Let the LLMs Talk: Replication of Conversational QA with LLaMA3 and GPT4-Turbo

Kalyani Vinayagam Sivasundari, Aarifah Ullah, Vaishavi Vijayakandan
Department of Computer Science
George Washington University
kalyani.vinayagamsivasundari, aarifahullah, vaishavi.vijayakandan@gwu.edu

Abstract

This study replicates Abbasiantaeb et al.'s "Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions" by presenting a method for simulating conversational question-answering (CQA) using two instances of a pre-trained large language model (LLaMA3 and GPT4-turbo), acting as a student and a teacher. Utilizing zero-shot learning, this approach aims to replicate human-like conversational dynamics while avoiding the constraints of dataset annotation. The study evaluates how these models autonomously generate and answer queries, focusing on their ability to sustain realistic dialogues. While the results highlight LLaMA3's potential in automated educational platforms and customer service applications, they also underscore the need for additional training to improve consistency and ensure completeness in responses. Compared to LLaMA3, GPT-4 demonstrates more effective performance, but further refinement is needed to enhance accuracy in both models. This research ultimately identifies significant prospects for enhancing conversational AI capabilities and improving the adaptability of large language models.

Keywords: conversational AI, question-answering systems, large language models, zero-shot learning, natural language processing, autonomous interactions.

1 Introduction

In the rapidly evolving field of natural language processing (NLP), the simulation of human-like conversational dynamics using machine learning models represents a significant technological advancement with wide-ranging applications. This paper explores a novel approach to conversational question-answering (CQA) systems through the utilization of two instances of a pre-trained large language model (LLaMA3 and GPT4-turbo), functioning in the roles of a student (questioner) and a teacher (answerer). Traditional CQA systems predominantly rely on labor-intensive human-annotated datasets to train models capable of generating and understanding interactive dialogue. This method, while effective, suffers from drawbacks related to scalability, speed, and consistency.

Advancements in zero-shot learning capabilities of large language models provide a compelling alternative, allowing models to generate and respond to questions based on their extensive pre-training without further specific tuning. This research leverages these capabilities to simulate teacher-student interactions, aiming to replicate and potentially enhance the human-to-human conversational experience. Such simulations not only circumvent the limitations associated with manual data annotation but also potentially improve the quality and depth of conversational interactions.

The primary aim of this study is to evaluate the feasibility and effectiveness of using LLMs in an automated CQA setting to understand their ability to conduct meaningful dialogues. This includes assessing their capacity for nuanced understanding and response generation that closely mimics human conversational patterns. Moreover, this research attempts to bridge the gap between theoretical NLP models and their practical applications, demonstrating the real-world utility of LLMs.

Emphasizing the importance of zero-shot learning, this study highlights the adaptability of pre-trained models to function in new contexts without the need for extensive retraining. This not only showcases the flexibility of LLMs but also marks a significant step towards the development of more dynamic and responsive AI systems. By examining the interactions between the student and teacher models, the paper also delves into the intrinsic capabilities of LLMs to generate knowledge-driven responses, thereby providing insights into their reasoning abilities and the challenges of ensuring response accuracy and relevance.

This paper begins with the Models and Methods section, outlining the methodology and technical setup for LLM-to-LLM interaction, setting a clear framework for their application. It then transitions into the Experiments section, detailing the practical execution of these methods, the simulations run, and the evaluation of generated data. Through a detailed analysis of the results obtained from these simulations, this study contributes to the broader discourse on the applicability of LLMs in CQA, highlighting both the strengths and limitations of this approach.

2 Models and Methods

This research employs a dual-instance setup of pre-trained large language models (LLaMA3 and GPT4-turbo), utilizing the advanced capabilities of the `replicate` and `openai` libraries to facilitate a teacher-student simulation for conver-

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

sational question-answering (CQA). The integration of these tools with the Natural Language Toolkit (`nltk`) and custom Python scripting forms the backbone of our methodology, aimed at emulating real-world conversational dynamics in an educational context. Below is a detailed breakdown of the models and methodologies applied:

Advanced NLP Tools Configuration

- **LLaMA3 Configuration**: Leveraging the `replicate` library, LLaMA3 is accessed and utilized in its pre-trained state, capable of understanding and generating responses across a broad spectrum of topics. This zero-shot learning approach enables the model to generate contextually relevant responses based solely on the input prompt without additional training.
- **GPT4-turbo Configuration**: The `run_gpt4` function uses the `openai` library to access GPT-4's chat completion capabilities. The model is configured through an API key, and responses are generated based on the conversation log provided. It returns both the role and content of GPT-4's response.
- **Python Environment Setup**: Critical Python libraries such as `nltk` for text processing and `re` for text string manipulation are installed and configured. The setup also includes secure handling of API tokens through environment variables, ensuring secure and reliable access to LLaMA3 and GPT4-turbo.

Data Preparation and Management

- **Text Input Handling**: The `get_topics` function is designed to read and parse structured articles from a specified file. This data is meticulously split and cleaned to ensure each article is clearly defined for the simulation.
- **Text Preprocessing**: Utilizing `nltk`, the text is tokenized, cleaned, and standardized, removing extraneous elements like punctuation and multiple spaces to prepare clean, usable text inputs for the simulation.

Robust Simulation Framework

- **Role Definitions and Dynamics**:
  - **Student Model (Questioner)**: The model is given `student_instruction` that directs it to be "a curious student who wants to explore" a topic in addition with prompts at spaced intervals, the model dynamically generates questions intended to probe deeper into the text, reflecting a natural curiosity and the evolving context of the document.
  - **Teacher Model (Answerer)**: This model is also given instruction, `instruction_teacher` and occasional prompts, this model instance operates strictly within the confines of the provided text, selecting responses that are accurate and textually grounded, using advanced text matching and selection algorithms.
- **Interactive Dialogue Simulation**: The simulation is structured to facilitate a continuous, coherent dialogue where the student model poses questions influenced by ongoing responses, promoting a realistic conversational flow.

Advanced Functionality and Validation

- **Text Processing and Validation Functions**:
  - `pre_process_res` and `get_grounded_answers` ensure that the responses are clean and directly derived from the text.
  - `check_overlap` verifies that the *teacher* generated its answer solely from copying the "Section" text. In this way, evaluation based on correctness was simpler and avoided potential cases of hallucination.
- **Error Handling and Adaptive Feedback**: Incorporates mechanisms and directional prompts to adjust questions and responses based on the feedback loops within the dialogue, enhancing the realistic and educational value of the simulation and ensures the models remain within the defined parameters.

Experimental Control Measures

- **Iteration and Patience Parameters**: The simulation allows four attempts to refine answers, defined by `TEACHER_PATIENCE`, enhancing the reliability of the conversational outputs. This control ensures that the dialogue maintains high standards of relevance and that the model stayed within the defined constraints.

This methodology harnesses state-of-the-art NLP tools within a robust simulation framework, enabling detailed exploration of conversational dynamics in educational settings. By employing LLaMA3 through `replicate` and GPT4 through `openai`, the study not only tests the models' ability to engage in meaningful dialogues but also their potential to adapt to complex conversational scenarios without prior specific training.

3 Experiments

We decided to replicate a subset of the experiments in Abbasiantaeb et al. (2024), specifically just the interaction between the two models: *student* and *teacher*. For example, additional experiments evaluated the *teacher* using the human annotated dataset QuAc!. A crowd-sourced task asked the workers to compare the *teacher* generated answers with the human-annotated ones from QuAc! with the three metrics: *naturalness*, *completeness*, and *correctness*, and finally asked for their overall preference.

The simulation was run twice with two different LLMs: LLaMA3 and GPT-4-turbo to assess whether our findings resemble that of the paper's. The setup, execution, and evaluation methodologies are described below, focusing specifically on the practical aspects of the simulations conducted.

Experimental Design

Much of the experimental design followed a set up that emulates the paper. A series of Wikipedia articles, recognized by its title is given to both systems. Each article contains the first paragraph of the article known as the "Background", a section "Header" that eludes to a specific topic within the general article, and the "Section" that serves as additional information with more details over the "Background."

The *student* has limited access to information and can only see the "Topic," "Background," and "Header" portions and must generate relevant questions to the "Header." The *teacher* on the other hand has access to all parts of the article, including the "Section" so it may provide detailed answers to the questions asked by the other system.

- **Simulation Runs:** Each simulation involved a sequence of conversational exchanges centered around a specific topic extracted from a subset of the SimQuAC dataset, the LLM-generated dataset used in Abbasiantaeb et al (2024).
- **Data Segmentation:** The text for each topic was divided into several components, including the title, background information, headers, and detailed sections. These segments served as the foundation for the questions and responses generated by the student and teacher models, respectively.

Data Collection and Setup

- **Source Material:** A curated selection of documents from the SimQuac dataset, mostly documents focusing on famous figures and cultural references were used to prepare the topics. A simulation file *data_for_simulation* contains the documents separated by the topics.

Execution of Simulations

- **Conversation Initiation:** The student model was given a prompt to ask a question which initiated each simulation. This question aimed to elicit an informative response, setting the stage for a natural conversational flow.
- **Dynamic Interaction:** Subsequent exchanges were driven by the student model's questions, which were influenced by the teacher's previous responses. This iterative questioning aimed to simulate a natural and coherent dialogue. The conversation ended after 12 question-answer sets were generated.

Evaluation Metrics

The *student* model will be judged in a slightly more abstract criteria, keeping in mind that a question in general is not inherently wrong. Instead, generated questions will be observed for their patterns. The *teacher* will be evaluated using stricter criteria because in real-world CQA systems, the answers come from the system and should be accurate and intuitive to use for individuals. Therefore the following metrics will be used to evaluate the *teacher*:

- **Naturalness:** The responses from the *teacher* model were evaluated for their control of the language, its fluency and flow. The model should avoid awkward phrasing and aim to mimic a person answering the question.
- **Correctness:** The *teacher* should also provide factual information, and inaccuracies must be avoided considering the necessity that its answers may affect individuals who use CQA systems in practical and vital settings.
- **Completeness** Not only does the answer need to be correct, it should encompass more than just a basic response to the question. For example, if a question asked for the address of a restaurant, simply stating the state or the zip code would be considered incomplete, but including the street is a more comprehensive answer.

4 Results and Discussion

We find that a portion of our results aligned with the findings of Abbasiantaeb et al (2024). The paper was published for the ACM conference this March, and the latest LLaMA model, LLaMA3 was released on April. When the paper was written, they found that LLaMA was unsuitable for generating answers to posed questions. It was not able to copy spans of text in a zero-shot manner. Our experiments aimed to check whether this still holds for the latest LLaMA model.

LLaMA3 was able to copy parts of the text in a zero-shot manner, but not consistently. Often-times, the model outputted "I cannot find the answer" even though the Section contained it. Occasionally, the model outputted this phrase, but then proceeded to answer the question. Of the questions that it did answer, the model was not able to gauge what constituted as the entire complete answer or gave unsatisfactory responses that contributed to further understanding, lacking in the evaluation metric of *completeness*.

Examples of LLaMA3's answers to the Student's questions

Question: What university did Bob Hayes attend for his college career?
Answer: Florida A&M University a historically black college

Question: What sport(s) did Bob Hayes excel in during his college career at Florida A&M University?
Answer: track & field

Question: I'll ask a new question. Did Bob Hayes win any championships or awards during his college career at Florida A&M University?
Answer: AAU 100 yard dash champion three years running, from 1962-1964, and in 1964 was the NCAA champion in the 200 meter dash

We then repeated the simulation using GPT4-turbo with the same set of documents listed under *data_for_simulation*. GPT4-turbo was not only able to copy the text, but was able to combine different spans to produce natural-forming sentences and narration that went above repeating the Section text verbatim. This aligned with the replicated paper's findings. There were some instances of incomplete-phrasing and the *teacher* outputted "I cannot find the answer" rather frequently. 45% responses were "I cannot find the answer" compared to 57% from LLaMA3. Interestingly, for both of the *teacher*s, they could not find answers more often near the end of the simulation, as if they were tired of answering questions.

Examples of GPT4-turbo's answers to the Student's questions

Question: How did the initiatives like SF HOT and Project Homeless Connect impact the homeless situation in San Francisco?
Answer: 5,000 more homeless people were given permanent shelter in the city

Question: Did Gavin Newsom face any significant challenges or controversies during his mayoralty?
Answer: in 2009, Newsom came under attack for the City of San Francisco's policy of illegally harboring juvenile criminal aliens

Question: What broad strategies did Gavin Newsom employ to address issues in San Francisco during his time as mayor?
Answer: focused on development projects in Hunters Point and Treasure Island

Overall, evaluating the *teacher* on the three metrics: *naturalness*, *completeness*, and *correctness*, GPT4-turbo generated more natural, more complete, and most importantly more correct answers over LLaMA3. However, we think in any case zero-shot learning may be inadequate for the *teacher* side since in deploy-able CQA systems, the expectation is for more natural conversation and explanation. Zero-shot may work, but perhaps future work could consider exploring how the model does without the constraint of copying the text verbatim as this seems to restrict the system from utilising its own knowledge. In general, LLMs provide great explanations if they are not restricted to following certain texts. Different metrics to evaluate correctness can be used, such as a human annotator which can also check for instances of hallucination.

As for evaluating the *student*, what we noticed is that the LLaMA3 *student* and the GPT4-turbo *student* asked entirely different kinds of questions, but both made quality-sounding and relevant questions. They focused on finding details from different information, which is note-worthy considering that both systems were given the exact same prompts. LLaMA3 tended to ask questions that could be found in the background primarily, while the GPT4-turbo model asked questions specifically keeping the subsection's title, the Header, in mind, so the responded answers came from the unseen Section. GPT4-turbo's student most-aligned with what the paper aimed to accomplish: to learn about an unknown topic, given some general information.

Examples of LLaMA3 and GPT4-turbo Student questions over the same topic, "The Residents" with the section from "2016-Present: The Real Residents"

LLaMA3

Q1: What does "The Real Residents" refer to in the context of the Residents' career?
Q2: What caught my attention in this document is that The Residents have released over sixty albums. Are they all avant-garde music?
Q3: Who are the members of The Residents, are they anonymous?
Q4: What is the main theme or focus of the multimedia works created by The Residents?

GPT4-turbo

Q1: What is the theme or concept behind "The Ghost of Hope" album released by The Residents in 2017?
Q2: What details can you provide about the themes or inspirations behind the upcoming blues-styled album mentioned during the "In Between Dreams" tour?
Q3: How was the reception of "The Ghost of Hope" album among fans and critics after its release?
Q4: What visual themes or styles were used in the video for "Rushing Like a Banshee" directed by John Sanborn?

5 Conclusion

This study explored a teacher-student simulation framework utilizing two pre-trained large language models, LLaMA3 and GPT4, to simulate conversational question-answering (CQA). The simulation replicated Abbasiantaeb et al (2024) using the SimQuAC dataset to evaluate how well the models autonomously generate and answer queries. Both models demonstrated the ability to engage in meaningful dialogues with relevant and informative responses. However, challenges like inconsistencies in response quality and limitations in zero-shot learning were evident. While GPT4 outperformed LLaMA3 in adaptability and coherence, both models need further training to enhance response accuracy and completeness, particularly in educational and customer service contexts.

6 Future Work

- **Training on Diverse Datasets**: Fine-tune both models using a broader range of annotated datasets to improve consistency and completeness, ensuring they can better handle variations in real-world conversational scenarios.
- **Dynamic Feedback Systems**: Introduce adaptive feedback loops where the teacher and student models learn iteratively, refining their questions and answers based on past interactions.
- **Bias Mitigation Strategies**: Explore methods to reduce inherent biases in training data, improving fairness, accuracy, and adherence to ethical standards.
- **Advanced Evaluation Metrics**: Develop more sophisticated evaluation metrics to better capture conversational coherence, topical relevance, and user satisfaction.

7 Ethical Statement

This study employed large language models (LLaMA3 and GPT4) for simulating conversational question-answering. While these models show promise, it's crucial to address ethical concerns. These models may reflect biases present in training data. We stress the importance of bias mitigation techniques and diverse datasets to ensure fair and non-discriminatory responses. Understanding how models generate responses is essential. Our work supports methods that increase transparency and help users understand AI-generated answers. Protecting user privacy is crucial. We handled simulated data responsibly and ensured no personally identifiable information was used.

References

Zahra Abbasiantaeb, Yifei Yuan, E. K., and Aliannejadi, M. 2024. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. *arXiv:2312.02913v1*.