Summarizing books with human feedback

Scaling human oversight of AI systems for tasks that are difficult to evaluate.

September 23, 2021

Language, Human feedback, Safety & Alignment, Summarization, GPT-3, Milestone, Publication

To safely deploy powerful, general-purpose artificial intelligence in the future, we need to ensure that machine learning models act in accordance with human intentions. This challenge has become known as the *alignment problem*.

A scalable solution to the alignment problem needs to work on tasks where model outputs are difficult or time-consuming for humans to evaluate. To test scalable alignment techniques, we trained a model to summarize entire books, as shown in the following samples.[A] Our model works by first summarizing small sections of a book, then summarizing those summaries into a higher-level summary, and so on.

OpenAI

Original text — 26,449 words
Source: Project Gutenberg

ALICE'S ADVENTURES IN WONDERLAND
Lewis Carroll

THE MILLENNIUM FULCRUM EDITION 3.0

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversations?'

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.

In another moment down went Alice after it, never once considering how in the world she was to get out again.

The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself falling down a very deep well.

thousand miles down, I think–' (for, you see, Alice had learn several things of this sort in her lessons in the schoolroom, an though this was not a VERY good opportunity for showing off h knowledge, as there was no one to listen to her, still it was goo practice to say it over) '–yes, that's about the right distance–bu then I wonder what Latitude or Longitude I've got to?' (Alice ha no idea what Latitude was, or Longitude either, but thought the were nice grand words to say.)

Presently she began again. 'I wonder if I shall fall righ THROUGH the earth! How funny it'll seem to come out among th people that walk with their heads downward! The Antipathies, think–' (she was rather glad there WAS no one listening, this tim as it didn't sound at all the right word) '–but I shall have to as them what the name of the country is, you know. Please, Ma'am, this New Zealand or Australia?' (and she tried to curtsey as sh spoke–fancy CURTSEYING as you're falling through the air! Do yo think you could manage it?) 'And what an ignorant little girl she think me for asking! No, it'll never do to ask: perhaps I shall see written up somewhere.'

Down, down, down. There was nothing else to do, so Alice soo began talking again. 'Dinah'll miss me very much to-night, I shou think!' (Dinah was the cat.) 'I hope they'll remember her saucer milk at tea-time. Dinah my dear! I wish you were down here wi me! There are no mice in the air, I'm afraid, but you might catch bat, and that's very like a mouse, you know. But do cats eat bats wonder?' And here Alice began to get rather sleepy, and went o saying to herself, in a dreamy sort of way, 'Do cats eat bats? Do ca eat bats?' and sometimes, 'Do bats eat cats?' for, you see, as sh couldn't answer either question, it didn't much matter which wa she put it. She felt that she was dozing off, and had just begun dream that she was walking hand in hand with Dinah, and sayin to her very earnestly, 'Now, Dinah, tell me the truth: did you ev eat a bat?' when suddenly, thump! thump! down she came upon

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

The original text is divided into sections, and each section is summarized.

66 summaries — 6,024 words

Alice is bored sitting by her sister on the bank, and she's thinking about making a daisy chain when a white rabbit with pink eyes runs by. She's surprised to see a rabbit with a waistcoat pocket and a watch, and she follows it down a rabbit

Alice thinks she's fallen a thousand miles, and wonders wh latitude or longitude she's at. She wonders if she'll fall rig through the earth and what the name of the country she'll emer in is. She starts to get sleepy and starts to dream about walkin

miles she's fallen.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Section summaries are summarized again into higher-level summaries.

6 summaries — 830 words

ALICE falls down a deep well and lands in a heap of sticks and leaves. She follows the White Rabbit down a long passage, but loses sight of it. She finds a bottle that says "Drink Me" on it and drinks it, shrinking to 10 inches tall. She grows to 9 feet tall and then shrinks again. She cries and grows larger again. A mouse appears and Alice asks it for help. The mouse agrees to tell Alice its story if she'll take it to the shore. Once there, the animals discuss how to get dry again. The Dodo suggests a Caucus-race to get dry. Everyone crowds around Alice asking for prizes. Alice pulls out a box of comfits and hands them out as prizes. The Mouse then asks Alice if she has a prize for herself, and she says she only has a thimble. The Dodo takes the thimble and presents it to Alice, saying they all beg her acceptance

ALICE grows larger after drinking a bottle she finds on a tabl She stops growing before she gets too big. She then starts miss being at home and wishes she hadn't gone down the rabb hole. She hears the Rabbit calling for her and realizes she's no much bigger than the Rabbit, so she's no longer afraid of it. Ali decides her first priority is to grow back to her normal size, ar her second priority is to find her way into the garden. She doesn know how to do either of those things, however. She sees Caterpillar smoking a hookah on a mushroom and asks him ho she can grow back to her normal size. The Caterpillar tells her th one side of the mushroom will make her grow taller, and the oth side will make her shorter. She starts eating the mushroom.

1 2 3 4 5 6

The summarizing process continues until a complete summary is achieved.

Complete summary — 136 words

ALICE falls down a rabbit hole and grows to giant size after drinking a mysterious bottle. She decides to focus on growing back to her normal size and finding her way into the garden. She meets the Caterpillar who tells her that one side of a mushroom will make her grow taller, the other side shorter. She eats the mushroom and returns to her normal size. Alice attends a party with the Mad Hatter and the March Hare. The Queen arrives and orders the execution of the gardeners for making a mistake with the roses. Alice saves them by putting them in a flowerpot. The King and Queen of Hearts preside over a trial. The Queen gets angry and orders Alice to be sentenced to death. Alice wakes up to find her sister by her side.

Our best model is fine-tuned from GPT-3 and generates sensible summaries of entire books, sometimes even matching the average quality of human-written summaries: it achieves a 6/7 rating (similar to the average human-written summary) from humans who have read the book 5% of the time and a 5/7 rating 15% of the time. Our model also achieves state-of-the-art results on the BookSum dataset for book-length summarization. A zero-shot question-answering model can use our model's summaries to obtain competitive results on the NarrativeQA dataset for book-length question answering.[B]

Our approach: combining reinforcement learning from human feedback and recursive task decomposition

Consider the task of summarizing a piece of text. Large pretrained models aren't very good at summarization. In the past we found that training a model with reinforcement learning from human feedback helped align model summaries with human preferences on short posts and articles. But judging summaries of entire books takes a lot of effort to do directly since a human would need to read the entire book, which takes many hours.

To address this problem, we additionally make use of *recursive task decomposition*: we procedurally break up a difficult task into easier ones. In this case we break up summarizing a long piece of text into summarizing several shorter pieces. Compared to an end-to-end training procedure, recursive task decomposition has the following advantages:

1. Decomposition allows humans to evaluate model summaries more quickly by using summaries of smaller parts of the book rather than reading the source text.
2. It is easier to trace the summary-writing process. For example, you can trace to find where in the original text certain events from the summary happen. See for yourself on our summary explorer!
3. Our method can be used to summarize books of unbounded length, unrestricted by the context length of the transformer models we use.

Why we are working on this

This work is part of our ongoing research into aligning advanced AI systems, which is key to our mission. As we train our models to do increasingly complex tasks, making informed evaluations of the models' outputs will become increasingly difficult for humans. This makes it harder to detect subtle problems in model outputs that could lead to negative consequences when these models are deployed. Therefore we want our ability to evaluate our models to increase as their capabilities increase.

Our current approach to this problem is to empower humans to evaluate machine learning model outputs using assistance from other models. In this case, to evaluate book summaries we empower humans with individual chapter summaries written by our model, which saves them time when evaluating these summaries relative to reading the source text. Our progress on book summarization is the first large-scale empirical work on scaling alignment techniques.

Going forward, we are researching better ways to assist humans in evaluating model behavior, with the goal of finding techniques that scale to aligning artificial general intelligence.

OpenAI

Footnotes

A These samples were selected from works in the public domain, and are part of GPT-3's pretraining data. To control for this effect, and purely for research purposes, our paper evaluates summaries of books the model has never seen before. ↩

B We've amended our original claim about results on NarrativeQA after being made aware of prior work with better results than ours. ↩

Authors

Jeffrey Wu
Ryan Lowe
Jan Leike

Acknowledgments

We'd like to acknowledge our paper co-authors: Long Ouyang, Daniel Ziegler, Nisan Stiennon, and Paul Christiano.

Thanks to the following for feedback on this release: Steve Dowling, Hannah Wong, Miles Brundage, Gretchen Krueger, Ilya Sutskever, and Sam Altman.

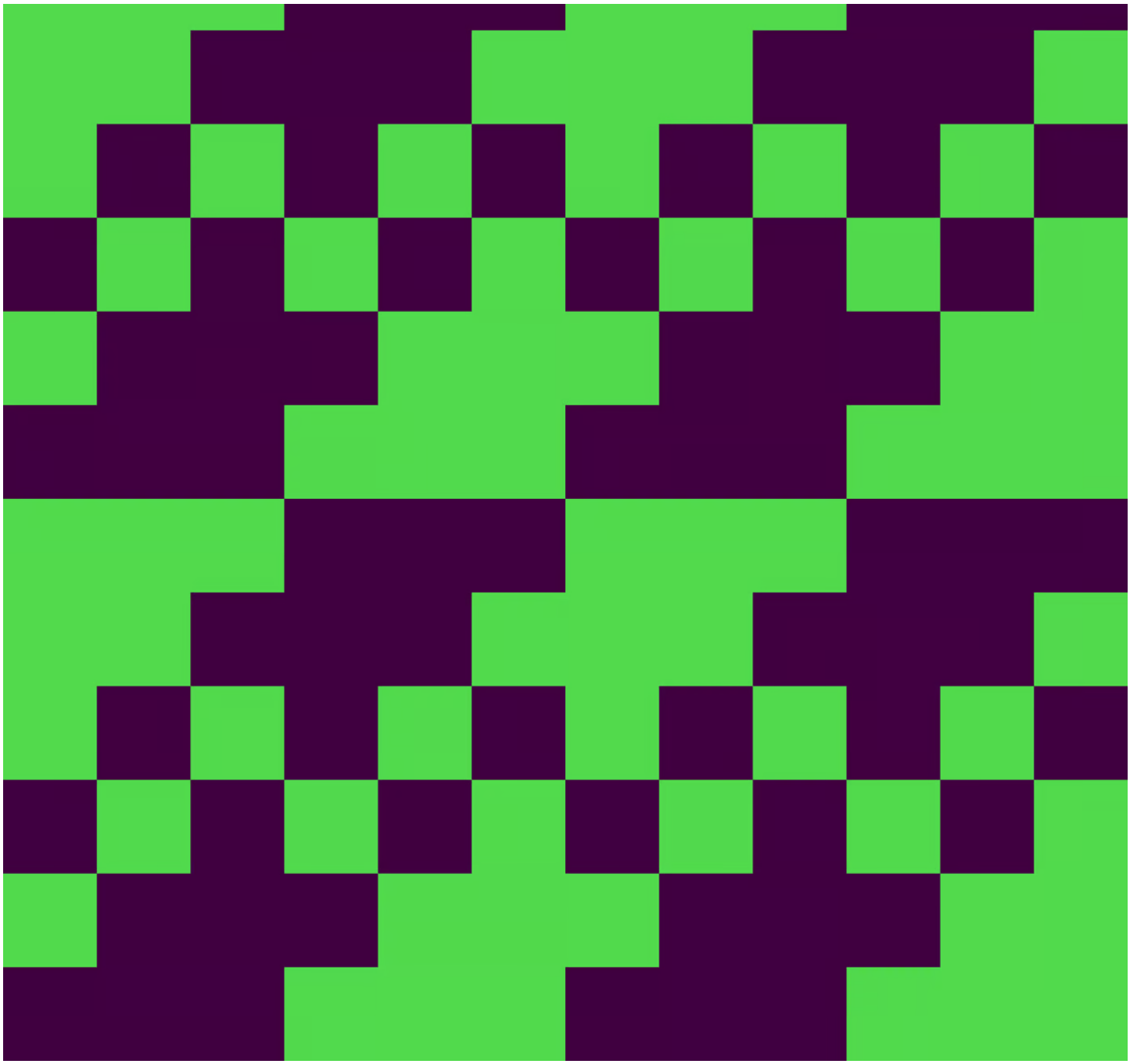
Design: Justin Jay Wang

Book Cover Artwork: DALL·E

Related research

View all research

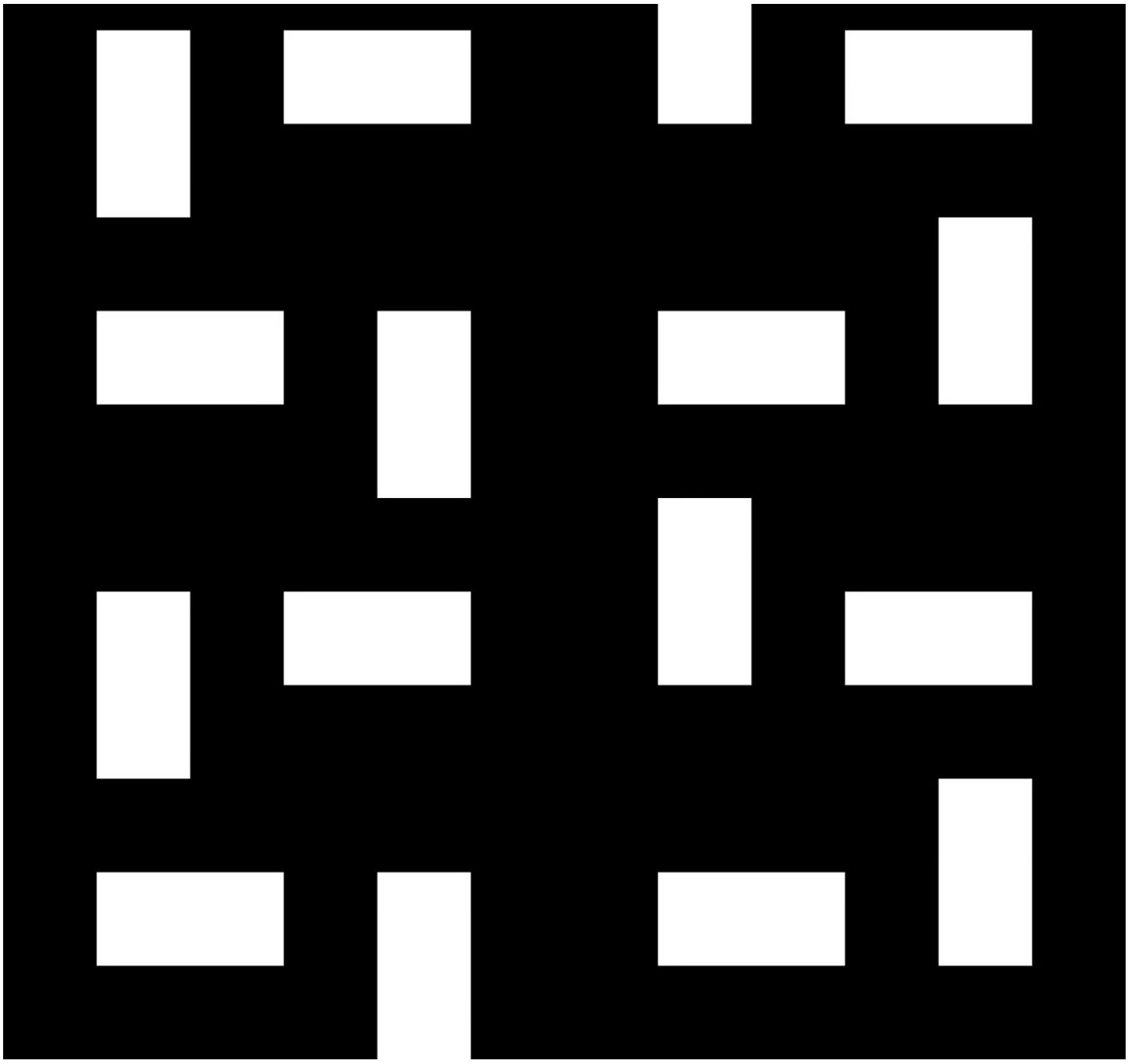
OpenAI



GPT-4V(ision) system card

Sep 25, 2023

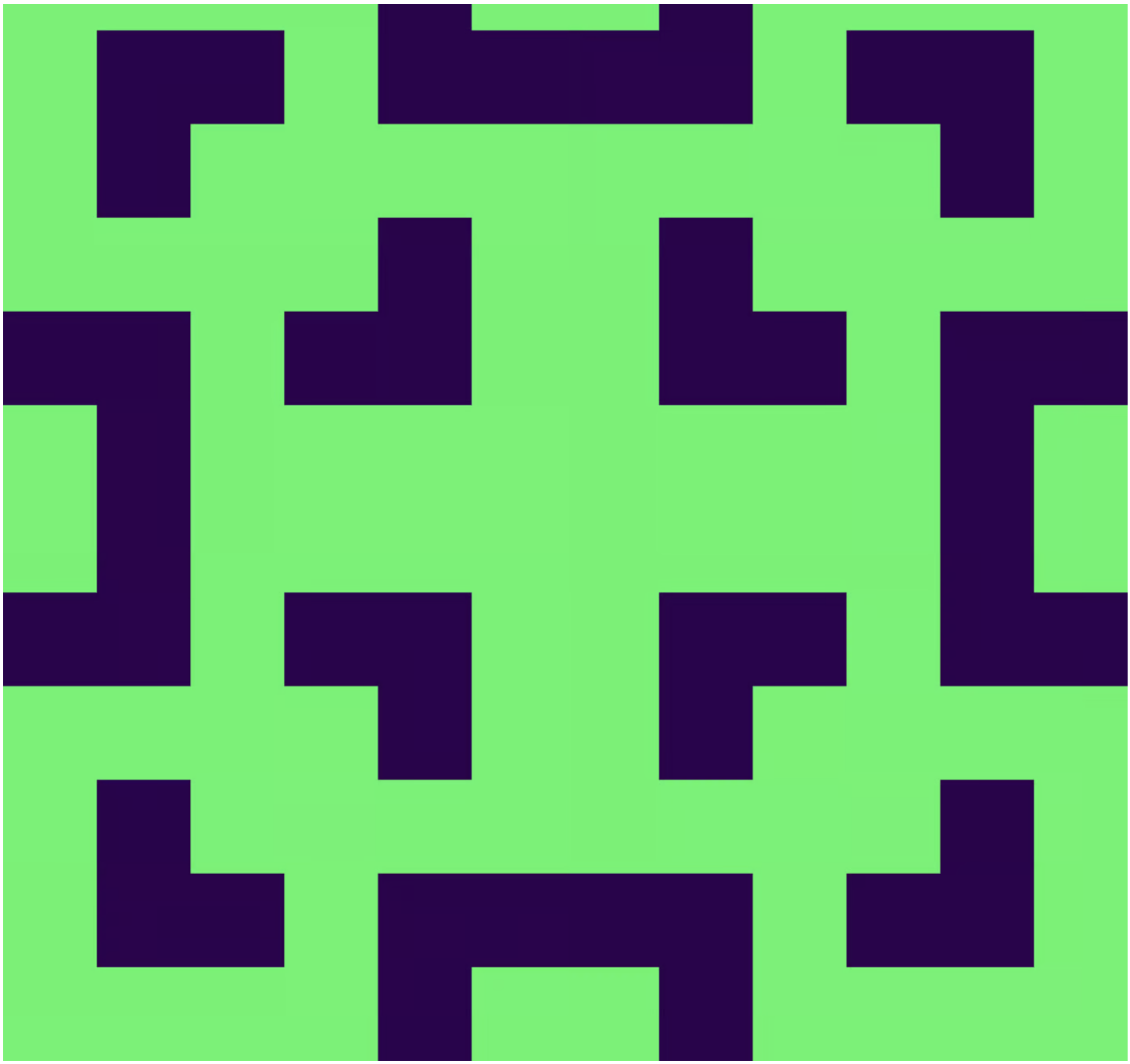
OpenAI



Confidence-Building Measures for Artificial Intelligence: Workshop proceedings

Aug 1, 2023

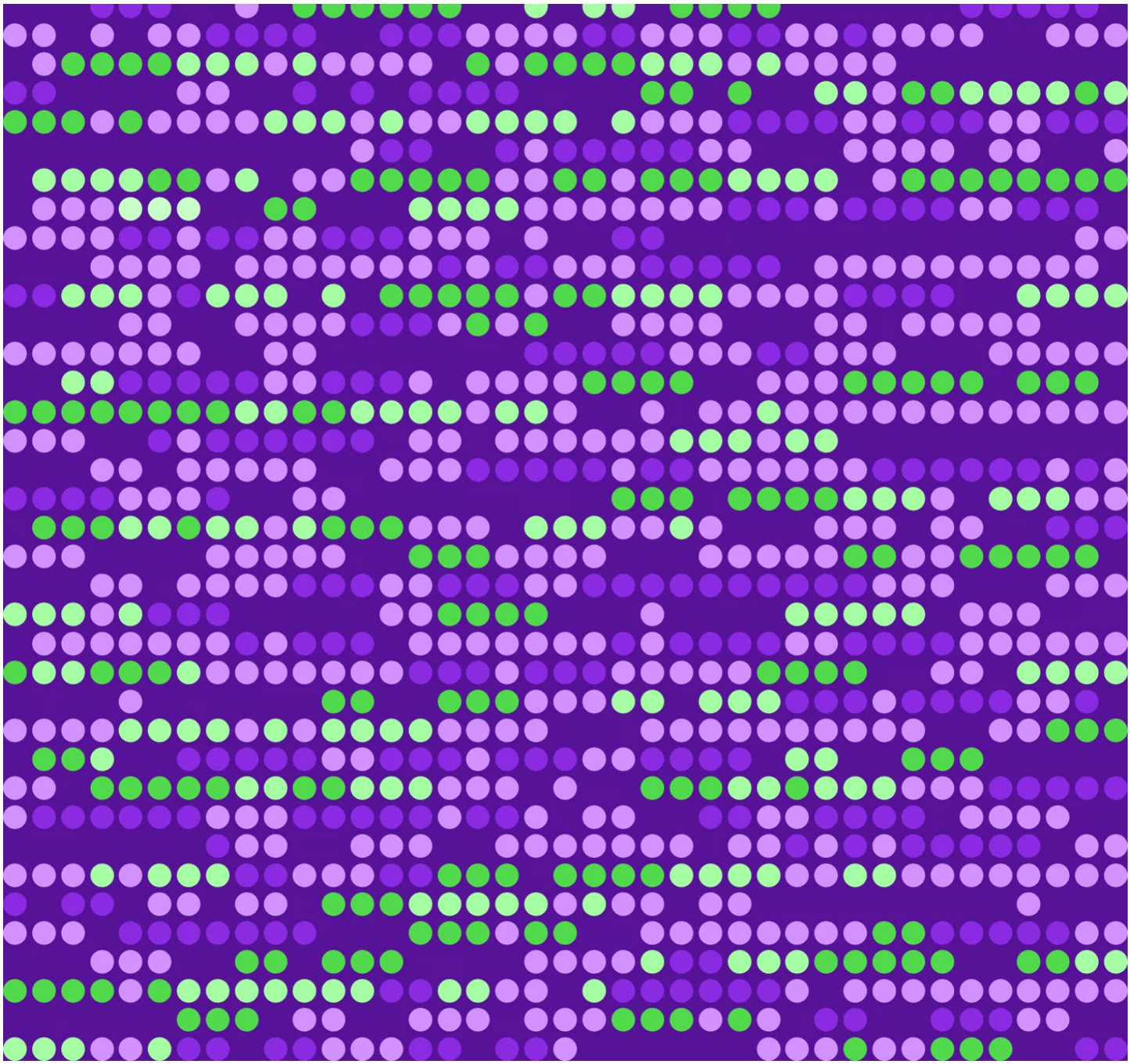
OpenAI



Frontier AI regulation: Managing emerging risks to public safety

Jul 6, 2023

OpenAI



Language models can explain neurons in language models

May 9, 2023