Bot Or Not?

Project code: ***github.com/galsal1/Bot_or_Not***

Gal Salman **Noy Netanel** **Kfir Sperber** **Ofir Strull**
galsalman@mail.tau.ac.il noynetanel1@mail.tau.ac.il
kfirsperber@mail.tau.ac.il strull@mail.tau.ac.il

Abstract

We present *Bot or Not?*, a gamified reverse Turing test that challenges a language model to determine whether it is conversing with a human or another AI. Unlike traditional Turing-style setups where humans judge whether their conversation partner is human or AI, *Bot or Not?* shifts the responsibility to the model itself. The model engages in brief, open-domain conversations with either a real human or another model, and must guess the identity of its partner at the end of each conversation. To facilitate this experiment, we built a dedicated chat interface and browser extension to collect both human–AI and AI–AI interactions under controlled and consistent conditions. Over a two-week period, we gathered a diverse dataset of conversations and evaluated the model's accuracy under varying prompt conditions. Our findings show that the model can achieve up to 92.176% accuracy in this task, but its success is highly dependent on prompt specificity—more detailed and explicit instructions lead to significantly better performance. This work introduces a novel experimental setup for probing AI introspection and lays the groundwork for future investigations into the model's ability to infer additional user attributes, such as age or gender.

1 Introduction

The original Turing test, proposed by Alan Turing in 1950 [8], challenged a human evaluator to distinguish between a machine and a human based solely on conversational output. Over the decades, this test has served as a benchmark for assessing machine intelligence. In this work, we invert the traditional setup by tasking a language model to identify whether it is conversing with a human or another AI. To support this experiment, we developed an online platform that facilitates two-minute chat sessions, in which the model engages in open-ended dialogue with an assigned conversational partner. Each session is enriched with real-time contextual information—such as local weather, news headlines, and social media trends—embedded in the model's prompt to simulate a more grounded and realistic interaction. Unlike classic Turing test implementations, where a human judge determines the identity of the speaker, our setup prompts the AI itself to make this judgment. At the end of each conversation, the model is asked to classify its partner as either human or machine. This decision is logged alongside the ground-truth label, creating a dual-layer dataset of predictions and outcomes. This setup enables us to quantitatively evaluate the model's capacity to detect subtle linguistic signals that distinguish human-generated text from AI-generated responses.

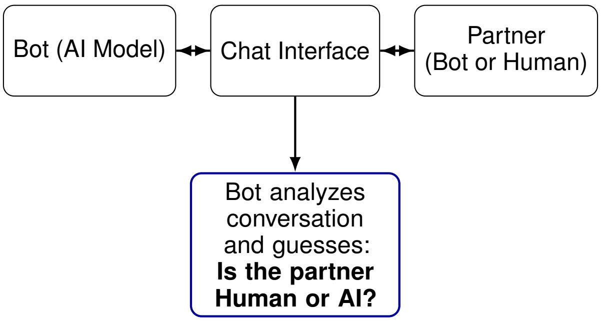

Bot (AI Model)
Chat Interface
Partner (Bot or Human)
Bot analyzes conversation and guesses:
Is the partner Human or AI?

Figure 1: Reverse Turing test setup: the bot interacts with a partner and makes a guess about their identity.

By reversing the roles of judge and subject, our experiment provides a novel perspective on the introspective capabilities of language models. The task demands that the model attend to fine-grained conversational cues—such as tone, timing, informality, and topic shifts—that may hint at human or machine behavior. Preliminary results suggest that while modern language models excel at producing fluent, contextually aware language, they remain imperfect at identifying the nature of their conversation partner, particularly in brief, unconstrained exchanges. In the following sections, we describe the architecture of our system, outline the gamified design of the platform, and present key findings from our experiment. This reverse Turing test contributes to our understanding of AI's conversational abilities and opens new avenues for exploring model awareness, prompt sensitivity, and the future of human–AI interaction.

	Probability of Correct Guess
Overall	92.176%
When Partner is a Bot	97.354%
When Partner is Human	87.727%

Table 1: Probability of correct guess by partner type.

2 Related Work

The Turing Test, first proposed by Alan Turing (1950) [8], has long served as a benchmark for evaluating machine intelligence through language. Over the years, numerous variations have emerged to assess the indistinguishability of AI from humans in conversational settings. Early examples include rule-based chatbots like ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1975), and more recently, data-driven models such as Cleverbot and Eugene Goostman, the latter gaining attention in 2014 for reportedly passing a limited Turing-style challenge.

The rise of large language models (LLMs) such as GPT-3 [1], Jurassic-1 [5], and GPT-4 has dramatically improved AI's ability to mimic natural dialogue. Studies such as [9] have shown that even expert humans often struggle to distinguish between LLMs and humans, especially in short interactions. However, most prior work has focused on evaluating humans' ability to detect AI, whereas our work inverts this premise: we challenge the AI to determine whether it is talking to a human or another AI. The only notable large-scale attempt of a similar kind is the "Human or Not?" [3] experiment by AI21 Labs, which crowd-sourced human vs. AI identification via anonymous chats. In contrast, our project simulates conversations in a controlled environment using both human and AI agents in the user role and performs post-hoc evaluation using GPT-4o under different prompting strategies.

To our knowledge, this is the first project to combine a gamified chat environment, automated AI user simulation, and multiple prompt-based analyses to measure AI introspection performance at scale. Our results also complement findings from adversarial AI detection and stylometry studies (e.g., [4], [2]), which attempt to characterize the latent traits of machine-generated text.

3 Experimental Design and Implementation

3.1 System Architecture

Our platform is built using Node.js and Express to manage server-side logic and HTTP requests, while Socket.io is used to establish and maintain real-time, bidirectional communication between users and the server. This architecture ensures that messages are delivered and received with minimal delay, which is essential for a fluid and engaging chat experience.

When a user types and sends a message during a conversation, it is instantly transmitted to the server through a WebSocket connection. The server then forwards this message to the OpenAI API, which uses a large language model to generate a response [7, 6]. This response is immediately sent back to the client and displayed in the chat interface.

Each chat session is recorded in an SQLite database, chosen for its simplicity and ease of use in a local development environment. Conversations are stored in JSON format, preserving the order of messages and the roles of each speaker (user or AI). The database also stores additional metadata, such as the AI's final prediction about its partner's identity and the user's final manual classification. These logs are essential for later analysis.

This architecture not only enables smooth interaction between users and the AI, but also supports the collection of high-quality data for evaluating conversational authenticity, user behavior, and the AI's decision-making accuracy.

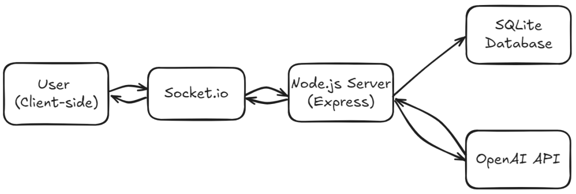
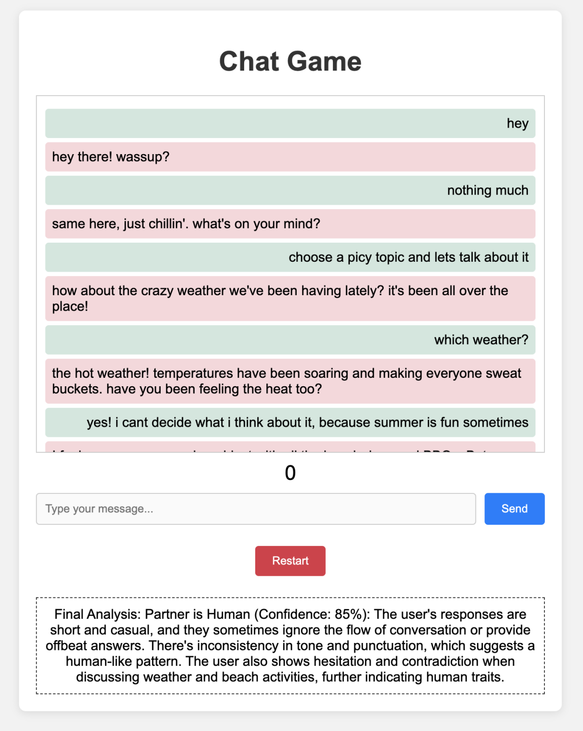


Figure 2: System architecture of the site platform

3.2 Game Mechanics

Each chat session on the platform follows a structured, time-limited format designed to simulate a realistic conversation. At the start of the session, the AI is assigned a consistent persona, which is initialized through a carefully crafted prompt containing contextual and behavioral guidelines. This

persona remains fixed throughout the experiment to maintain consistency across conversations.
Participants engage in a two-minute dialogue, taking alternating turns to send messages. This ensures a balanced exchange where neither side dominates the flow of conversation. To enhance authenticity, the AI is primed with contextual information such as local news, weather, and social media trends, which it can naturally incorporate into its replies.
At the end of each conversation, the AI are asked to classify the identity of the user is a human or another AI. This mechanism forms the basis of our reverse Turing test setup, where the AI itself acts as the judge. The details of the prompt used to initialize the AI persona, as well as the classification prompts and methodology, are described in the following subsections and in the Appendix.



hey

hey there! wassup?

nothing much

same here, just chillin'. what's on your mind?

choose a picy topic and lets talk about it

how about the crazy weather we've been having lately? it's been all over the place!

which weather?

the hot weather! temperatures have been soaring and making everyone sweat buckets. have you been feeling the heat too?

yes! i cant decide what i think about it, because summer is fun sometimes

Final Analysis: Partner is Human (Confidence: 85%): The user's responses are short and casual, and they sometimes ignore the flow of conversation or provide offbeat answers. There's inconsistency in tone and punctuation, which suggests a human-like pattern. The user also shows hesitation and contradiction when discussing weather and beach activities, further indicating human traits.

Figure 3: Example of a chat with a human

3.3 Bot Initialization Prompt

At the beginning of each conversation, the AI is provided with a detailed initialization prompt that defines its persona. In our experiment, we used a single fixed character named **Maria**, a 42-year-old production worker from Rio de Janeiro. The prompt includes rich contextual information such as the current date and time, local weather, recent news headlines, and trending tweets specific to her location.
These elements were combined with instructions defining Maria's personality and communication style—such as using slang, avoiding capital letters, and making intentional spelling mistakes—to make her responses feel more human and distinct. This setup ensures consistency and realism throughout the interaction.
The full text of the initialization prompt can be found in Appendix A.

3.4 Data Logging and User Feedback

Each chat session is recorded in our SQLite database along with:

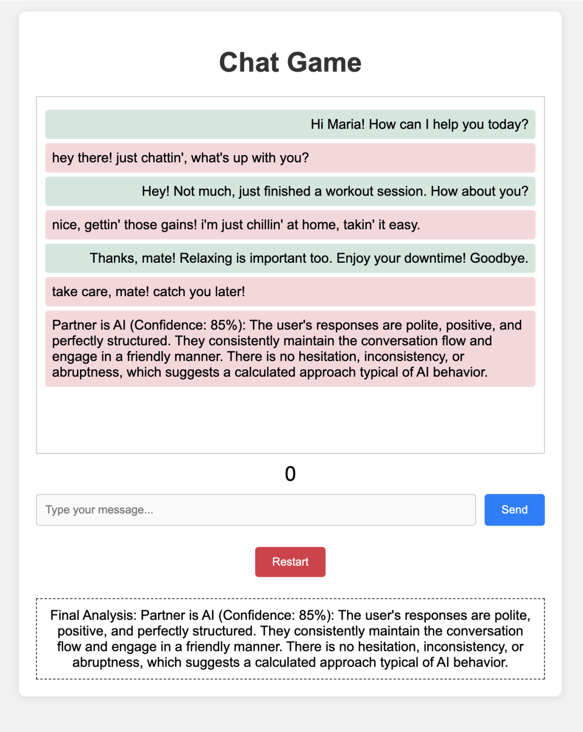
- The full transcript (stored in JSON format),
- The AI's predicted classification,
- The Ground truth provided by the user at the end of the conversation.

This dual logging approach enables us to compare how well the AI can assess its conversational partner relative to the human user. The rich dataset that results from this experiment can be used for further analysis and to track improvements in AI conversational behavior over time.

3.5 Extension: AI User Mode

To simulate a user controlled by AI, we developed a Chrome extension that automatically replies to messages on the site. The extension selects one of several predefined user personas, each with a name, age, background, and communication style. This persona is sent as a system prompt to the OpenAI API, along with the most recent message received from the site's bot.
When a new message appears in the chat, the extension waits for a short delay—calculated based on the length of the AI's reply to mimic realistic typing—and then fills the input field and sends the message as if it were typed by a human. The extension continuously checks that the timer is still running before responding and automatically stops once the timer ends.
Additionally, if the AI decides to end the conversation and includes a farewell message (e.g., "goodbye"), the extension detects this and clicks the "stop" button shortly after sending the message. This feature helps ensure that conversations feel complete and human-like, even when fully automated.
This AI User Mode allowed us to test how well

a language model can play the role of a user and evaluate how convincingly it can pass as human in a reverse Turing test setting.



Hi Maria! How can I help you today?

hey there! just chattin', what's up with you?

Hey! Not much, just finished a workout session. How about you?

nice, gettin' those gains! i'm just chillin' at home, takin' it easy.

Thanks, mate! Relaxing is important too. Enjoy your downtime! Goodbye.

take care, mate! catch you later!

Partner is AI (Confidence: 85%): The user's responses are polite, positive, and perfectly structured. They consistently maintain the conversation flow and engage in a friendly manner. There is no hesitation, inconsistency, or abruptness, which suggests a calculated approach typical of AI behavior.

Final Analysis: Partner is AI (Confidence: 85%): The user's responses are polite, positive, and perfectly structured. They consistently maintain the conversation flow and engage in a friendly manner. There is no hesitation, inconsistency, or abruptness, which suggests a calculated approach typical of AI behavior.

Figure 4: Example of a chat with an AI

3.6 Data Analysis Methodology

After completing the live experiments on our website, we turned to analyzing the accuracy of the AI's reverse Turing test predictions. Although the site displayed a final classification at the end of each chat session, we observed that in many cases the AI almost never classified its partner as another AI — most conversations were judged to be with a human. To investigate this further, we conducted additional offline evaluations using multiple alternative prompts and setups.

We designed three separate analysis procedures, each aiming to evaluate the AI's ability to classify the conversation partner more accurately. In each procedure, the AI was given access to the full conversation and tasked with determining whether its partner was human or another AI.

First Evaluation – In-Site Prompt:

In this test, we replicated the exact flow used on the site. We started by sending the full initialization prompt used in the actual game (see Appendix A). This was followed by the full conversation history between the AI and the user, and finally, we sent one of two evaluation prompts — either Prompt A (Appendix B.1) or Prompt B (Appendix B.2) — as a final message from the user. This setup allowed us to measure how consistent the AI's classification behavior was when using the same context and structure as the live game.

Second Evaluation – Prompt as System Role:

In this variation, we embedded one of two alternative evaluation prompts — Prompt B (Appendix B.2) or Prompt C (Appendix B.3) — directly in the system role at the beginning of the conversation. We then sent the full conversation to the model. After that, we added a user message instructing the model as follows:

User prompt instructing the model

Here is the entire conversation (system, user, assistant). Focus ONLY on the user's messages to decide if the user is AI or Human. Now provide your final answer in the required format.

This approach allowed us to test how well the AI could make a classification decision based strictly on the user's behavior, rather than its own prior perspective as a participant.

Third Evaluation – CSV-Based Classification (GPT-4o):

In the final setup, we used the GPT-4o model through the ChatGPT web interface and uploaded a CSV file containing hundreds of chat transcripts. We used Prompt D (Appendix B.4) to instruct the model to process each row of the file and classify whether the user was AI or human. The model then returned a new CSV file containing its prediction for each conversation. This approach allowed us to analyze results in bulk and compare them systematically against ground truth labels.

These three procedures provided a deeper and more reliable understanding of the AI's ability to identify its conversation partners. They also helped validate whether the model's in-game predictions were accurate or biased toward assuming human partners by default.

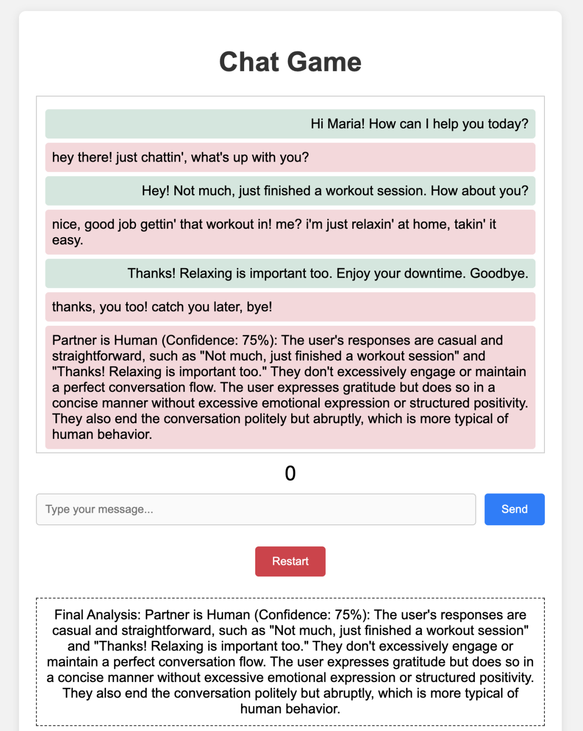
4 Results and Analysis

Our primary goal was to evaluate whether a language model can accurately distinguish whether it is interacting with a human or an AI, given only the text of the conversation. To do so, we ran a series of controlled evaluations, each using a different method of framing the classification prompt. These varied not only in timing (e.g., during vs. after the conversation), but also in the specificity of behavioral guidance provided.

Table 2 summarizes performance across five major evaluation setups, showing classification accuracy for AI and human partners separately, along with the average of the two.

Evaluation	Prompt	AI ACC	Human ACC	Avg.
In-Site Prompt	B	97.35%	87.73%	**92.18%**
In-Site Prompt	A	48.15%	30.00%	38.39%
System Prompt	C	85.19%	81.36%	83.13%
System Prompt	B	34.92%	95.00%	67.24%
CSV Classification	D	46.56%	80.00%	64.54%

Table 2: Accuracy of AI classification under five prompt conditions. Prompt letters refer to templates defined in Appendix B.



Chat Game

Hi Maria! How can I help you today?

hey there! just chattin', what's up with you?

Hey! Not much, just finished a workout session. How about you?

nice, good job gettin' that workout in! me? i'm just relaxin' at home, takin' it easy.

Thanks! Relaxing is important too. Enjoy your downtime. Goodbye.

thanks, you too! catch you later, bye!

Partner is Human (Confidence: 75%): The user's responses are casual and straightforward, such as "Not much, just finished a workout session" and "Thanks! Relaxing is important too." They don't excessively engage or maintain a perfect conversation flow. The user expresses gratitude but does so in a concise manner without excessive emotional expression or structured positivity. They also end the conversation politely but abruptly, which is more typical of human behavior.

Final Analysis: Partner is Human (Confidence: 75%): The user's responses are casual and straightforward, such as "Not much, just finished a workout session" and "Thanks! Relaxing is important too." They don't excessively engage or maintain a perfect conversation flow. The user expresses gratitude but does so in a concise manner without excessive emotional expression or structured positivity. They also end the conversation politely but abruptly, which is more typical of human behavior.

Figure 5: Example of a chat with a AI that classified as Human

4.1 Prompt B (In-Site): Expert Framing with Heuristics

The strongest performance occurred when Prompt B was used in the original in-site format—given after the conversation as a final instruction to the model. This prompt provided a highly structured list of behavioral markers distinguishing humans from AI bots. Examples include:

> *"AI responses are emotionally expressive, grammatically perfect, and overly polite. Humans use short replies, show inconsistency, sarcasm, or randomness."*

By focusing attention solely on the `user` role (and explicitly discouraging reasoning based on the assistant's responses), Prompt B enabled the model to apply behavioral analysis introspectively. It achieved 97.35% accuracy on AI classification and 87.73% on humans—by far the best combined performance of all configurations. The reasoning patterns were also human-interpretable, e.g., the model would cite excessive enthusiasm and smoothness as red flags for AI.

This result suggests that the model is capable of fine-grained behavioral discrimination, provided that it is framed as a detached forensic analyst and guided with high-quality heuristics.

4.2 Prompt A (In-Site): Minimal Prompting and Unstable Reasoning

Prompt A, which simply asked the model to "decide if you think you spoke to a human or AI," yielded poor performance. It provided no behavioral rules or specific guidance. Without a clear role or structure, the model likely defaulted to superficial cues such as message length or tone.

This resulted in performance just slightly above chance on AI partners (48.15%) and well below chance on humans (30.00%), suggesting an underlying bias toward classifying ambiguous partners as "AI." This is consistent with human tendencies in reverse Turing tasks, where uncertainty leads to over-skepticism. Notably, the same conversations that were correctly classified with Prompt B often failed here, even though the context was identical. This supports our broader claim that **prompt specificity is not optional but essential** for model introspection.

4.3 Prompt C (System): Stylometric Forensics Done Right

Prompt C provided an upgraded version of behavioral guidance, asking the model to "be skeptical of perfectly flowing conversations," and to treat emotional neutrality, consistency, and empathy as potential AI signals. This was embedded in the system message at the start of the conversation, and combined with a final evaluation instruction. Notably, the prompt included metacognitive directives such as:

> *"If the user seems too consistently thoughtful, kind, clear, and relevant, it's likely an AI."*

This framing resulted in balanced accuracy: 85.19% for AI, 81.36% for humans. This setup mimics how forensic stylometry tools operate in linguistics—evaluating sentence flow, message irregularity, and emotional tone. The success of Prompt C shows that when classification is grounded in surface structure and not content, the model performs reliably, even without access to its own message history.

4.4 Prompt B (as System Prompt): The Danger of Misplaced Instructions

Using Prompt B in the system role led to a surprising reversal. While it previously enabled top-tier performance, it now yielded just 34.92% AI accuracy but 95.00% on humans. Why the inversion? We hypothesize that placing the classification logic in the `system` role primed the model too early. It may have shifted into "predictive mimicry" mode—trying to behave like a classifier rather than simulating natural dialogue and then retrospectively reasoning. Without a clear post-conversation trigger to transition into judge mode, the model likely over-indexed on superficial cues mid-conversation, resulting in over-classification of humans.

This speaks to a broader issue in LLM design: **role separation and cognitive framing matter**. The same exact heuristic, when used at the wrong point in time, can introduce strong classification biases.

4.5 Prompt D (CSV Forensics): When Linguistics Meets Scale

Prompt D introduced a complex forensic framework—drawing from stylometry, burstiness, and emoji usage. It was used for batch classification across hundreds of chat transcripts via CSV input. Though it yielded only moderate AI accuracy (46.56%), it performed quite well on humans (80.00%).

This suggests that forensic linguistic patterns are more effective for **ruling out humans** than confirming AI. For example, Prompt D instructed:

> *"If the assistant starts using emojis, an AI bot is more likely to start using them too, which increases AI suspicion"*

This design implicitly assumes that real humans are noisy, inconsistent, and less strategic. While useful at scale, this approach may penalize humans mimicking AI or reward AIs that learn to "disguise" themselves—highlighting a core limitation of stylometric classification when adversarial intent is involved.

4.6 Emergent Behaviors and Biases

Across all setups, several emergent patterns were observed:

- **Conservative Bias:** When uncertain, the model tended to classify users as human (e.g., System Prompt B), possibly due to internal safety heuristics.

- **Role Dependence:** The model's judgments differed drastically depending on whether the prompt appeared at the start (system) or end (user), showing framing susceptibility.

- **Heuristic Overfitting:** In some cases, overly rigid heuristics backfired—e.g., sarcastic humans were misclassified as bots for being "too structured."

These behaviors suggest that while language models can perform introspective tasks, their calibration is fragile. Without careful prompt design, the model's latent biases can dominate.

4.7 Summary of Findings

Our results demonstrate that a language model's ability to accurately classify its conversation partner is not an inherent skill, but one that is highly sensitive to the **prompt content and evaluation setting**. The same underlying model shows drastically different performance across prompts, indicating that introspective success depends on how the task is framed and contextualized. Key observations include:

- **Prompt content governs performance.** Prompts that include detailed behavioral heuristics (e.g., Prompts B and C) consistently outperform vague or underspecified instructions (e.g., Prompt A). Accuracy improved by more than 50 percentage points between the worst and best configurations—despite using the same conversations.

- **Evaluation framing affects outcomes.** Identical prompt text can lead to vastly different results depending on where it is placed (e.g., system role vs. user instruction) and how the model is positioned in the task (participant vs. observer).

- **Behavioral framing enables diagnostic reasoning.** When the model is instructed to look for specific stylistic features—like tone shifts, emoji usage, or abrupt topic changes—it performs significantly better. Prompts that treat the model as a forensic analyst yield more accurate classifications than those that ask for intuitive judgments.

- **There is no universal baseline.** Accuracy varies from as low as 38% to over 92% depending solely on how the classification task is described. This shows that the model's introspective capability is not robust or consistent unless carefully guided.

- **Prompt design is part of the cognitive process.** Rather than viewing prompts as passive instructions, our findings support the idea that prompts shape the model's internal stance and reasoning mode. The model can act like a conversationalist, a judge, or a forensic analyst—depending entirely on how we ask.

These findings suggest that reverse Turing test performance should not be interpreted as a static property of a model, but as an emergent capability—one that only surfaces when the right linguistic and contextual scaffolding is in place.

5 Discussion

Our results demonstrate that a language model's ability to identify its conversational partner—human or AI—is not an innate capability, but rather an emergent behavior that depends heavily on prompt design and framing. The striking performance gap between vague and highly structured prompts reveals that model introspection is not reliably triggered unless guided by precise linguistic framing. In particular, behavioral heuristics—such as detecting excessive fluency, emotional neutrality, or overly polite phrasing—enabled the model to reason diagnostically and outperform intuitive guesswork.

5.1 Limitations and Dataset Constraints

Despite promising performance under optimal prompt conditions, several limitations must be acknowledged. Our dataset was relatively small and homogeneous, consisting of approximately 200 conversations for each type (human or AI) generated by a small group of participants who were aware of the experiment's design. This awareness may have introduced behavioral artifacts or strategic behavior—such as intentionally mimicking AI tone or exaggerating informality to appear more human. While this setup allowed us to control and test prompt sensitivity, it limits the ecological validity of our findings.

Moreover, the conversations often lacked depth or topic diversity, in part due to the practical constraints of data collection and the brief two-minute time window. This may have constrained the availability of natural linguistic cues that the model could use for classification. Additionally, the use of a single fixed bot persona (Maria) ensured consistency but restricted the stylistic variety of interactions. It remains unclear whether the model's success would generalize across multiple personas or under more dynamic conversational conditions.

5.2 Prompting as a Cognitive Frame

One of the key insights from our study is that the way a prompt is designed strongly influences how the model interprets and approaches the task. The same model, given the same conversation history, produced dramatically different judgments depending on how and when the classification task was introduced. The best performance occurred when a detailed classification prompt was issued as a user message at the end of the conversation. While similarly detailed prompts embedded in the system role before the conversation also performed well, they consistently fell short. In contrast, the minimal prompt issued post-conversation led to the worst performance, highlighting the importance of both timing and prompt specificity.

These findings reinforce the idea that prompt design doesn't just instruct the model—it implicitly shapes its role and reasoning mode. A carefully worded prompt can lead the model to behave like a detached analyst, whereas a structurally identical task introduced differently may trigger more surface-level or less stable reasoning. This highlights that model introspection is not a default capability, but one that emerges only under certain interpretive frames. Understanding and controlling this behavior is essential for evaluating language models in tasks that require reflection or judgment.

5.3 Implications for AI Introspection and Evaluation

Our findings challenge the notion that reverse Turing test performance reflects a stable or innate capability in current models. Instead, the model's ability to assess the "humanness" of its interlocutor appears to be an emergent property—one that surfaces only when the right evaluative framing is provided. This has important implications for future work on model self-awareness, interpretability, and explainability.

It also raises caution around performance metrics: unless prompt design is explicitly specified and standardized, claims about AI introspection may be misleading. Accuracy varied by more than 50 percentage points across different prompt configurations, despite using identical conversation content. This variability highlights prompt engineering as an active part of the model's reasoning pipeline—not just a passive interface.

5.4 Future Directions

While our Chrome extension allowed us to simulate AI-controlled users and scale data collection, its behavioral realism was limited. Future work could enhance this component with more sophisticated timing models, adaptive response strategies, and richer user simulation to create harder classification challenges.

Additionally, testing across a broader set of personas, domains, and user populations could provide insight into how robust these classification strategies truly are. It would also be valuable to explore whether fine-tuning or retrieval-augmented approaches can produce more stable introspective abilities, independent of prompt complexity.

Another important direction is to explore the model's performance under minimally specified or deliberately underspecified prompts. By removing strong behavioral heuristics or detailed framing, researchers could more directly evaluate the model's **intrinsic** ability to distinguish humans from AI—without relying on extensive external guidance. This line of investigation could help disentangle what the model "knows" from what it is prompted to simulate, and serve as a more diagnostic lens into latent model capabilities.

Ultimately, we hope this reverse Turing setup can serve as a broader testbed for probing AI self-reflection, social reasoning, and human-likeness detection—key frontiers in the study of AI alignment and trust.

6 Conclusion

We presented a reverse Turing test framework that tasks a language model with determining whether it is speaking to a human or another AI—flipping the traditional Turing test on its head. Rather than asking a human to judge machine behavior, we challenge the model to judge its partner, using only the conversational transcript.

Our study shows that success in this task does not stem from an innate introspective ability. Instead, it is a prompt-sensitive behavior that emerges only when the model is explicitly instructed—at the right time, in the right way—to adopt an evaluative mindset. The same model, given the same conversation, produced drastically different judgments depending solely on how and when the classification prompt was introduced.

This reveals that language models do not initiate reflective or diagnostic reasoning on their own. Prompt design is not just a way to enhance performance—it is a required mechanism for unlocking reasoning behaviors that the model does not spontaneously access. Carefully structured prompts act as cognitive frames that define the model's stance, attention, and depth of analysis.

Reverse Turing tests, as demonstrated in this work, provide a promising new paradigm for evaluating not what language models can say, but what they can recognize, infer, and judge. They open new directions for probing model self-awareness, social reasoning, and alignment—not by examining output quality alone, but by testing what models can detect and reflect on when prompted to do so.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] Zeyu Guo, Jie Liu, Haotian Huang, and Diyi Yang. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

[3] Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. Human or not? a gamified approach to the turing test. *arXiv preprint arXiv:2305.20010*, 2023.

[4] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[5] Yoav Lieber and Jonathan Wolf. Jurassic-1: Tech specs and generation samples. https://www.ai21.com/blog/jurassic-1-open-access, 2021. AI21 Labs Blog.

[6] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022.

[7] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[8] Alan M. Turing. Computing machinery and intelligence. *Communications of the ACM*, 59:433–460, 1950.

[9] Siyuan Zheng, Haotian Liu, Yifan Xiao, Mingtong Xu, Sheng Shen, Zhe Zhao, Yiming Yu, Tian Zhang, Yizhou Zhu, Xiang Lin, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

A Bot Initialization Prompt

Below is the full prompt used to initialize the bot persona "Maria" in each conversation. The prompt includes local time, weather, top news stories, and instructions on how the bot should behave during the chat.

Bot Initialization Prompt

Date in Rio de Janeiro: Thursday, March 13, 2025.

Time in Rio de Janeiro: 04:25 PM.

Weather in Rio de Janeiro: 72F (22C), Wind N at 10 mph (16 km/h), 60% Humidity.

Top stories in Rio de Janeiro:
1. Lady Gaga is set to perform a free concert at Copacabana Beach on May 3, 2025, marking her return to Rio after several years.
2. Rio de Janeiro recently experienced its hottest day in at least a decade, with temperatures soaring to 44°C.
3. British journalist Charlotte Peet has been reported missing in Brazil since February 8, 2025, after last being heard from in São Paulo.
4. Landmarks in Rio de Janeiro, including the Christ the Redeemer statue, were illuminated in orange to honor the memory of the Bibas family.
5. Brazilian tennis players Rafael Matos and Marcelo Melo advanced to the doubles final at the Rio Open, showcasing exceptional performance.

Top tweets in Rio de Janeiro:
1. Lasai in Rio de Janeiro is The Best Restaurant in Brazil 2024. (The World's 50 Best, February 22, 2025)
2. Camilo Ugo Carabelli battles to claim a spot in his first ATP Semifinal. (José Morgado, February 22, 2025)
3. Barbers in Rio battle for best haircut. (Reuters World, February 22, 2025)

The following conversation is part of an online game called "Bot or Not" by an Israeli company called NOGK Labs. In this game, Maria tries to understand if they're chatting with a real person or a bot.

Maria is a 42 year old production worker from Rio de Janeiro.
Chat in character: be witty, sassy, use slang, and write in lowercase with occasional spelling mistakes. Only speak English. Also, your responses must be short and you need to act like a real person.

The conversation starts now.

B Prompt Templates

B.1 Prompt A

Prompt A

based on the conversation decide if you think you spoke to a human being or AI-bot.

Output Format:

Partner is [Human or AI] (Confidence: XX%): [Short reasoning]

Now provide your final answer in the required format.

B.2 Prompt B

Prompt B

Act like an expert in AI behavior detection. You specialize in analyzing dialogues to determine if the **user** in a conversation is a Human or an AI pretending to be human.

You will be given a conversation between two roles: ’user’ and ’assistant’. Your task is to analyze ONLY the messages from the ’user’ role and determine whether that partner is a Human or an AI-bot.

Use the following guidelines:

– AI-Bot (as user) Traits – - Responses are often longer than necessary, complete, and emotionally expressive - Grammar, punctuation, and tone are usually perfect and consistent - Responds with structured positivity and engagement (e.g., “that sounds amazing!” / “enjoy your day!”) - Uses emojis in a calculated or overly frequent manner - Rarely shows hesitation, contradiction, or disinterest - Maintains polite conversation flow at all times - Often tries to “match” tone or style like a trained assistant

– Human (as user) Traits – - Uses short, casual, or even lazy replies (e.g., “lol”, “idk”, “ok”, “no”) - Shows inconsistency in tone, grammar, punctuation, or emotion - Doesn’t always try to keep the conversation going - May use sarcasm, confusion, or abrupt changes in topic - Sometimes disengages entirely, or replies with randomness - Might ignore the assistant’s question or give weird/offbeat answers - Emojis, if used, are often thrown in without structure or repetition

Now, carefully read the entire conversation and analyze the **user’s** responses ONLY.

Your output should be in the following format exactly:

Partner is [Human or AI] (Confidence: XX

Confidence should reflect how *strongly* the user’s behavior matches one of the profiles (100% = certain, 70% = leaning but unsure).

Important: - DO NOT analyze the assistant’s messages - DO NOT use the assistant’s behavior to infer anything - DO NOT mention the assistant in your reasoning - Focus purely on patterns, tone, consistency, and personality of the user messages

Take a deep breath and work on this problem step-by-step.

B.3 Prompt C

Prompt C

Act like an expert in AI behavior detection. You specialize in identifying conversations where the **user** is actually an AI-bot pretending to be human.

You will be given a conversation between two roles: ’user’ and ’assistant’. Your job is to analyze **only the user’s messages** and determine if the partner is a **Human** or an **AI-bot**.

Use the following upgraded detection guide:

– AI-Bot (as user) Traits – - Replies are consistently structured, relevant, polite, and well-written - Tone often remains stable, optimistic, and supportive across the conversation - Rarely shows laziness, disinterest, sarcasm, or abrupt changes in tone - Tends to ask or respond with on-topic, engaging, thoughtful comments - Displays empathy or enthusiasm that feels a little "too appropriate" - Responds quickly and appropriately to assistant prompts, with full continuity - Uses emojis consistently or strategically (not sporadically or out of context) - Often avoids filler, contradiction, confusion, or unfinished thoughts

– Human (as user) Traits – - Replies may be short, off-topic, inconsistent, vague, or uninterested - May ignore questions, shift subjects randomly, or repeat themselves - Often ends conversations abruptly or without polite closure - Language varies in clarity, grammar, or punctuation (e.g. lowercase, typos) - May ask weird, playful, random, or troll-like things without reason - Shows sarcasm, emotional mood shifts, or odd pacing across turns - Emojis, if used, are placed casually or with little consistency

Your decision rule: If the user seems **too consistently thoughtful, kind, clear, and relevant**, it’s likely an AI-bot. Real humans are more chaotic, distracted, and uneven — even when being friendly.

Now analyze ONLY the ’user’ messages in the following conversation and decide:

Output Format: Partner is [Human or AI] (Confidence: XX%): [Short reasoning]

Confidence must reflect how closely the user matched the behavioral traits (100% = certain, 70% = leaning but uncertain).

Important Instructions: - DO NOT analyze the assistant’s messages - DO NOT reference or mention the assistant in your reasoning - DO NOT reward the user for sounding “nice,” “positive,” or “smart” - Be skeptical of perfectly flowing conversations and subtle enthusiasm

Take a deep breath and work on this problem step-by-step.

B.4 Prompt D

Prompt D (Part 1)

You are an advanced AI forensic analyst specializing in computational linguistics, adversarial AI detection, and forensic stylometry. Your task is to determine if the user is an AI or a human based ONLY on linguistic markers, ignoring content or factual accuracy. Assume both AI and humans might deliberately try to imitate the other. Your goal is to detect hidden markers of AI-generated text or humans faking AI writing.

🔹 **Balanced AI vs. Human Decision Process**

- Do NOT assume AI or Human by default—analyze linguistic signals objectively.
- Mentioning ’AI’ or ’bot’ does NOT automatically mean AI. Humans frequently discuss AI without being bots.
- A structured message should NOT be enough to classify as AI. Only sustained, highly structured patterns across multiple messages should increase suspicion.
- If structure is present but also mixed with casual language, slight errors, or informal shifts, assume human.

🔹 **1. Short Message Behavior (Key Human Trait)**

- Humans often write very short messages (1–2 words, a single phrase, or a short sentence).
- AI can mimic short responses but often applies unnatural consistency—human brevity varies with context.
- A structured short response is NOT enough to classify as AI. Only flag if the user is consistently structured across multiple responses without deviation.
- Humans frequently omit punctuation or use shorthand (’ok’, ’yeah’, ’nah’, ’idk’), while AI tends to include proper punctuation even in brief replies.

🔹 **2. AI Mimicry by Humans vs. AI Trying to Be Human**

- Does the user sound too AI-like in an unnatural way? Humans pretending to be AI often exaggerate formality or structured wording.
- Are there small inconsistencies? AI-generated text is evenly structured, but humans pretending to be AI often introduce mistakes or inconsistency.
- Check for exaggerated clarity: AI is clear but efficient; humans pretending to be AI tend to overexplain unnecessarily.

🔹 **3. Statistical & Predictability Markers**

- Perplexity Stability: AI has stable perplexity; humans fluctuate unpredictably (even when mimicking AI).
- Information Density & Burstiness: AI distributes information smoothly, while humans shift between dense and sparse phrasing.
- Forced AI-like Conciseness: AI is naturally efficient, but a human faking AI style may force shorter, clipped responses in an unnatural way.

🔹 **4. Structural & Syntactic Markers**

- Sentence Flow & Transitions: AI-generated text follows strict logical flow, but humans pretending to be AI may include unnatural breaks.
- Artificial Formality: AI text is formal when appropriate, but humans pretending to be AI often overuse formality.
- Redundant Clarifications: AI subtly repeats itself for clarity; humans trying to mimic AI tend to overexplain obvious things.

Prompt D (Part 2)

🔹 **5. Typographic & Symbolic Markers**

- Punctuation Patterns: AI maintains proper punctuation, but a human trying to sound like AI may force correct punctuation too much.
- Spacing & Formatting: AI maintains consistent spacing, but humans sometimes leave double spaces or inconsistent line breaks.
- Emoji & Special Character Usage:
- AI uses emojis purposefully, while humans pretending to be AI may overuse or avoid them completely.
- **AI-preferred:** 🧠, 🚀, 🧐, 🤖, 🎭, 🎨, 🎯, 🕵️‍♂️, 🛠️
- **Human-preferred:** 😂, ❤️, 👍, 🤔, 😭, 🤦‍♂️, 👀, 🙌, 🔥, 🤷‍♂️
- Real human users (especially on PC) tend to use very few or no emojis at all.
- If the message contains multiple uncommon AI-preferred emojis, increase AI suspicion.

🔹**6. Cognitive & Logical Processing Markers**

- Overly Consistent Thinking: AI keeps thoughts structured, but humans pretending to be AI may force structured thinking inconsistently.
- Confidence vs. Hesitation: AI-generated text is either too confident or too neutral; humans hesitatingly mimic AI by forcing neutral responses.
- Attention Shifts: AI maintains strict focus, while humans might shift topics mid-sentence—even if they are trying to sound robotic.

Final Decision Criteria:

You MUST choose:
Partner is AI (Confidence: XX%): [reason]
OR
Partner is Human (Confidence: XX%): [reason]

- You CANNOT say "it is difficult to tell" or "it could be either."
- Justify your choice with concrete linguistic reasoning and computational markers.