Self-Generated Culture

François Fleuret

July 5, 2024

(work in progress, to be updated)

https://fleuret.org/public/culture/culture.pdf

1 Introduction

The hypothesis behind this experiment is that high-level abstract thinking is fueled by social competition. A group of communicating agents that try to demonstrate their cognitive superiority would end up developing a rich and consistent culture.

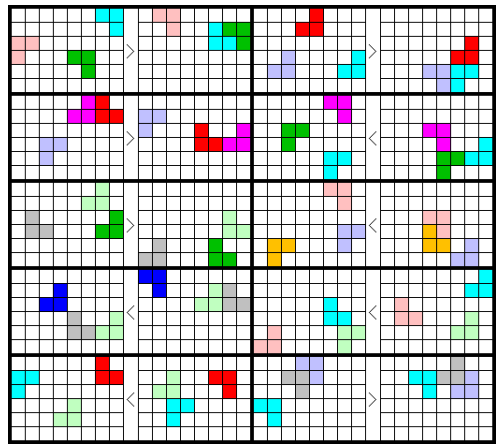
The experiment is designed with a group of GPTs that alternatively learn to solve quizzes and generate new ones.

A "quiz" is a triplet of the form (A, d, B) where A and B are two sequences and d is a token indicating if the direction is forward or backward. Given (A, d), the challenge is to generate B.

The experiments starts with a set of quizzes, that is going to be progressively enriched.

2 Bird World

The initial set of quizzes consist of predicting the dynamics of a very simple world: A 6×8 grid with three colored "birds" moving in a straight line, possibly bouncing on the grid's borders. There are ten different colors.



In each on these quizzes, A is the left image serialized in raster-scan order as a sequence of $6 \times 8 = 48$ tokens, d is either the token "forward" or the token "backward", and B is the right image, also serialized. The direction of prediction is chosen at random.

3 Generating Quizzes

Given a set of N GPTs, we can generate new quizzes as follows: Select one of the models, and use it to generate the 97 tokens of a triplet (A, d, B).

Then with each one of the $N-1$ other models, predict B from (A, d), and A from (B, d') where d' is the direction token opposite of d.

A quiz is validated if **all the other GPTs but one predict it deterministically correctly in both directions.**

This criterion assures that the new quizzes are both solvable and sophisticated, and incrementally complexify the culture. Imposing both direction prevents the generation of quizzes which are not trivial only because the prompt has been randomly degraded.

4 Overall Process

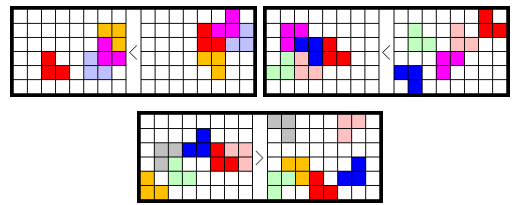
The overall process consists of training the GPTs from scratch by iterating the following steps:

- select the GPT with the lowest recorded test accuracy, train it through one epoch,
- if its test accuracy gets above 97.5%, generate 1'000 new quizzes, add them to the training set, re-compute the accuracy of all the models

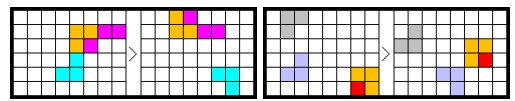
5 Results

This procedure results in the discovery of patterns which are not present in the original quizzes:

More birds



New bird shapes



Occlusions