GPTuring Test

Avi Rahimov & Elon Ezra

January 2024

> *"A computer would deserve to be called intelligent if it could deceive a human into believing that it was human."*
>
> Alan Turing, 1950

Abstract

This project involves developing a web-based platform for a Turing-Test experiment where users engage in bot-human chat interactions. Upon entering the website, users input their names and wait in a virtual lobby for another participant. Once paired, one user becomes a tester, and the other an experimenter.

The tester engages in a five-minute chat with two candidates: one is a human (the experimenter), and the other is a bot (powered by the GPT and Llama models). After the chat, the tester provides feedback, identifying which candidate was human and which was the bot, along with reasons for their choices. The experimenter's role is to prove they are human.

This study explores the capabilities of contemporary AI models in mimicking human conversation, providing insights into the effectiveness of models like GPT-3.5, GPT-4o, and Llama 3 in passing a Turing-like test.

1 Introduction

The Turing Test, conceptualized by Alan Turing in 1950, is a cornerstone in the field of artificial intelligence (AI), assessing a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. Turing proposed an "imitation game," where a human evaluator converses with both a machine and another human, without knowing which is which. The machine passes the test if the evaluator cannot reliably distinguish it from the human participant. This test has been instrumental in shaping research in AI and cognitive science, particularly in the development of conversational agents and natural language processing (NLP) technologies.

In this project, we aim to create a practical implementation of the Turing Test through a web-based platform designed for bot-human chat interactions. Our system enables users to engage in conversations where one participant is a human, and the other is an AI bot, powered by models such as GPT-3.5, GPT-4o, and Llama 3. The primary objective is to evaluate the ability of these AI models to convincingly mimic human conversation and to collect user feedback on their experiences and perceptions.

Upon accessing the website, users are prompted to enter their names and wait in a virtual lobby until another participant joins. Once matched, one user is assigned the role of the tester, and the other becomes the experimenter. The tester engages in separate five-minute text-based conversations with both the human experimenter and the AI bot. Following these interactions, the tester is directed to a feedback page to provide their insights on which conversation partner they believe was human and which was a bot, along with their reasoning.

The experimenter, on the other hand, has the sole task of proving their human identity during the conversation. This dynamic introduces a real-time challenge for the AI bot to convincingly emulate a human participant, leveraging advanced NLP techniques and contextual understanding.

The significance of this project lies in its potential to contribute valuable data and insights into the current state of AI conversational abilities. By analyzing user feedback and interaction patterns, we can assess the effectiveness of models like GPT-3.5, GPT-4o, and Llama 3 in passing a Turing-like test and identify areas for improvement. Additionally, this project provides a robust framework for future studies on human-AI interaction, helping to refine AI models and enhance their performance in real-world applications.

The development of this platform involves several key components, including user interface design, backend server management, and integration of AI models for real-time conversation. Ensuring a seamless and secure user experience is paramount, necessitating meticulous attention to data privacy and system reliability.

In conclusion, this project not only aims to recreate the Turing Test in a modern digital context but also aspires to push the boundaries of AI's conversational capabilities. By engaging users in structured yet open-ended interactions, we seek to gather empirical evidence on the indistinguishability of AI-generated text from human communication, advancing our understanding of AI's potential and limitations.

2 Related Work

The concept of the Turing Test, as originally proposed by Alan Turing in his seminal paper [2], has inspired numerous studies and variations. Turing's test, often referred to as the "Imitation Game," involves an interrogator who must determine which of two participants, one human and the other a machine, is the human based on their responses to questions. This test aims to assess a machine's ability to exhibit intelligent behavior indistinguishable from that of a human.

2.1 Gamified Approach to the Turing Test

One notable variation is the gamified version of the Turing Test, as explored in [1]. In this approach, participants engage in a single-window chat interface. After a set period, they must decide whether they are conversing with a human or a bot. This method differs significantly from Turing's original design, as it lacks the separation of the tester's room into two distinct spaces for the human and the bot. Additionally, the roles of tester and experimenter are not explicitly defined in the gamified version.

2.2 Our Project: A Web-Based Implementation

Our project seeks to adhere more closely to Turing's original vision [2]. We developed a web-based platform where participants, referred to as the tester and the experimenter, enter a virtual room. The tester's room is split into two separate spaces: one for the experimenter and one for the bot. The tester, unaware of which room contains the human, interacts with both. Conversely, the experimenter is placed in a single room and must convince the tester of their humanity. This setup provides a more accurate representation of Turing's test by maintaining the distinct interactions between tester and the two entities being evaluated.

2.3 Comparison with Gamified Version

The primary distinction between our project and the gamified version [1] lies in the structural design and role definitions. While the gamified version employs a single chat interface and lacks explicit roles, our implementation ensures a clear separation of spaces and roles, aligning with Turing's original methodology. This separation is crucial for maintaining the integrity of the test, as it prevents any bias or confusion that might arise from a unified interface.

Furthermore, our project emphasizes the interaction dynamics described by Turing [2], providing a robust framework for evaluating machine intelligence. The experimenter's role is critical, as they must actively engage in convincing the tester of their humanity, thereby enhancing the depth and rigor of the evaluation process.

3 Basic Prompt build

The field of prompt engineering is complex and dynamic, focusing on defining instructions for AI models to shape their behavior according to specific goals. This process requires a delicate balance between providing precise and content-rich instructions that stimulate the model to act in a certain way, and avoiding information overload, which could cause the model to forget or ignore essential parts of the instructions.

In the current prompt, we began with the phrase "*Enter RP mode,*" a common opening in role-playing environments that forces the model to assume a specific character, which we define later. This definition is crucial as it creates a coherent framework within which the model needs to operate. Additionally, using terms familiar to the model from the role-playing domain enhances its suitability for the specific context, leading to interactions that feel more natural.

Next, we explicitly defined the character with the instruction *"You are now a {gender} named {bot_name}, and the person talking to you is named {human_name},"* which assigns the model a clear role with a defined name and gender. This allows the model to adopt a more personalized approach to the conversation, increasing the interlocutor's feeling that they are speaking with a real person. In previous interactions with models, we've seen how defining a clear character enables the model to generate more personal and meaningful responses, contributing to a deeper and more engaging conversation.

We then closed the technological context in which the model operates by providing instructions that position it in a virtual chat environment where it cannot see or hear the user. These guidelines not only limit the model's response mode but also give it an opportunity to demonstrate personalization by referencing the limitations of the conversation. Through this, the model reflects the environment in which it operates, enhancing the user's sense of reality and helping the model appear more convincing.

Finally, we emphasized the writing style that the model should use. We asked it to adopt a natural and informal speech style, including the use of slang, spelling mistakes, and emojis. These choices aim to break the rigid and organized structure of language typically associated with artificial intelligence, thereby bringing the conversation as close as possible to a real human conversation. The challenges of prompt engineering in this area include the need to create prompts that allow flexibility and creativity in the model's responses while maintaining a defined behavioral framework that aligns with the conversation's goals.

By combining all these elements, we succeed in building a conversation model that merges a well-defined framework with a human-like interaction capability, addressing the unique challenges of the prompt engineering field.

4 Experiment Results

In this section, we present the results of the Turing Test experiments conducted using different AI models. The objective was to assess the effectiveness of these models in fooling human participants into believing that they were conversing with a human rather than an AI.

4.1 GPT 3.5 Turbo No Prompt

The experiment using GPT 3.5 Turbo without any specific prompt engineering resulted in only 1 out of 22 participants being fooled. This corresponds to a very low success rate of 4.55%, indicating that without tailored prompts, this model struggles to convincingly mimic human conversation.

4.2 GPT 3.5 Basic Prompt

With basic prompt engineering applied to GPT 3.5, the model's performance improved significantly. Out of 24 participants, 5 were fooled, resulting in a success rate of 20.83%. This suggests that even simple prompt modifications can enhance the model's ability to deceive human evaluators.

4.3 GPT 4o Basic Prompt

When applying basic prompts to the more advanced GPT 4o model, the number of fooled participants increased to 6 out of 25, yielding a success rate of 24.00%. This indicates that GPT 4o, with prompt engineering, has a slightly better performance than GPT 3.5, but the improvement is marginal.

4.4 Llama 3

The Llama 3 model demonstrated the most impressive results. With 12 out of 17 participants being fooled, the success rate reached 70.59%. This substantial increase highlights the advanced capabilities of Llama 3 in simulating human-like conversations, particularly when compared to the other models tested. we can assume that

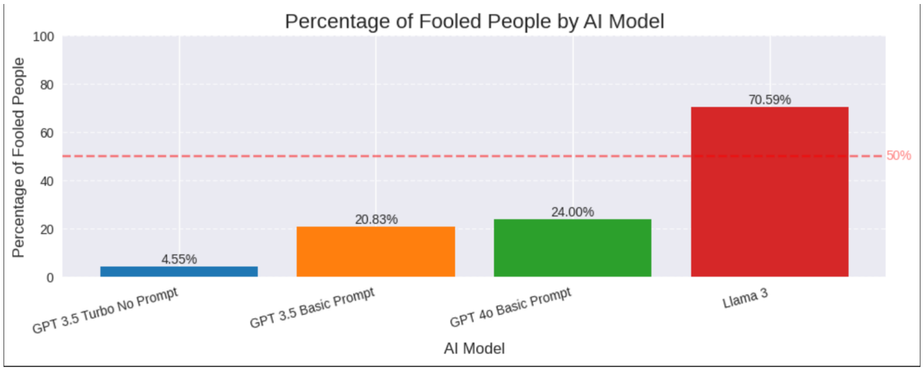

Figure 1: Percentage of Fooled People by AI Model

Figure 1 illustrates the percentage of fooled participants across the different AI models tested. The red dashed line represents the 50% threshold, above which a model can be considered more likely to deceive participants than not.

LLama 3 appears to achieve superior performance in mimicking human behavior and fulfilling conversational roles, largely due to extensive training and fine-tuning on dialogue-based data. Meta's access to data, such as conversations within groups and posts, enables the model to capture linguistic nuances that are often unattainable through training on traditional text sources like articles. This conversation-focused training also enhances the model's ability to generate culturally and contextually relevant responses, aligning closely with the goals of conversational AI assessments.

Moreover, advancements in the model's architecture and training methodologies further enhance data utilization. The refined feature extraction processes lead to a more accurate and contextually appropriate understanding of conversational dynamics, thereby improving the model's overall effectiveness in dialogue settings.

5 Conclusion

This study aimed to evaluate the effectiveness of various AI models in passing a Turing-like test, assessing their ability to mimic human conversation convincingly. The experiments demonstrated that prompt engineering can significantly influence the performance of AI models, as observed in the improved results of GPT 3.5 and GPT 4o with basic prompts. However, the most notable outcome was achieved with the Llama 3 model, which successfully deceived 70.59% of participants, surpassing the 50% threshold that typically signifies successful human mimicry.

These findings underscore the potential of advanced AI models like Llama 3 in approaching human-like conversational abilities. The stark contrast between the results of Llama 3 and those of GPT 3.5 Turbo without prompt engineering highlights the critical role of both model architecture and fine-tuning techniques in enhancing AI performance.

Looking ahead, the exploration of more sophisticated methods, such as soft prompt tuning, holds promise for further improving AI models' capacity to engage in indistinguishable human-like conversations. The insights gained from this research provide a strong foundation for future work aimed at refining AI models and advancing our understanding of human-AI interaction.

6 Future Work

The success of the Llama 3 model in this study opens up new avenues for research in the field of AI-driven conversation. Future work will focus on exploring advanced prompt tuning techniques, such as soft prompting, to further enhance the conversational abilities of AI models. Soft prompting involves the optimization of prompt embeddings rather than the raw text, offering a more nuanced approach to guiding AI responses. This technique has the potential to significantly improve the coherence and context-awareness of AI-generated text, making it even more challenging for human evaluators to distinguish between human and machine interlocutors.

Additionally, expanding the scope of the experiment to include a wider range of AI models and more diverse participant demographics will provide a more comprehensive understanding of the factors that influence the effectiveness of AI in passing the Turing Test. This future research will not only contribute to the development of more sophisticated AI models but also to the broader field of human-AI interaction, shedding light on the evolving relationship between technology and human communication.

References

[1] A game approach to the turing test. *arXiv*, 2023. https://arxiv.org/pdf/2305.20010.

[2] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, October 1950. A Quarterly Review of Psychology and Philosophy.