interactive fiction experiences.

Conclusion

ChatEval provides a comprehensive evaluation approach for chatbot development, combining both automatic and human-like metrics to assess quality. Its framework has the potential to enhance interactive storytelling by adapting these metrics to evaluate narrative quality, relevance, and coherence, fostering more immersive and engaging experiences. By applying ChatEval's principles, future interactive storytelling platforms can improve their quality assessment, leading to richer and more dynamic interactive story generation.

Thank you