Large Language Model (LLM) driven Turing Game Bot

Asude Ebrar Kızıloğlu - Advisors: Onur Güngör & Suzan Üsküdarlı
Computer Engineering, Bogazici University

Play the Game!

Problem Statement

To implement a modern Turing Test [1] where an **LLM-driven chatbot** attempts to **convince humans** of its humanity through natural group chat conversations.

Motivation

- Evaluate LLMs' ability to generate **human-like conversational** patterns
- Explore prompt engineering strategies for natural dialogue
- Contribute to the ongoing discussion of **AI capabilities** and **limitations**

Key Requirements

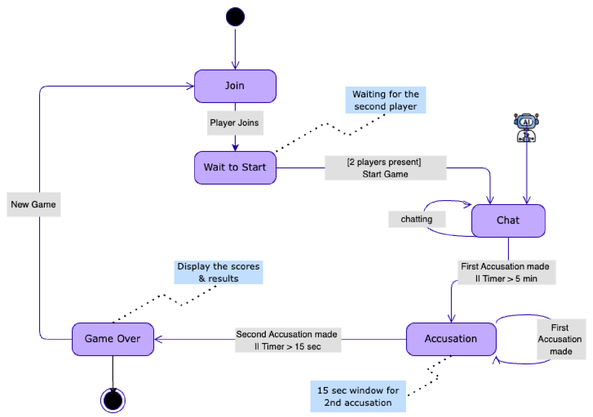
- Bot's Behavior 🤖:
  - Generate human-like chat responses
  - Maintain a consistent persona to avoid detection
- Chat Application ✉️:
  - Support 3-party chat sessions (2 humans + 1 bot)
  - Manage accusation and session timings
- Data Collection 📊:
  - Record game data, accusations, and outcomes
  - Track performance metrics and interaction patterns

System Design

- 5-minute game sessions with **2 humans** and **1 bot**
- Players engage in natural conversation
- Human players can accuse others of being the bot
- Accurate bot detection is rewarded.
- Bot aims to maintain its **human persona**.

We assessed Bot's performance according to the following criteria:

- Do humans successfully detect the bot?
- Does bot manage to talk in human-like manner using natural chat tone?



User Tests

Bot's Prompt:

- *You'll chat like a Turkish university student who is ESL:*
  - *simpler vocabulary & occasional small grammar mistakes*
- *use mostly lowercase and do not use many punctuation marks*
- *stay casual if asked if you're a bot*

Bot's Persona for the User Tests:

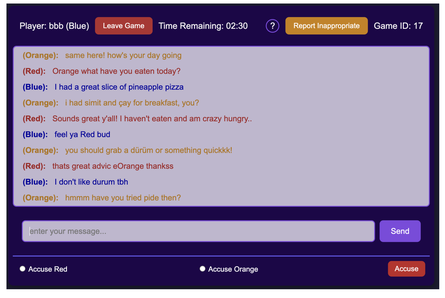
- You're a 20-yo math student at Bogazici University (you call it boun)
- You talk about games, uni life, travels, and what's popular these days

Experiment Setup:

- Conducted 4 experiment sessions
- 24 human participants completed 120 games

Tracked Metrics:

Bot Detection Rate | Conversation Length | Vocabulary Range | Accusation Patterns

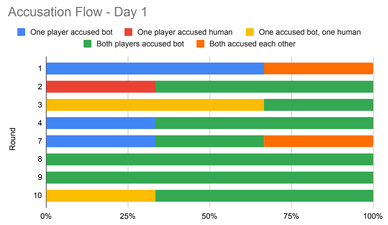


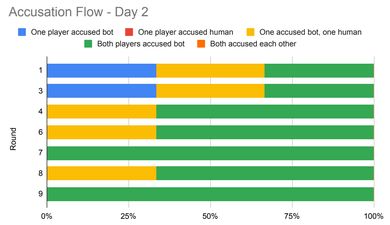
Detection Results

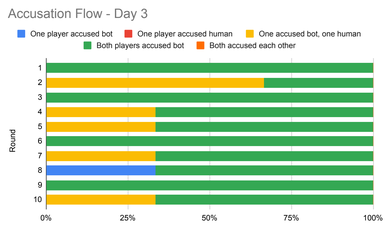
The data displays how players' bot-hunting strategies changed as they played more games. We made key improvements to the bot over 4 days of experiments, and observed how these changes affected human players' accusation patterns.

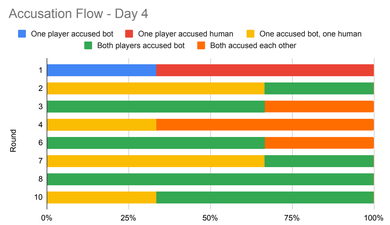
Bot deceived **0.49 people per game**, on average.

Accusation Pattern	Games	%
Both accused human	6	6.45
One player accused bot	8	8.60
Both accused bot	56	60.22
One accused human	3	3.23
One accused bot, one accused human	20	21.51
Total	93	100.00









Day 1: Players quickly caught onto the bot's patterns in early games. By later rounds, they often agreed on who the bot was.
Day 2: Things got more interesting - players started disagreeing more about who the bot was.
Day 3: After we added more natural and varied typing mistakes to the bot, players had a harder time catching it.
Day 4: With random accusation button placement, players often ended up suspecting each other.

Conclusion

- The bot deceived at least one person in about **31%** of games, but its **message length**, **vocabulary**, and **writing style** made it mostly detectable.
- Modifications we made between the experiment days improved the bot's performance of mimicking human-like chat patterns.

Future Work

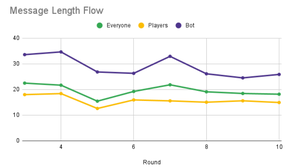
- Prompt improvements for better human-like behavior.
- N-user game set up.
- "I cannot decide" button.
- Bot detection bot for more automated analysis.
- Introduction of hand-raising feature to maintain conversation order.

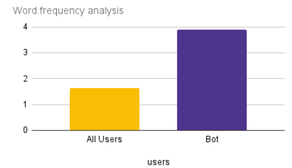
References

[1] A.M. Turing, Computing machinery and intelligence. *Mind*, 1950.

Discourse Results

Messages Analysis





- Bot wrote **longer messages** than humans.
- Bot **repeated words** much more frequently.

Word Clouds Comparison

- Bot kept using the same words (especially "chilling") - humans quickly noticed it!
- Humans showed greater vocabulary diversity.
- While both used common chat words, bot used them in a more robotic, repetitive way.
- Bot used Turkish words ('nasılsın,' 'selam'), but their placement felt forced and mechanical.

ebrarkiziloglu@gmail.com