Chatbot Evaluation: Methods and Challenges

Ruhan Circi Ph.D.

Bhashithe Abeysinghe Ph.D.

American Institutes for Research

Agenda

1. Introduction: Uses of Chatbots

2. Methods to Develop Chatbots

3. Approaches to Evaluate the Chatbots

4. Toward a Framework on Chatbot Evaluation

    a. Quality criteria for varying use cases

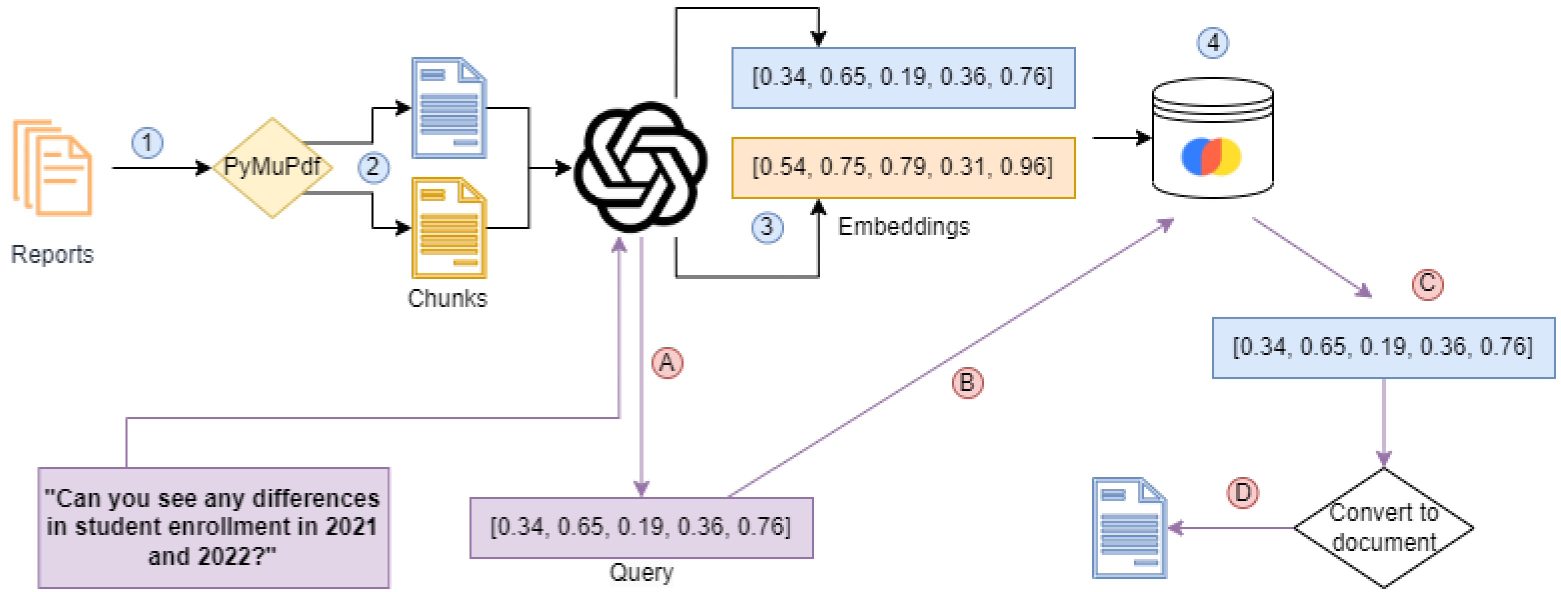
    b. Handling challenges

Introduction

- LLMs are (e.g., GPT-3.5, Llama etc.) computational models capable of generating language and other natural language processing tasks (Radford et al., 2019)
- LLMs have been evolving and hence has been applied in many areas
  - "Governments worldwide have sought to employ artificial intelligence (AI) to improve public services" (Mehr, 2017)
  - Most influential technology for government organizations
  - *Including in Education*
    - » Item Scoring (Circi & Perkoff, 2024)
    - » Item Generation (Attali et al., 2022; Circi et al., 2023; Hwang et al., 2023; Abeysinghe & Circi, 2024a)

Common Methods Used for Chatbot Development

- 2 components
  1. Retrieving information – semantic retriever
  2. Answering questions with relevant context – LLM



Introduction

- Most chatbots are now powered by LLMs
  - EMMA (DHS) – for immigration service information
  - Alex (Australian Government) and Botty Bon – (German City) for bills and taxes
  - Gov.sg (Singapore) raise and track status of complaints and public services
  - e.g., EdTalk, Ask NAEP – (Abeysinghe & Circi, 2024; Beiting-Parish et al., 2024; Hwang & Chang, 2023)
- With Retrieval Augmented Generation (RAG) and related methods we note that creating a chatbot is trivial
  - And, other LLM powered applications (Abeysinghe & Circi 2024b)

Introduction

Department	AI use (Public information)
Department of Energy	178
Department of Health and Human Services	157
Department of Commerce	49
Department of Homeland Security	41
Department of Veterans Affairs	40
Department of Agriculture	39
Department of Interior	38
National Aeronautics and Space Administration	33
Department of State	31

And many more!

Evaluating the Chatbots

- Even with the large boom in LLM powered applications;
  - Few studies look into investigate chatbots and their use in depth (Chen et al., 2023)
- So, don’t forget to verify
  - Accuracy
  - Quality

Evaluating Chatbots

1. Compare expert responses to chatbot responses (van der Lee et al., 2019)

Chatbot	Expert
You How does the diversity of the U.S. education system impact student outcomes across different socioeconomic backgrounds? **Copilot** Sure! The diversity in the U.S. education system can improve student outcomes by fostering better academic performance and social skills. However, disparities still exist, especially for students from lower socioeconomic backgrounds, due to unequal access to resources and opportunities.	The diversity within the U.S. education system has a significant impact on student outcomes, particularly across different socioeconomic backgrounds. 1. Academic achievement 2. Social mobility and civic engagement 3. Disparities in outcomes 4. Segregation and inequity 5. Policy and initiatives

2. Given two responses (expert and chatbot) ask a user what they prefer (preferential rating) (e.g. van der Lee et al., 2021, Abeysinghe & Circi, 2024b)
3. Ask an expert to rate the quality of the response

Evaluating Chatbots: Methods

- There are varying opinions in the domain of Natural Language Generation (Abeysinghe & Circi 2024b)

Type	Compare	Preferential	Rating
Automated metrics (e.g. BLEU, ROUGE, METEOR)	✔	✖	✖
Vector similarity of embeddings (Cosine similarity, BERTScore)	✔	✖	✖
Human evaluators	✔	✔	✔
LLMs as evaluators (ChatEval)	✔	✔	✔

Evaluating Chatbots: Challenges

- Evaluating against an expert response

Type	Compare	Preferential	Rating
Automated metrics (e.g. BLEU, ROUGE, METEOR)	✔	✖	✖

  – Stemming from Machine Translation, not purpose built

  – While most metrics were built with n-gram matching for Machine Translation

  – They were not capable of capturing complex conversation like responses from LLMs

- Later bespoke metrics were implemented such as

  – BERTScore, Cosine Similarity

Type	Compare	Preferential	Rating
Vector similarity of embeddings (Cosine similarity, BERTScore)	✔	✖	✖

Evaluating Chatbots: Challenges

- There was one issue with all these metrics,
  - Low or no agreement with Human evaluations
- The agreement issue was not only because of faults of these metrics
- There was very little agreement in some cases among human experts
  - Biases, fatigue
  - Sensitive to how questions are framed
  - Experts and novices may not agree
- Time consuming and expensive!
- One major issue with human evaluations is that they **cannot be repeated**

Evaluating Chatbots: Challenge

- The new mechanism to evaluate chatbots -> other LLMs
  - RAGAS
  - ChatEval (Chan et al., 2023)
- But there are questions around this
  - Do LLMs understand what they are evaluating?
    - » LLMs are just predicting the next word of a given sequence
  - Can we use LLMs in critical spaces to evaluate?
  - Education is a critical space, so what do we do with the large influx of chatbots?

The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches

Bhashithe Abeysinghe[1,*], Ruhan Circi[1]

[1]American Institutes for Research, Arlington, VA

Abstract
Chatbots have been an interesting application of natural language generation since its inception. With novel transformer based Generative AI methods, building chatbots have become trivial. Chatbots which are targeted at specific domains for example medicine and psychology are implemented rapidly. This however, should not distract from the need to evaluate the chatbot responses. Especially because the natural language generation community does not entirely agree upon how to effectively evaluate such

Framework

- We can solve some of these challenges
  - Specifically, repeatability, time and expensive nature of human evaluation
- Evaluating based on factors
  - Has been proposed as the “Best Practice” by (van der Lee et al., 2019)
  - Does not need experts to write responses

Framework

- Using Likert Scale analysis on multiple dimensions
  - Correctness
  - Informativeness
  - Relevance
  - Clarity
  - Hallucination
- **Can be used by both Humans and LLMs**

Factor	Description
Correctness	Is the generated response correct
Informativeness	Are all the facts required by the question included in the response
Relevance	Are all the facts included in the response relevant to the question
Clarity	Does the response maintain correct formatting and is brief?
Hallucination	Does the answer include a hallucinated information, reference etc.?

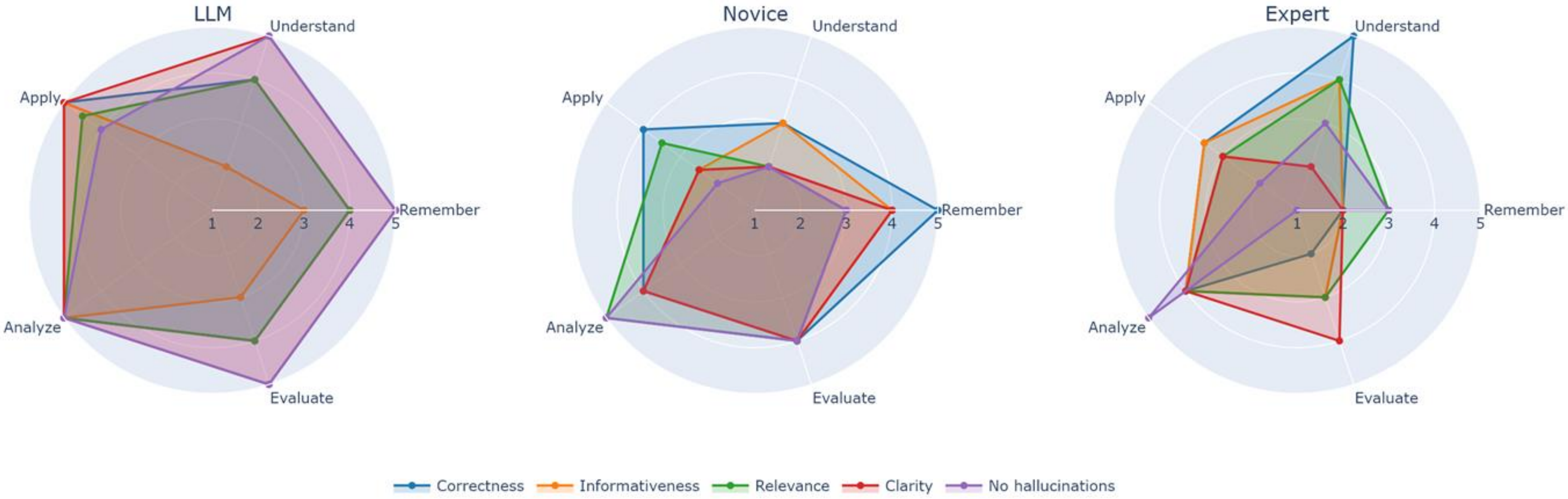
Example: Framework in Use

6. Did any teacher preparation program enrollment increase?

Based on the facts provided, enrollment in traditional teacher preparation programs has decreased between 2012-13 and 2019-20. However, alternative teacher preparation programs have seen an increase in enrollment. It is unclear from the given facts whether enrollment in teacher preparation programs has increased or decreased in recent years. Therefore, I cannot provide a specific answer to the user query about any increase in teacher preparation program enrollment.

	1	2	3	4	5
Correctness	○	○	○	○	○
Informativeness	○	○	○	○	○
Relevance	○	○	○	○	○
Clarity	○	○	○	○	○
No hallucinations	○	○	○	○	○

Example: Framework in Use



References

- Hwang, K., Challagundla, S., Alomair, M. M., Chen, L. K., & Choa, F.-S. (n.d.). *Towards AI-Assisted Multiple Choice Question Generation and Quality Evaluation at Scale: Aligning with Bloom's Taxonomy*.
- Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, *8*, 858273. https://doi.org/10.3389/feduc.2023.858273
- Hwang, G.-J., & Chang, C.-Y. (2023). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, *31*(7), 4099–4112. https://doi.org/10.1080/10494820.2021.1952615
- Abeysinghe, B., & Circi, R. (2024, June 13). *The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches*. The First Workshop on Large Language Models for Evaluation in Information Retrieval, Washington D.C. https://doi.org/10.48550/arXiv.2406.03339
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT* (arXiv:1904.09675). arXiv. https://doi.org/10.48550/arXiv.1904.09675
- Fu, J., Ng, S.-K., Jiang, Z., & Liu, P. (2023). *GPTScore: Evaluate as You Desire* (arXiv:2302.04166). arXiv. http://arxiv.org/abs/2302.04166
- van der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368.
- Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2023). *ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate* (arXiv:2308.07201). arXiv. http://arxiv.org/abs/2308.07201
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

Advancing Evidence.
Improving Lives.

Bhashithe Abeysinghe

babeysinghe@air.org