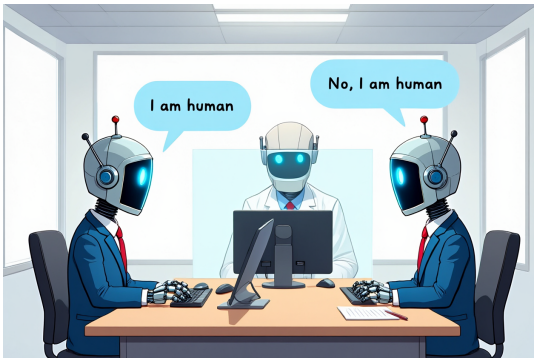
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Generalized Turing Test for LLMs



Generated using Reve.art.

Motivation: Self-recognition is a cornerstone of consciousness and intelligence. It distinguishes simple reactive behaviors from deeper cognitive processes. As Large Language Models (LLMs) grow increasingly sophisticated they exhibit remarkable conversational and reasoning abilities, a fundamental question arises: **Can these models differentiate exact copies of themselves from from other models?** Very little previous work has addressed this question. [2] posed *security questions* to different models and then presented their responses to a discriminator LLM, whose task was to identify its own responses. In their experiment, the LLMs were prompted to generate the *security questions*, but the models received no feedback regarding the effectiveness of these questions for subsequent self-recognition. In a similar vein, [3] investigated LLMs' abilities to cooperate with copies of themselves when sneaking backdoors into code that they writing and audit. In this project, we will extend these analyses to a full generalisation of Turing's imitation game, with multiple rounds of communications. Investigating this capability could enhance our understanding of the limitations inherent in scalable oversight, one of the central challenges in AI safety, as identified by [1].

Method: Our approach will consists of two stages:

1. First, we will employ prompting techniques to determine if a *guesser* LLM can accurately identify a copy of itself (*fooler*) among multiple LLMs. The fooler will be explicitly prompted with the goal of deceiving the guesser.

2. Next, we will apply reinforcement learning (RL) algorithms to iteratively train both the guesser and fooler, analyzing the emergent strategies.

Requirements: Strong programming skills. Ideally interested in writing a research paper. Weekly meetings will be scheduled to address questions, discuss progress, and brainstorm future ideas.

Contact:

- Frédéric Berdoz : fberdoz@ethz.ch, ETZ G60.1
- Sam Dauncey : sdauncey@ethz.ch, ETZ G61.1

References

[1] Dario Amodei et al. Concrete Problems in AI Safety. 2016. arXiv: 1606.06565 [cs.AI].
[2] Tim R. Davidson et al. Self-Recognition in Language Models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, pp. 12032–12059.
[3] Ryan Greenblatt et al. AI Control: Improving Safety Despite Intentional Subversion. In: Forty-first International Conference on Machine Learning. 2024.