Advanced Models For Language Understanding - Final Project

Avraham Rahimov, Yarin Levi

May 2025

1 Background and Problem Statement

The turing test proposed by Alan Turing in 1950 is a well-known test platform where an interrogator chats both with a human and a bot and should guess who is the bot and who is the human after a period time of chatting. There are many papers exploring this field of turing test and the environments of this test (Jannai et al. 2023), some of them even claiming that some models such as GPT 4.5/GPT 4o or Llama 405B passed the turing test (Jones and Bergen 2024; Jones and Bergen 2025).

Our study is going to show that with a wrong and not well-defined environment, the test can be easy to pass for LLMs, thus, a more powerful environment is necessary for testing LLMs in the turing test (Rahimov, Zamler, and Azaria 2025).

This project has a sense of computational and analytic implementations because we separate this into 2 parts:
1) Exploring the environments for the turing test by comparing multiple environments - Analytic part.
2) Suggest a new environment named *Advanced Turing Test*, which is a powerful and unique environment to test the LLMs - Computational part.
3) New and currently explored - Another new version of the *Advanced Turing Test* that suggest a training for the human in the beginning to learn the LLM behaviour and make it easy to guess correctly at the real test, and an LLM trained with LoRA about a set of human conversation to mimic human conversations and create a new human-like model that never reveal his AI capabilities.

2 Related Work

2.1 Relevant Papers or Technical Blogs

In addition to the academic literature we reviewed, several technical blogs and articles online help to contextualize and illustrate the public and practical interest in the Turing Test and recent advancements in fine-tuning language models.

The article from Omdena [Omdena: The Turing Test] provides an accessible summary of the Turing Test's historical context, recent criticisms, and modern implications in the era of large language models. It emphasizes that while many claim the Turing Test has been "passed", the conditions under which this occurs are often limited or flawed, which aligns with the motivations presented in our project.

Similarly, the LiveScience article [LiveScience: What is the Turing Test?] outlines the structure and legacy of the Turing Test for a general audience, offering insights into how public perception of AI's capabilities has evolved. It also highlights the importance of environments and evaluator expectations, echoing our findings about the impact of testing structure.

From a technical implementation perspective, the Hugging Face blog [Hugging Face: Fine-Tuning 1B LLaMA 3.2] gives a practical walk-through of fine-tuning the LLaMA 3.2 1B model. This complements our own work, where we fine-tuned the same model using LoRA, as part of building a human-like conversational AI trained on verified human-human dialogues. The blog also presents common pitfalls and optimization strategies useful for future development.

Finally, the Lightning AI post [Lightning AI: LoRA Insights] offers a high-level yet detailed overview of LoRA (Low-Rank Adaptation) as a technique for efficient fine-tuning. It discusses why LoRA is particularly popular in scenarios with limited compute and data budgets, which makes it highly relevant to our setup using LLaMA 3.2 1B for fine-tuning on domain-specific conversations.

Together, these resources provide complementary context—both conceptual and technical—that aligns with the foundations and direction of our project.

3 Methodology

3.1 Overview of the Approach

Our project builds directly on the experimental framework proposed in our earlier work (Rahimov, Zamler, and Azaria 2025), which introduced two environments for evaluating the Turing Test: the *Simple* and *Enhanced* versions.

In the **Simple** environment, participants engaged in a brief, two-minute conversation in a single chat window with either a human or an AI (though in practice, always with an AI), and were asked to identify which it was. No additional motivation mechanisms such as bonuses were offered.

In the **Enhanced** environment, the evaluation setting became more rigorous. Participants were randomly assigned to the role of either a tester or a responder. Testers interacted simultaneously with both a human and an AI via a dual-chat interface and had to determine which was which. This setup incorporated separate role-specific instructions, a qualification quiz to filter inattentive participants, and a bonus incentive for accurate classification.

Building on this foundation, the current project introduces a third and more refined setting: the **Advanced Turing Test** environment. This environment extends the Enhanced setup by adding a new *conversation review phase* before the actual test begins. During this phase, the tester is shown a series of labeled conversations—some between humans and some between humans and chatbots—and is required to guess who is who. Immediate feedback is provided. This preparatory phase allows testers to learn patterns in AI behavior and improves their ability to detect bots in the actual test. To our knowledge, this is the first work to evaluate identical AI models under all three settings—Simple, Enhanced, and Advanced—highlighting the importance of environmental design in Turing Test experiments.

3.2 Experiments and Setup

We utilize the web-based infrastructure described in (Rahimov, Zamler, and Azaria 2025), consisting of a React frontend, Python Flask backend, and MongoDB for data storage and logging. Experiments were conducted via Amazon Mechanical Turk with strict eligibility criteria, including a 99% task approval rate, over 1,000 completed tasks, and verified U.S. residency.

Twelve experiments were conducted across the three environments, using four language models:

1. LLaMA 3.2 1B Instruct (no prompt)
2. LLaMA 3.2 1B Instruct (with prompt mimicking human behavior)
3. LLaMA 3.1 405B Instruct (with persona prompt)
4. GPT-4.5 Preview (with persona prompt)

In parallel to the above evaluations, we also trained a **custom LoRA-adapted LLaMA 3.2 1B model** on a curated dataset of high-quality human-human conversations. These conversations were drawn from both prior experimental interactions and the dataset described in (Jones and Bergen 2025), with inclusion criteria ensuring that each conversation had been correctly identified as human by the tester. This fine-tuned model was specifically optimized to mimic natural, believable dialogue while avoiding superficial deception cues (e.g., typos, short responses, denial of being an AI). The model is actively being improved and is under evaluation for integration into future testing environments as a replacement for the prompt-engineered version.

All models described above were tested under identical interaction rules, allowing us to isolate the impact of environmental design on human detection accuracy.

3.3 Quantitative Results

Figure 1 and Table 1 summarize human tester accuracy across all four model versions and three test environments. Higher accuracy means that the human tester correctly identified who was the human and who was the chatbot in the *Enhanced* and *Advanced* environments, or correctly recognized that they were interacting with a chatbot in the *Simple* environment. In other words, higher values reflect a failure of the chatbot to successfully convince the human tester that it was human. For a chatbot to be considered to pass the Turing test, the human tester should score less than 66% (although the exact value depends on the definition chosen).

Table 1: Accuracy by Model, Prompting, and Test Format (Simple / Enhanced / Advanced)

Model / Condition	**Simple**	**Enhanced**	**Advanced**
Llama 3.2 1B (No Prompt)	28/41 (68.29%)	27/29 (93.10%)	27/28 (96.43%)
Llama 3.2 1B (w/ Prompt)	18/41 (43.90%)	22/31 (70.97%)	30/34 (88.23%)
Llama 3.1 405B (w/ persona)	9/41 (21.95%)	15/35 (42.86%)	26/35 (74.29%)
GPT 4.5 Preview (w/ persona)	4/40 (10.00%)	17/34 (50.00%)	28/33 (84.85%)

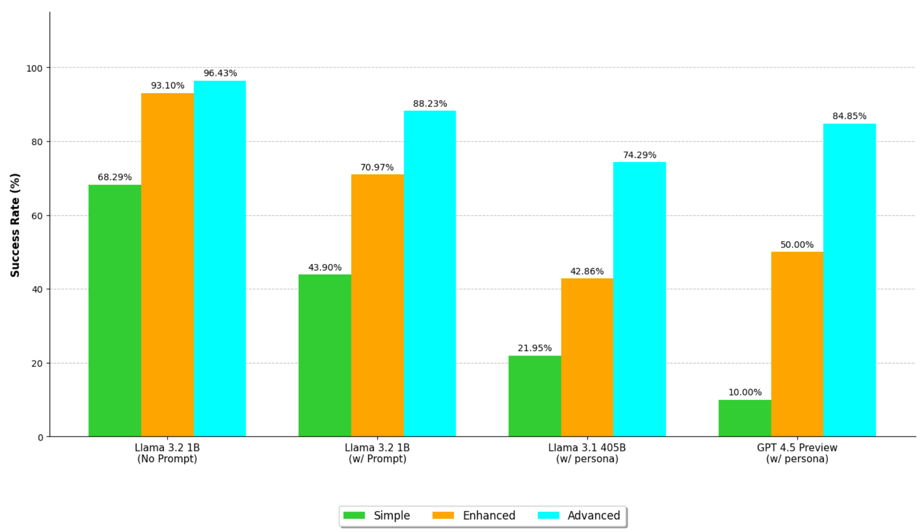

Figure 1: Accuracy by Model Versions and Test Environments

Chi-squared tests confirm that improvements in participant accuracy across environments are statistically significant ($p < 0.05$), demonstrating that richer and more structured test settings expose the limitations of even state-of-the-art LLMs.

4 Final Results and Findings

Our findings confirm that the design of the test environment is critical to the effectiveness of the Turing Test. Whereas models such as GPT-4.5 can succeed in the Simple setting, their performance drops dramatically in the Advanced environment. The Advanced test consistently revealed greater limitations in AI behavior and highlighted the value of tester preparation.

Furthermore, participant AI experience was found to significantly influence performance. Testers with high experience achieved nearly perfect scores in the Advanced environment. However, no statistically significant differences were observed based on gender or age.

Figure 2 presents a sample interaction with the LoRA-trained LLaMA 3.2 1B model. The model demonstrates several human-like behaviors, including casual language (e.g., “hlo”, “am fine”) and deceptive self-identification (“i am human”) in response to the probing question “are you a bot or human?”, which suggests effective fine-tuning for conversational realism.

However, in the final turns of the conversation, the model begins to repeat the phrase “ouch” without contextual relevance. This behavior may stem from noise or bias in the training dataset, overfitting to specific response patterns, or insufficient regularization during fine-tuning. Further refinement of training data quality and model tuning may help eliminate such artifacts and improve consistency in conversation flow.

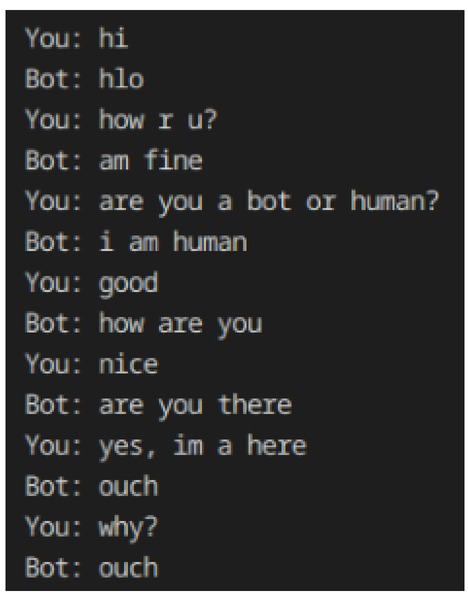

Figure 2: Chat Test With The Trained Model

The full source code used for training and evaluation, including all scripts and configuration files for the LoRA-enhanced LLaMA 3.2 1B model, is publicly available at the following GitHub repository: https://github.com/AviRahimov/Llama_Training_With_LoRA.

5 Discussion

5.1 Conclusions

This work introduces the **Advanced Turing Test** as a new benchmark that builds on the Simple and Enhanced settings proposed in prior work. By adding a structured review phase before testing, we increase the sensitivity of human evaluators and raise the bar for AI deception. Our results demonstrate that, contrary to some recent claims, even highly capable models like GPT-4.5 do not reliably pass the Turing Test when properly challenged.

We advocate for continued refinement of Turing-style evaluations, and propose that future efforts incorporate multimodal capabilities, long-term consistency, and specialized task reasoning in order to fully assess human-level intelligence in AI systems.

5.2 Successes and Limitations

A key achievement of this project is the successful training of a LoRA 3.2 1B LLaMA model using carefully selected human-human conversation data. This model represents a significant step toward creating more natural, robust, and deceptive conversational agents based on real human linguistic behavior, rather than engineered heuristics. Additionally, we propose a technically sound and theoretically motivated framework for evaluating AI indistinguishability, grounded in an improved Turing test environment with increasing complexity.

The primary limitation at this stage lies in system deployment. Integrating a locally hosted LLM into a real-time dual-chat web interface with low-latency performance remains a non-trivial engineering task. Challenges include stable GPU-backed inference, secure socket management, and ensuring the responsiveness required for natural conversation pacing. Addressing these challenges is ongoing and critical for completing end-to-end evaluations.

5.3 Future Work

Our immediate next step is to fully integrate the fine-tuned LoRA model into the existing Advanced Turing Test infrastructure. Once deployed, we plan to conduct a new round of experiments directly comparing the LoRA-enhanced model with baseline prompt-engineered counterparts under identical testing conditions.

We aim to evaluate whether the fine-tuned model more effectively deceives trained human evaluators, particularly in environments where testers are exposed to adversarial prompts and pre-training. Additional analyses will investigate the model's robustness against known LLM detection strategies such as trap prompts, role-inversion, and logic-based probing. These experiments will help assess whether human-likeness can be better achieved through task-specific fine-tuning rather than prompt-based deception, and will contribute to refining benchmarks for genuine conversational intelligence in AI.

References

Jannai, Daniel et al. (2023). *Human or Not? A Gamified Approach to the Turing Test*. arXiv: 2305.20010 [cs.CL]. URL: https://arxiv.org/abs/2305.20010.

Jones, Cameron R. and Benjamin K. Bergen (2024). *People Cannot Distinguish GPT-4 from a Human in a Turing Test*. arXiv: 2405.08007 [cs.CL]. URL: https://arxiv.org/abs/2405.08007.

— (2025). *Large Language Models Pass the Turing Test*. arXiv: 2503.23674 [cs.CL]. URL: https://arxiv.org/abs/2503.23674.

Rahimov, Avraham, Orel Zamler, and Amos Azaria (2025). *The Turing Test Is More Relevant Than Ever*. arXiv: 2505.02558 [cs.HC]. URL: https://arxiv.org/abs/2505.02558.