
Using ARC-AGI puzzles as CAPTCHa task

Mikolaj Kniejski

With
Apart Research

Abstract

As large language models (LLMs) like GPT-4 become increasingly powerful, their potential to bypass traditional CAPTCHA systems and automate access to websites poses a growing security risk. With LLMs gaining the ability to solve many CAPTCHA tasks, new techniques are necessary to differentiate between human and AI agents. One promising solution lies in using ARC-AGI (Abstraction and Reasoning Corpus for Artificial General Intelligence) puzzles as CAPTCHA tasks. Originally designed to evaluate general intelligence in AI, ARC puzzles require abstract reasoning and problem-solving skills that current AI models lack. In this paper, we explore the potential of using ARC puzzles to protect websites from unauthorized AI access, the challenges of designing such a system, and its implications for the broader field of Agent Security, where AI agents—sophisticated systems integrating language models and other programs—are transforming the landscape of digital security.

Keywords: Agent Security, AI Safety

1. Introduction

LLMs are now proficient at tasks like recognizing distorted text, identifying objects in images, and solving simple arithmetic, which undermines the effectiveness of traditional CAPTCHAs. These methods were built on the premise that humans possess superior pattern recognition and cognitive flexibility, but LLMs have caught up in many of these areas. As outlined in the **Agent Security Hackathon**, AI agents are becoming more sophisticated, capable of autonomously interacting with digital environments. Ensuring they don't gain unauthorized access is increasingly difficult with conventional CAPTCHA systems. **How can we make sure that AIs don't get unauthorized access to digital resources?**

2. Overview

ARC-AGI puzzles, created to test abstract reasoning and generalization in AI, present a unique opportunity for CAPTCHA systems. These puzzles require users to infer rules from a minimal set of examples and apply them to new, unseen cases—something current AI, including LLMs, struggles with. **ARC-AGI leaderboard reports that the best solutions achieve 48% success rate on private ARC-AGI dataset, while some papers report that usual CAPTCHA tasks can be solved with 63.5% accuracy (Deng et al.) or above 90% accuracy (Searles et al.) by some forms deep learning.**

3. Code

In this project, no new code was developed, as existing ARC puzzles and online implementations already provide sufficient resources for experimentation. ARC-based puzzle frameworks are accessible online, and coding new systems from the ground up was deemed unnecessary given the wealth of available tools and author’s interest in pursuing different projects in this hackathon. Example ARC-AGI puzzles can be viewed on this link <https://arcprize.org/play>

4. Discussion and Conclusion

ARC-AGI puzzles provide a novel approach to CAPTCHA design that leverages the current limitations of AI in abstract reasoning and problem-solving. In the broader context of **Agent Security**, preventing AI-driven agents from unauthorized web access is only one part of the larger challenge of building secure, trustworthy AI systems for the future.

5. References

Deng, Gelei, et al. “OEDIPUS: LLM-enhanced Reasoning CAPTCHA Solver.”

2024, <https://arxiv.org/pdf/2405.07496>.

Searles, Andrew, et al. *An Empirical Study & Evaluation of Modern CAPTCHAs*.

2023, <https://arxiv.org/pdf/2307.12108>.