

Natural Language Generation for Creative Applications

Course: AI for Media, Art & Design
(A.k.a. “Intelligent Computational Media”)

Christian Guckelsberger
christian.guckelsberger@aalto.fi
 @creativeEndvs

A'

Aalto-yliopisto
Aalto-universitetet
Aalto University



Mario Klingemann
@quasimondo

...

These-movies-do-not-exist. Or do they? #thread #gpt3

A teenage girl and her father travel to California to stay with her dad's relatives, but they are actually horrific aliens in disguise.

A man is taught to fight zombies by a martial arts master in the Arizona desert.

11:06 PM · Jul 15, 2020 · Twitter Web App

18 Retweets 11 Quote Tweets 74 Likes



Mario Klingemann @quasimondo · Jul 15

...

Replying to @quasimondo

A man from the past is sent into the future to help his friend, but ends up leaving the future in ruins.



Mario Klingemann @quasimondo · Jul 15

...

A scientist invents a drug that brings out a person's innate ability to deal with any crisis.



Mario Klingemann @quasimondo · Jul 15

...

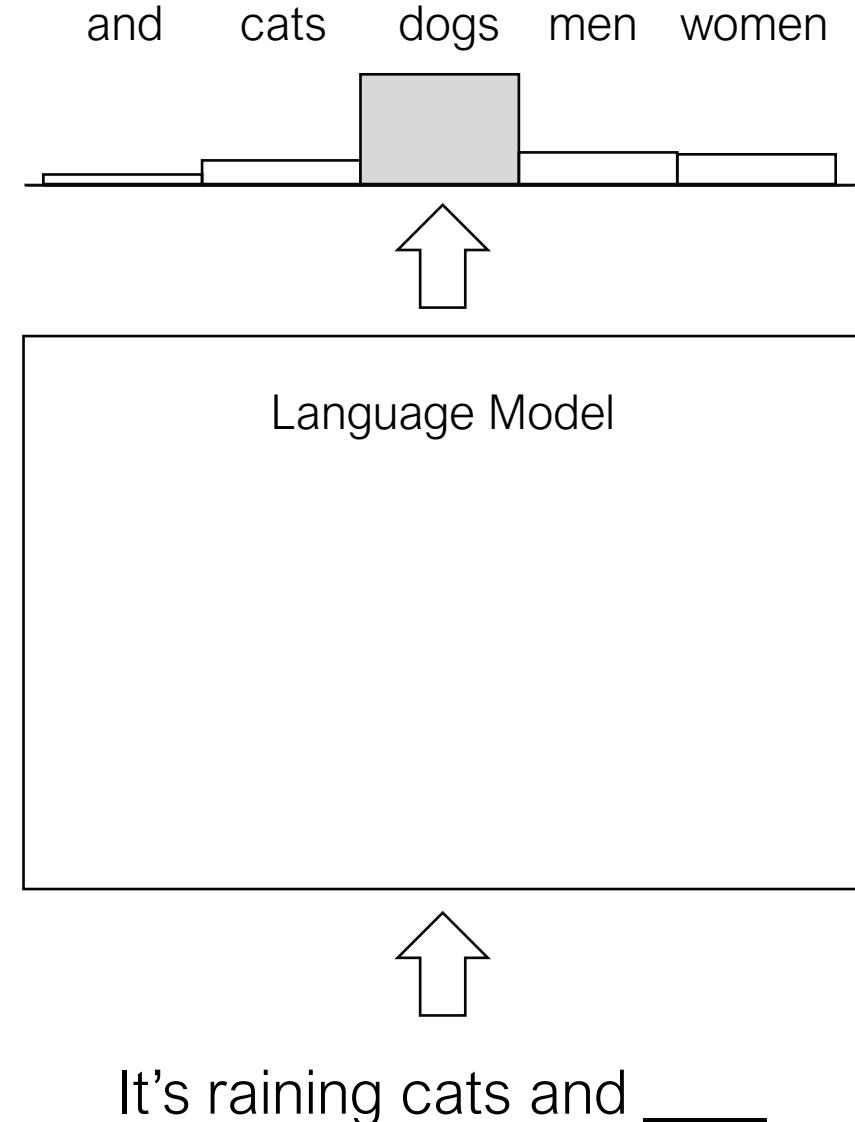
An alcoholic former sports star becomes a children's book author and

Contents

- What is natural language generation?
- Model evolution: attention is all you need!
- Hands-on: Text generation with GPT-3
- Showcase: Applications and risks
- Summary & further reading

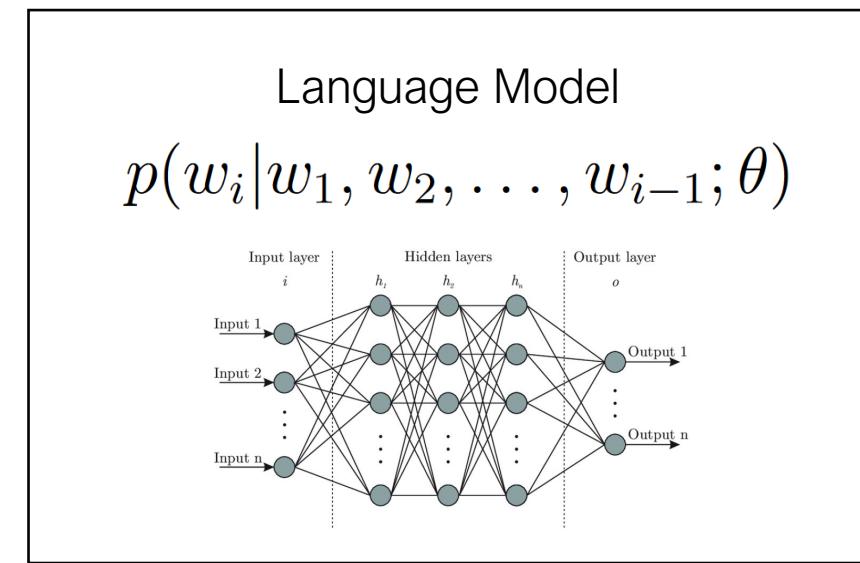
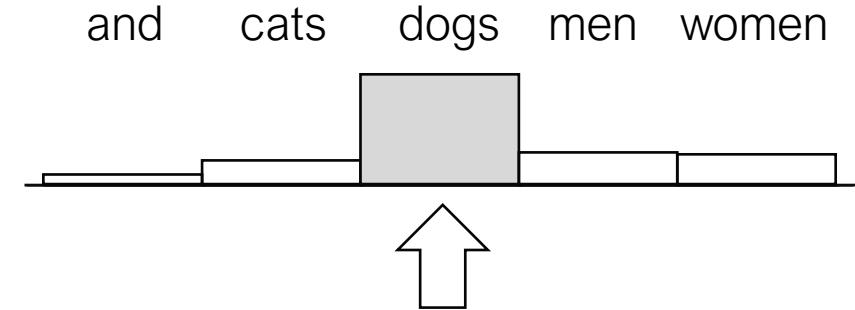
What is natural language generation?

- Informally: Given an (unfinished) sentence, tell me which word (token) is likely to come next. And then the word after...



What is natural language generation?

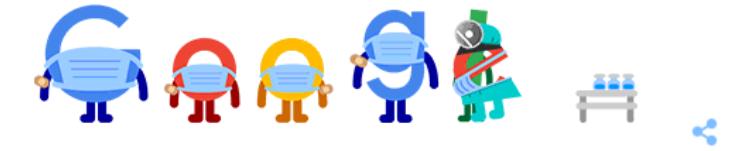
- **Informally:** Given an (unfinished) sentence, tell me which word (token) is likely to come next. And then the word after...
- **Formally:**
 - Conditional probability distribution:
 $p(w_i|w_1, w_2, \dots, w_{i-1}; \theta)$
 - Next word sampled from distribution:
 $w_i \sim p(w_i|w_1, w_2, \dots, w_{i-1}; \theta)$
 - Parameters θ e.g. weights of a neural net (-> lecture 2) learned from data.



It's raining cats and dogs
 $w_1 \quad w_2 \quad \dots \quad w_{i-1}$

What is natural language generation?

- Exercise: What use-cases of natural language generation can you think of?
 - Search query completion
 - ...



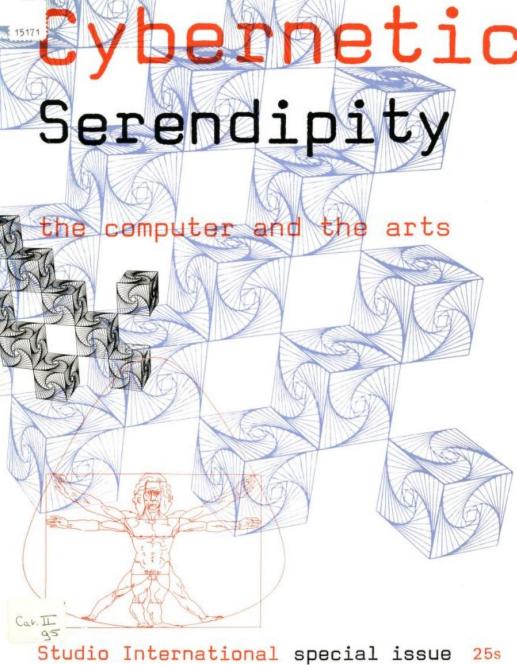
A screenshot of a Google search interface. The search bar contains the text "Use cases for |". Below the search bar is a list of suggested queries, each preceded by a magnifying glass icon:

- use cases for **nfts**
- use cases for **blockchain**
- use cases for **kafka**
- use cases for **atm system**
- use cases for **mongodb**
- use cases for **smart contracts**
- use cases for **nosql**
- use cases for **ethereum**
- use cases for **graph databases**
- use cases for **cryptocurrency**

At the bottom of the search interface are two buttons: "Google Search" and "I'm Feeling Lucky". A small link at the bottom right says "Report inappropriate predictions".

Model Evolution

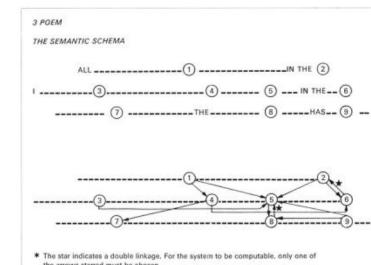
- For the model output to be **creative** (Runco & Jäger, 2012), it must be:
 - Novel:** not in training set
 - Value:** syntactically correct, meaningful, ... + domain-specific metrics.



Computerized Japanese haiku

These are examples, produced by on-line man-machine interaction at the Cambridge Language Research Unit, of one use of a computer for producing poetry. The programme is written in the TRAC language. The user can type in a formula, in which the operator types words. In '1 Poem' and '2 Poem', the operator chooses his words as he wishes. In the two '3 Poem' poems were produced by Margaret Masterman and Robin McKinnon Wood.

Slot 1 (→ 4) (→ 5)	Slot 2 (→ 5) (→ 6)	Slot 3 (→ 5)	Slot 4 (→ 6)	Slot 5 (→ 8)	Slot 6 (→ 7)	Slot 7	Slot 8 (→ 5) (→ 8)
White	Buds	See	Snow	Trees	Spring	Bang	Sun
Blue	Twigs	Trace	Tall	Full	Hush	Moon	Fit
Red	Leaves	Glimpse	Peaks	Pall	Swift	Star	Fled
Black	Hills	Flash	Dark	Hills	Cloud	Dimmed	
Green	Peak	Faint	Streams	Cold	Pffft	Cracked	
Snow	Snow	White	Heat	Specks	Flick	Streak	
Brown	Ice	Hear	Clear	Shade	Shoo	Shrunk	
Bright	Sun	Seize	Red	Arcs	Dawn	Smashed	
Pure	Rain		Blue	Grass	Shoo	Flower	Blown
Crowded	Cloud		Green	Dusk	Tree	Bud	Giving
Crowned	Sky		Grey	Cows	Day	Look	Crashed
Starred	Dawn		Black	Mist	Crash	Child	Gone
Fog	Mat		Deer	Trees	Crane	Crane	Fogged
Spring	Souls		Round	Stars	Woods	Bird	Burst
Heat	Clouds		Straight	Clouds	Hills	Plane	
Cold	Curved		Curved	Flowers	Pools	Moth	
	Slim		Slim	Buds			
	Burst		Burst	Leaves			
	Thin		Thin	Trees			
	Bright		Bright	Pools			
				Drops			
				Stones			
				Bells			
				Trails			



1 Poem
eons deep in the ice
I paint all time in a whorl
bamboo sludge has cracked
2 Poem
eons deep in the ice
I see gelled time in a whorl
pffft the cloud takes on the well known
mushroom shape
3 Poem
all green in the leaves
I smell dark pools in the trees
crash the moon has fled
3 Poem
all white in the buds
I flesh snow peaks in the spring
hang the sun has fogged

Computer poetry from CLRU

Robin McKinnon Wood and Margaret Masterman

Output resulting from a bug in the segmentation programme

The programme is designed to cut continuous strings of characters corresponding to the rhythmic divisions of speech or spoken prose. These units usually include two stress-points and a terminal intonation feature containing breath-groups which are also sense-groups.

The phonetic evidence and its relevance to silent reading was studied by David Shillan, the generalization of a two-stressed structure as a semantic form is the work of

Margaret Masterman. The segmentation programme, which in effect simulates human perception of these phrasings, is by John Dobson at the CLRU. This output was produced by the titan computer at Cambridge. In the case of this particular output, a bug in the programme effectively randomized the text given to it. We think this was caused by the program taking as its text the first word on each line of the original text, but we are not sure as we failed to make the programme do it again. At any rate this 'poem' is all the computer's own work!

Job Title (JED744/Phrasing) 10.7.67 Stream 1/0 (Phrasings) 1/0/1 I 1/2/6 a senator for a nineteen 1/2/11 provide to focus 1/2/12 in the need 1/2/13 some years 1/2/14 was is.

1/1/4 the people of them the 1/1/6 in the 1/1/7 these authorises services 1/1/8 for the deliberations 1/1/9 it kingdom associations 1/2/1 to the development people natural larger 1/2/2 by the fifty to the development 1/2/3 1/2/4 organisations cities, we welfare whole provided 1/2/5 1/2/6 a senator for a 1/2/7 for nineteen 1/2/8 provide to focus 1/2/9 in the need 1/2/10 some years 1/2/11 visiting there 1/2/12 homemaker 1/2/13 some years 1/2/14 was is.

Tape Mark I

into six lines of four metrical units each. The work of programming was on 322 punched cards, with 1,200 instructions. Much of the work is illustrated. The full experiment was performed on an IBM 7070 computer at the Electronic Centre of the Lombard Provinces Savings Bank in Milan in October 1961.

Nanni Balestrini

The basic text is made up of three extracts taken from:

(1) Maeda Hisashi's *Hiroshima diary*.

(2) Paul Goldwin's *The mystery of the elevator*: head pressed on shoulder hair between the lay motionless without shaking off his proved his fingers slowly trying to grasp.

(3) Lao Tzu's *Tao Te Ching*: while the multitude of things come into being I envy nothing return among things flourish they all return to their roots.

The working instructions for the computer are as follows:

(a) Make combinations of ten elements out of the given fifteen, without permutations or repetitions.

(b) Construct chains of elements taking account of the head-codes and end-codes.

(c) Avoid juxtaposing elements drawn from the same extract.

(d) Subdivide the chains of ten elements

Head pressed on shoulder, thirty times brighter than the sun I envisage their return, until he moved his fingers slowly and while the multitude of things comes into being, at the summit they all return to their roots and take on the well known mushroom shape endeavouring to grasp.

Hair between lips they all return to their roots, in the blinding fireball I envy their return, until he moves his fingers slowly and although things flourish on the well known mushroom shape endeavouring to grasp while the multitude

of things comes into being, head pressed on shoulder, thirty times brighter than the sun I envisage their return, until he moves his fingers

slowly and although things flourish on the well known mushroom shape endeavouring to grasp while the multitude of things comes into being, head pressed on shoulder, thirty times brighter than the sun I envisage their return, until he moves his fingers

slowly and although things flourish on the well known mushroom shape endeavouring to grasp while the multitude of things comes into being, head pressed on shoulder, thirty times brighter than the sun I envisage their return, until he moves his fingers

slowly and although things flourish on the well known mushroom shape endeavouring to grasp while the multitude of things comes into being, head pressed on shoulder, thirty times brighter than the sun I envisage their return, until he moves his fingers

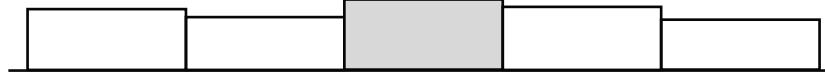
slowly and although things flourish on the well known mushroom shape endeavouring to grasp while the multitude of things comes into being, head pressed on shoulder, thirty times brighter than the sun I envisage their return, until he moves his fingers

slowly and although things flourish on the well known mushroom shape endeavouring to grasp while the multitude of things comes into being, head pressed on shoulder, thirty times brighter than the sun I envisage their return, until he moves his fingers

Model Evolution

- For the model output to be **creative** (Runco & Jäger, 2012), it must be:
 - **Novel**: not in training set
 - **Value**: syntactically correct, meaningful, ... + domain-specific metrics.
- We get **novelty** from sampling through recombination. But not necessarily value.

and cats dogs men women



Assumption: similar distribution irrespective of token

Input: It's raining ...

1st sampling:

It's raining dogs ...

It's raining and ...

It's raining women ...

2nd sampling:

It's raining dogs dogs ...

It's raining and women ...

It's raining women and ...

3rd sampling:

It's raining dogs dogs men ...

It's raining and women cats ...

It's raining women and and ...

Model Evolution

- For the model output to be **creative** (Runco & Jäger, 2012), it must be:
 - **Novel**: not in training set
 - **Value**: syntactically correct, meaningful, ... + domain-specific metrics.
- We get **novelty** from sampling through recombination. But not necessarily value.
- **Value**: learn better distributions via:
 1. better and more **training data**,

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
(2018) 355M parameters

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language

{jacobdevlin, mingweichang, kentonl, kristout}@google.com

Language Models are Unsupervised Multitask Learners

Alec Radford *[†] Jeffrey Wu *[†] Rewon Child[†] David Luan[†] Dario Amodei **[†] Ilya Sutskever **[†]
(2019) 1.5B parameters

Abstract

Natural language processing tasks, such as ques-

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training

Journal of Machine Learning Research 21 (2020) 1-67

Submitted 1/20; Revised 6/20; Published 6/20

(2019/20) 11B parameters

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel*

CRAFFEL@GMAIL.COM

Language Models are Few-Shot Learners

(2020) 175B parameters

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

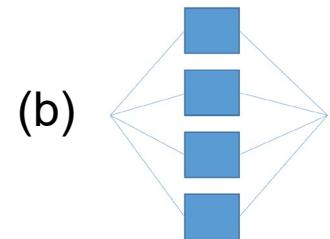
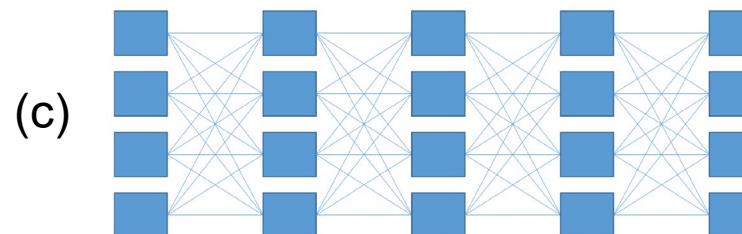
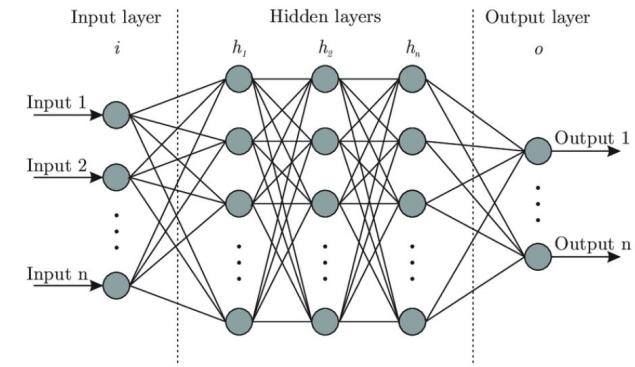
Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Model Evolution

- For the model output to be **creative** (Runco & Jäger, 2012), it must be:
 - **Novel**: not in training set
 - **Value**: syntactically correct, meaningful, ... + domain-specific metrics.
- We get **novelty** from sampling through recombination. But not necessarily value.
- **Value**: learn better distributions via:
 1. better and more **training data**,
 2. increased **model complexity**,

Exercise: (1) How many paths can input data take through networks (a)-(c)?
(2) How many nodes are there for a given width and depth?



Model Evolution

- For the model output to be **creative** (Runco & Jäger, 2012), it must be:
 - **Novel**: not in training set
 - **Value**: syntactically correct, meaningful, ... + domain-specific metrics.
- We get **novelty** from sampling through recombination. But not necessarily value.
- **Value**: learn better distributions via:
 1. better and more **training data**,
 2. increased **model complexity**,
 3. by incorporating (more) **context**!

“**Context**”: how many of the preceding words to take into consideration when sampling the next?

$$w_i \sim p(w_i | \underbrace{w_1, w_2, \dots, w_{i-1}}_{\text{context}}; \theta)$$

Why does it matter? Complete the following sentence:
“The animal didn't cross the street because it ...”

“The animal didn't cross the street because it is wet”

“The animal didn't cross the street because it is tired”

Required context size: 2 vs. 6 words!

Existing architectures incorporate context to various degrees: Markov chains, Recurrent Neural Networks (RNNs), Long-Short-Term Memory (LSTMs), ...

Model evolution: larger contexts, less forgetting!

Transformers: Attention is All You Need

- Architecture revolutionised natural language processing (Vaswani et al., 2017)

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [2] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

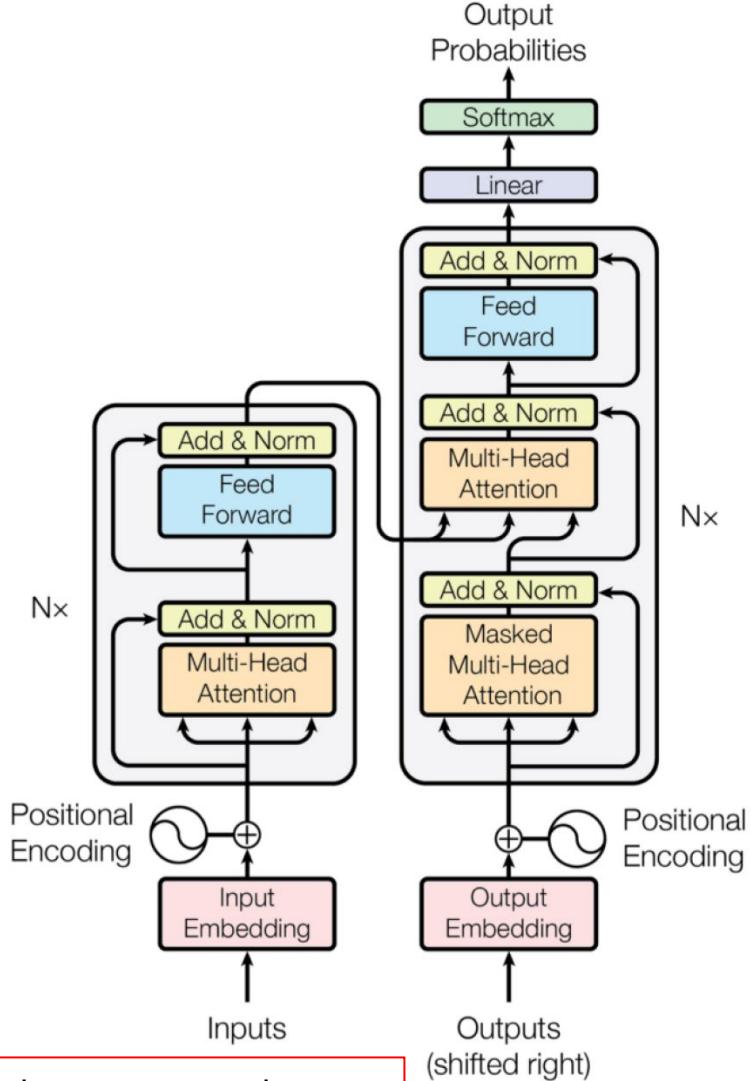
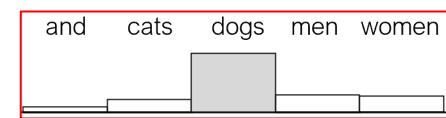
*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

Transformers: Attention is All You Need

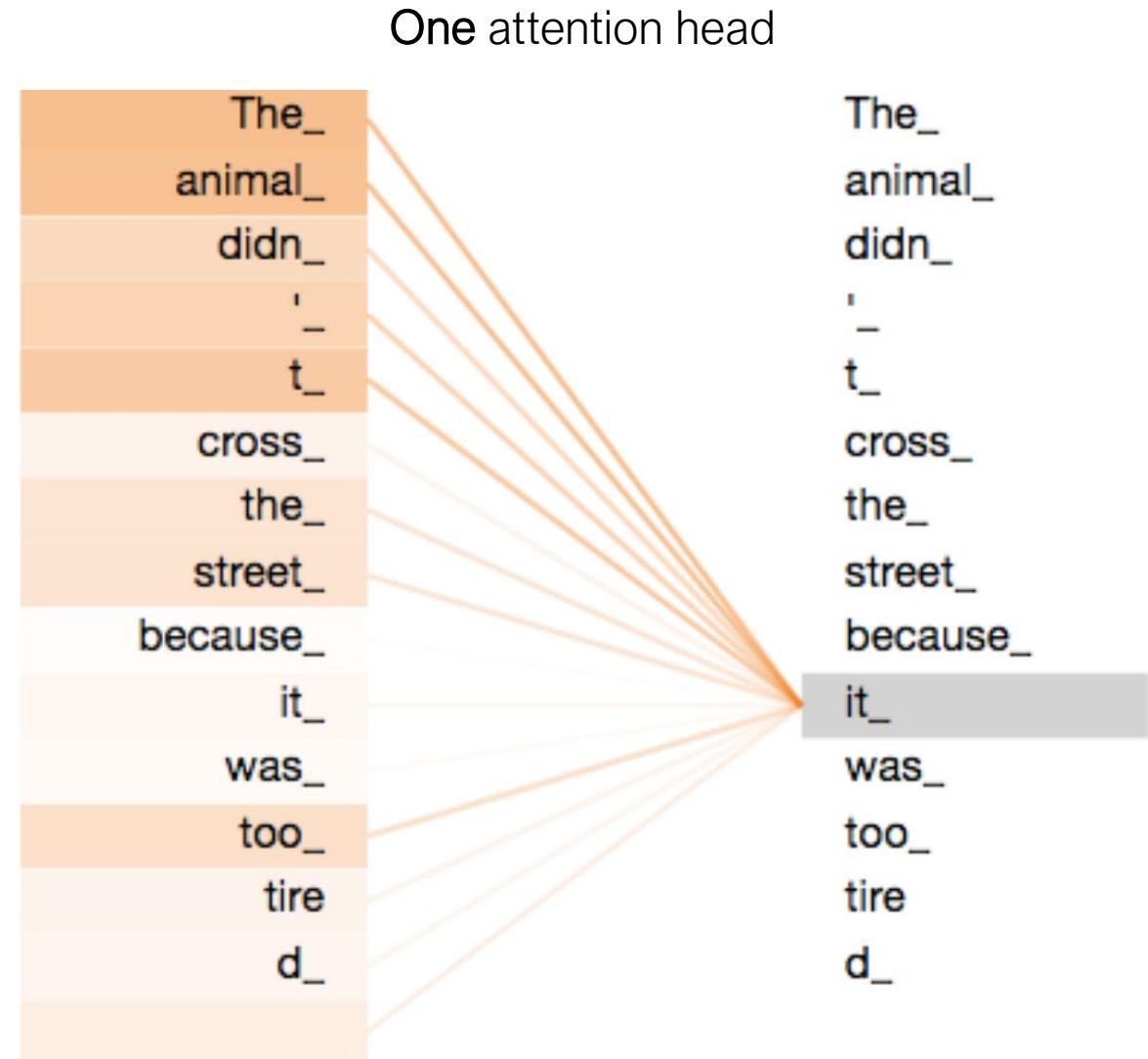
- Architecture **revolutionised** natural language processing (Vaswani et al., 2017)
- Most important **components**:
 - Feed forward networks (**blue**, -> lecture 2)
 - New element: self-attention (**orange**).



It's raining cats and _____

Transformers: Attention is All You Need

- Architecture **revolutionised** natural language processing (Vaswani et al., 2017)
- Most important **components**:
 - Feed forward networks (**blue**, -> lecture 2)
 - New element: self-attention (**orange**).
- “**Self-attention**” here: learns to look at other tokens in input to better predict word.
- Transformer has 8 attention heads.
Wider context, less forgetting!



Transformer self-attention
(Illustration by Jay Alammar:
<https://jalammar.github.io/illustrated-transformer/>)

Generative Pre-Trained Transformers (GPT)

- OpenAI's large Transformers, unsupervised training on huge, highly curated datasets.
- **Pre-training:** high performance on various tasks with little/no task-specific fine-tuning.
- **GPT-2** (Radford et al., 2019):
 - 1.5B parameters.
 - Available for download & fine-tuning.
- **GPT-3** (Brown et al., 2020):
 - 175B parameters
 - accessible only through OpenAI's web interface / API.

Radford, Wu, Child, Luan, Amodei & Sutskever. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), p.9.

Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell & Agarwal. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Language Models are Unsupervised Multitask Learners

Alec Radford ^{* 1} Jeffrey Wu ^{* 1} Rewon Child ¹ David Luan ¹ Dario Amodei ^{** 1} Ilya Sutskever ^{* 1}

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al., 2017), reading comprehension systems (Jia & Liang, 2017),

Language Models are Few-Shot Learners

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan [†]	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess		Jack Clark	Christopher Berner	
Sam McCandlish		Alec Radford	Ilya Sutskever	
		Dario Amodei		

OpenAI

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task agnostic

Hands-On: Text Generation With GPT-3

- **Exercise (15 mins):**
 - Register on OpenAI's GPT3 Playground
 - Co-create a videogame plot with GPT-3 as "imagination engine". Provide prompts for the model to complete.
 - Consider: how do different modifications of the input prompt affect the output?
- **Interaction:** Found something cool?
Share the prompt + output (mark which is which) on Discord!
- **Two strategies for co-creation:**
 - Prompt only (zero-shot learning)
 - Several examples (few-shot learning)

The screenshot shows the OpenAI GPT-3 Playground interface. At the top right, there are links for Help, Aalto University, and a menu icon. Below that, the word "Playground" is displayed next to a "Load a preset..." dropdown and several action buttons: Save, View code, Share, and settings. The main area contains three generated text snippets separated by "===".
1. Let me tell you about the indiegame that I developed over the past two years. It's a strategy game on the theme of "lost and found". The player swings a lasso to catch and store animals, which he can then trade for score. The player himself is a small creature which can be found in one of the levels. If you are able to capture him, you get a bonus. The game also has a level editor, so you can make your own levels and challenge your friends. The game ends when the player has caught all animals.
2. Let me tell you about the indiegame that I developed over the past two years. It's a strategy game on the theme of "lost and found". The player is lost in space and hopes to be found by the crew of his lost ship. To this end, he must follow a signal and navigate through space debris. The debris can injure the player, but they can recharge energy through the sun. The game ends when the player finds the rescue ship.
3. Let me tell you about the indiegame that I developed over the past two years. It's a strategy game on the theme of "lost and found". The player is on a strange island, which is inhabited by animals. The player is able to communicate with them and trade with them. However, if the player attempts to communicate with an animal that he hasn't traded with, the animal will attack him. The game ends when the player trades with all animals.
At the bottom left, there is a green "Generate" button and three circular icons for undo, redo, and refresh.

OpenAI Gpt-3 Playground:
<https://beta.openai.com/playground>

Debriefing: How can we shape the output?

- Exercise: Tell us what you've found!
- ...

The screenshot shows the OpenAI GPT-3 Playground interface. At the top right, there are links for Help, Aalto University, and a menu icon. Below that, the word "Playground" is displayed next to a "Load a preset..." dropdown and several action buttons: Save, View code, Share, and three more options represented by ellipses and a gear icon.

In the main area, a question is entered into the input field: "How can I get something meaningful out of you?". Below the input, the generated response is shown in a large text block:

This is a difficult question. It depends upon what you are looking for. If you are looking for someone to talk to, I am a good listener. If you are looking for someone to help you with a problem, I may be able to help. If you are looking for someone to do something for you, I may be able to help or I may be able to recommend someone who can.

At the bottom of the playground window, there are three buttons: "Generate" (in green), a refresh icon, and another refresh icon.

OpenAI Gpt-3 Playground:
<https://beta.openai.com/playground>

Third Strategy: Fine-Tuning GPT

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Susanna Värtinen

Generating Role-Playing Game Quest Descriptions With the GPT-2 Language Model

Master's Thesis
Espoo, December 10th, 2021

Supervisor: Professor Perttu Hämäläinen, Aalto University
Advisor: Doctor Christian Guckelsberger

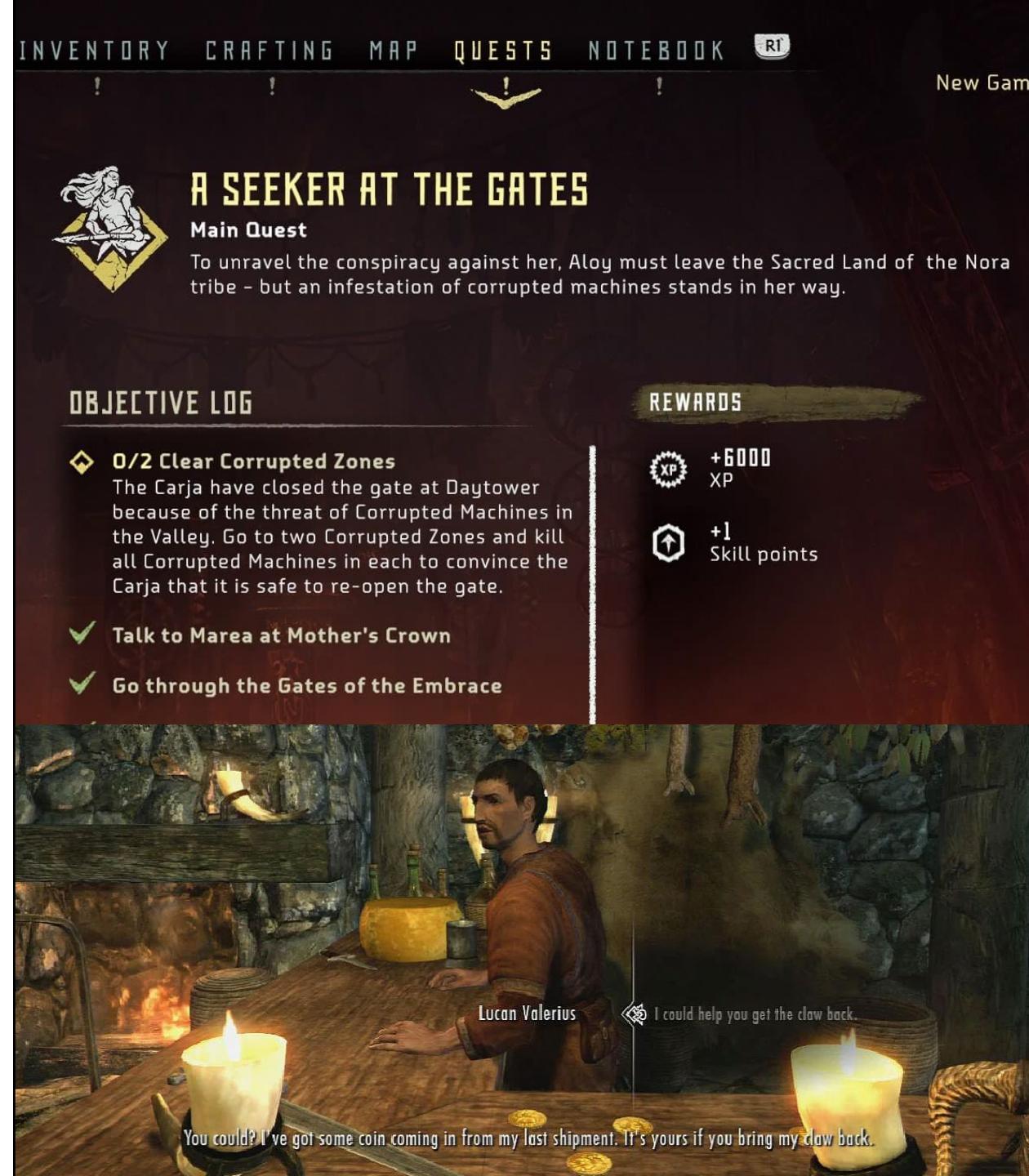
I am Officer Vai, a member of the Flaming Fist. I'm here to talk about something that has been plaguing us for quite some time now. We've had reports of bandits operating in the area around Sword Coast for quite some time now. They seem to be targeting travelers and merchants alike. We're not sure who they work for, but we know it isn't good. It seems as though these bandits are working for someone else. If you could track down the person or people behind this, we would reward you with 50 gold pieces per bandit scalp.

I am Jauffre, military leader here at Cloud Ruler. I've been hearing rumors about strange people coming into our town. We need to find these people before someone else does! We don't want any trouble with the Countess of Bruma, but we do need to get rid of these people. You're going to help us track down these people. You'll go to Cloud Ruler and speak to Steffan, one of the gate guards. He'll tell you what he knows. Then you'll head over to Bruma and speak to Captain Burd, the captain of the guard. Tell him what you found. If there's anything suspicious, let me know.

Case-Study: Generating Role-Playing Game Quests.
Source: Master Thesis Susanna Värtinen (Aalto)

Fine-Tuning Case Study: Generating RPG Quests

- **Quests** in Role-Playing Games (RPGs): Challenging tasks for the player.
- Focus: **quest descriptions**. Short texts that explain the quest to the player from the perspective of a quest-giving character
- **Motivation**: growing demand for more game content and support continuous player engagement.
- **State-of-the-art**: procedurally generate quests to satisfy demand; but: existing approaches formulaic and repetitive.
- **Idea**: GPT language models to the rescue?



Fine-Tuning Case Study: Generating RPG Quests

- Method: fine-tuning, i.e. re-train model on smaller dataset for better results.
- GPT-3 unsuitable:
 - GPT-3 not available for fine-tuning at beginning of project
 - GPT-3 too big to be trained / sampled from on game dev / player hardware. Thus not standalone but requires OpenAI API.
 - Fine-tuning GPT-3 is expensive.
- Thus fine-tuning GPT-2 (1.5B parameters).

Fine-tuning

Learn how to customize a model for your application.

Introduction

Fine-tuning lets you get more out of the models available through the API by providing:

- 1 Higher quality results than prompt design
- 2 Ability to train on more examples than can fit in a prompt
- 3 Token savings due to shorter prompts
- 4 Lower latency requests

GPT-3 has been pre-trained on a vast amount of text from the open internet. When given a prompt with just a few examples, it can often intuit what task you are trying to perform and generate a plausible completion. This is often called "few-shot learning."

Pricing

When training a fine-tuned model, the total tokens used will be billed at 50% the [base model inference rate](#) per 1k tokens. The total tokens used to fine-tune are the tokens in your [training file](#) multiplied by the number of [epochs](#). A 1M token file is roughly 5MB and fine-tunes default to 4 epochs.

MODEL	PRICE PER 1K TOKENS	1M TOKEN FILE TRAINED FOR 4 EPOCHS
Fine-tuned Davinci <small>Beta</small>	\$0.03	\$120
Fine-tuned Curie	\$0.003	\$12.00
Fine-tuned Babbage	\$0.0006	\$2.40
Fine-tuned Ada	\$0.0004	\$1.60

As a reminder, pricing is subject to OpenAI's [terms of service](#), which prohibit you from using the API to generate content that is subject to OpenAI's [content policies](#) or otherwise violates them. You are responsible for ensuring that your use of the API complies with these policies.

MODEL	PRICE PER 1K TOKENS
Fine-tuned Davinci <small>Beta</small>	\$0.12
Fine-tuned Curie	\$0.012
Fine-tuned Babbage	\$0.0024
Fine-tuned Ada	\$0.0016

Fine-tuned Davinci is a beta feature. It is ready for production use, but rates and features are subject to change.

Fine-Tuning Case Study: Generating RPG Quests

- Method: fine-tuning, i.e. re-train model on smaller dataset for better results.
- GPT-3 unsuitable:
 - GPT-3 not available for fine-tuning at beginning of project
 - GPT-3 too big to be trained / sampled from on game dev / player hardware. Thus not standalone but requires OpenAI API.
 - Fine-tuning GPT-3 is expensive.
- Thus fine-tuning GPT-2 (1.5B parameters).
- Custom-made dataset, publicly available:
 - 978 quests and descriptions
 - From 6 RPG games

The screenshot shows a GitHub repository page for 'rpg-quest-data-set'. At the top, it displays the repository name 'svartinen / rpg-quest-data-set' (Public), with options to Watch (1), Fork (0), Star (0), and more. Below the header, there are tabs for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, and more. A dropdown menu shows 'main' selected. To the right, there's a sidebar with sections for About, Releases, and Packages, all currently empty. The main content area shows a list of recent commits:

Author	Commit Message	Date	Commits
svartinen	Update README.md	on 7 Dec 2021	4
BG	add quests in generic JSON form...	2 months ago	
MC	add quests in generic JSON form...	2 months ago	
TES	add quests in generic JSON form...	2 months ago	
TL2	add quests in generic JSON form...	2 months ago	
train_text	add training data that was used i...	2 months ago	
README.md	Update README.md	2 months ago	

Below the commits, the README.md file is shown with the title 'rpg-quest-data-set'. It describes the dataset as a collection of quests from Baldur's Gate, Baldur's Gate II, The Elder Scrolls IV: Oblivion, The Elder Scrolls V: Skyrim, Minecraft, and Torchlight II. The quests from BG, BGII, TES IV, and TES V were pieced together from dialogue files extracted with modding tools. The Minecraft quests were written manually by us. The quests from TLII were originally by Stegeren and Theune ([repo](#)) and edited by us.

Table I
THE QUEST DATA SET (978 QUESTS)

Game	Sourcing	Quests
Baldur's Gate [25]	collected (game files)	100
Baldur's Gate II: Shadows of Amn [26]	collected (game files)	94
The Elder Scrolls IV: Oblivion [27]	collected (game files)	215
The Elder Scrolls V: Skyrim [28]	collected (game files)	389
Minecraft [29]	written by the authors	100
Torchlight II [30]	collected [31]	80

Fine-Tuning Case Study: Generating RPG Quests

- **Evaluation:** online questionnaire
 - 500 quest descriptions generated from 50 quests (test set) from 6 RPGs.
 - 349 participants, each rating 25 quest descriptions (5 descriptions x 5 quests)
 - 4-point Likert scale (Strongly Disagree – Strongly Agree): “I would be happy to see this quest description in a video game”.
 - Recruited from RPG sub-reddits.

Research Study: Generating Video Game Quest Descriptions With Natural Language Models

 Mandatory fields are marked with an asterisk (*) and must be filled in to complete the form.

Below you find one hand-authored role-playing game quest, and five corresponding, AI-generated quest descriptions. Please read the quest and rate how happy you would be to see each of the quest descriptions in a videogame.

A note about the questions: some mobile browsers might show questions titled "Show Quest" due to a bug in the survey platform. You should ignore those questions: they are meant to control the presentation logic of the study.

Quest 5: Vai's Bounty Upon Bandits

Objective: find out why the bandit raids have been getting worse

Tasks: kill all bandits

Task locations: Sword Coast, a region

Quest-giver: Officer Vai, an officer of the Flaming Fist, located at the Sword Coast

Rewards: 50 gold pieces per bandit scalp

Groups: Flaming Fist, a mercenary organization promoting order and law

Facts:

the bandits are working for someone

Quest Description 3 (Quest: Vai's Bounty Upon Bandits) *

I am Officer Vai, an officer of the Flaming Fist. I was sent here to bring order to these lands. I've heard that the bandits are working for someone else. If you're willing to help us with our investigation, we'll reward you well. You must go into the Sword Coast and kill every single one of them. We'll give you 50 gold per bandit scalp.

I would be happy to see this quest description in a video game.

1	2	3	4
Strongly Disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly Agree			

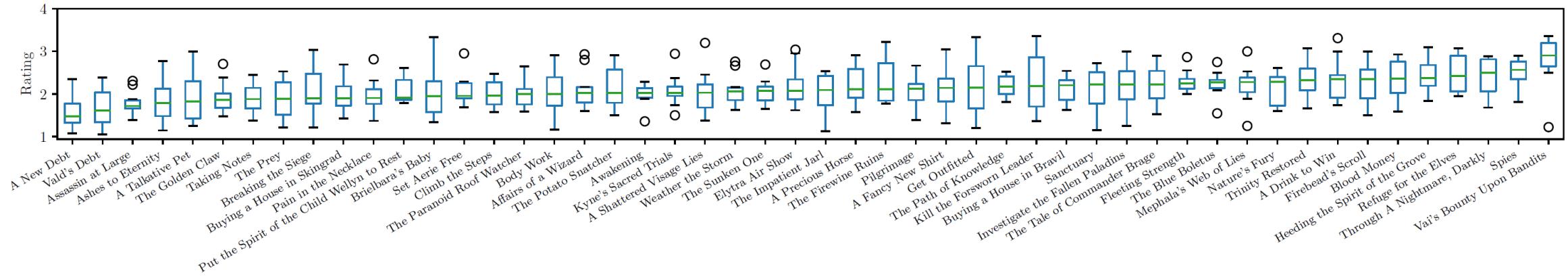
Quest Description 1 (Quest: Vai's Bounty Upon Bandits) *

Greetings! I am Officer Vai, an officer of the Flaming Fist. We have received word that the bandits have been making more attacks than usual. I would like you to go into the Sword Coast and see what we can do about it. I'll pay you 50 gold per bandit scalp.

I would be happy to see this quest description in a video game.

1	2	3	4
Strongly Disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly Agree			

Fine-Tuning Case Study: Generating RPG Quests



- Findings:**
 - Strong variation in perceived quest quality.
 - Curation coefficient (rating >2.5): 0.22.
I.e. 1 in 5 quests descriptions acceptable.
 - Participants generally more critical the more they played.
Exception: “hard-core” players (>16 hours / week).
- Further findings** on participant’s rating criteria, best prompt formatting, etc. (ask for the paper if interested)

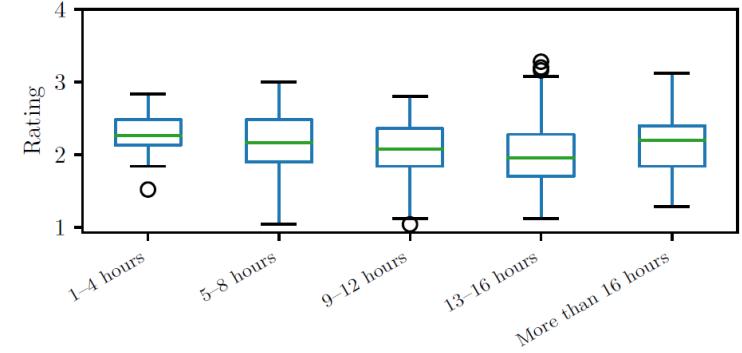


Figure 9. Box plots of averaged ratings per participant, grouped by their average weekly playtime (groups holding <5% participants were omitted).

Fine-Tuning Case Study: Generating RPG Quests

- Conclusion: fine-tuning largest GPT-2 variant does not allow for autonomous generation of high-quality quest descriptions. Curation still indispensable!
- Use-cases:
 - Offline generation + curation
 - Assistant for co-creative quest writing
 - Quest idea generator

I am Mogrul, a money-lending merchant. I'm not sure if you've heard of me before, but I'm one of those people who can lend a hand to anyone in need. My name is Mogrul. You might know me as Mogrul, or maybe as Drovax Relvi. I was hired by Neloth to help him out with a problem he had. It turned into quite the headache for both of us. You're going to help me deal with it. I'll give you something called 1,000 septims. This is a new form of currency that we're trying out here at my location. We don't have much of it yet, so I'm giving you a small amount to help you get started. If you ever need anything else, just ask.

I am Officer Vai, a member of the Flaming Fist. I'm here to talk about something that has been plaguing us for quite some time now. We've had reports of bandits operating in the area around Sword Coast for quite some time now. They seem to be targeting travelers and merchants alike. We're not sure who they work for, but we know it isn't good. It seems as though these bandits are working for someone else. If you could track down the person or people behind this, we would reward you with 50 gold pieces per bandit scalp.

I'm not sure if you're aware of it, but I've been looking for someone to help me out with my current situation. My name is Vex. You might know me as Vald. I was once your lover, until we were betrayed by our mutual friend Mercer Frey. We had a good run together, but now he has turned on us both. He owes us money, so I need to get his debtors off our backs. If you could speak to Maven Black-Briar about getting rid of Vald's debts, that would be great. If you can't do that, then I'll have to kill him myself.

I am Jauffre, military leader here at Cloud Ruler. I've been hearing rumors about strange people coming into our town. We need to find these people before someone else does! We don't want any trouble with the Countess of Bruma, but we do need to get rid of these people. You're going to help us track down these people. You'll go to Cloud Ruler and speak to Steffan, one of the gate guards. He'll tell you what he knows. Then you'll head over to Bruma and speak to Captain Burd, the captain of the guard. Tell him what you found. If there's anything suspicious, let me know.

I'm sorry... I was just having a bad day. I don't know what happened. It all started when someone tried to kill me. It was a Thalmor. They were after my job and they thought I was one of them. I had to fight them off but it took too long. I couldn't get away fast enough. I ran into a Dunmer named the Khajiit. He wanted to kill me so he came after me. He was going to kill me anyway, but now he doesn't want to. He says if I go back to Windhelm, I'll die. I told him I'd stay out of Morrowind. I didn't tell him where I was going. If you ever need anything, come see me.

I've got a problem here in Arcane University. A student named Henantier. He was studying with me last semester, but he got himself into too much trouble. I'm not sure what went wrong, but it seems like he's been doing some very dangerous things. He doesn't seem to care about anything anymore. He just keeps getting into more and more trouble. It's all my fault. I told him to stop messing around with magic, but he didn't listen. Now he's stuck in a place called Leyawiin. There's nothing I can do for him. If you could get Henantier out of there, I'd give you a sizeable reward.

(a) The worst rated quest descriptions for the three worst rated quests: "A New Debt", "Vald's Debt", and "Assassin at Large".

(b) The best rated quest descriptions for the three best rated quests: "Vai's Bounty Upon Bandits", "Spies", and "Through A Nightmare, Darkly".

Fine-Tuning Case Study: Generating RPG Quests

- Conclusion: fine-tuning largest GPT-2 variant does not allow for autonomous generation of high-quality quest descriptions. Curation still indispensable!
- Use-cases:
 - Offline generation + curation
 - Assistant for co-creative quest writing
 - Quest idea generator
- Future work / silver lining: non-fine-tuned GPT-3 produces more coherent output for worst-rated quests in our experiments.

This is an RPG quest from The Elder Scrolls V: Skyrim.
The quest-giver is called Mogrul. Mogrul is a money-lender.
The quest-giver gives a quest to the player. The player's objective is to help Drovas Relvi with his debt to Mogrul.
The player should first pay Drovas Relvi's debt to complete their objective.
The player will receive the following rewards for completing the quest objective: Mogrul will stop sending his debt collectors after the player.
The following characters are related to this quest: Drovas Relvi (a man the player recruited for Neloth, and thus became untouchable), Neloth (a powerful male wizard).
The following items are related to this quest: 1,000 septims (a currency).
There are some important facts concerning this quest. Drovas Relvi owes Mogrul 1,000 septims: Mogrul thinks that the player should pay Drovas Relvi's debt.
This is the quest description, the quest-giver explaining the quest to the player:

I hear you've been helping out Neloth. I'm not surprised. Neloth isn't very good at repaying his debts. I hear you helped out Drovas Relvi, too. Drovas Relvi owes me 1,000 septims. I'd like for you to pay his debt. Then I'll stop sending my debt collectors after you.

I'm sure you're wondering what you're doing here. Well, I need your help. Drovas Relvi owes me 1,000 septims, but I can't go after him myself. He works for Neloth, you see. If I attack him, Neloth will know.
I'm sure you can understand. I'm willing to pay you 500 gold to help me. I will follow you to Relvi's house. You just have to deliver the money.
My associate will take care of the rest.

Fellow, what's the word on the street? Drovas Relvi owes me 1,000 septims. That can't be good for his health. Has he talked to you?
I don't care what you get him to do. Just get my money.

There's a fellow in town who owes me a fair bit of coin. Drovas Relvi. I'd like to collect, but he's Neloth's friend and I don't want to make trouble.
I'm trying to help him out, but I'm not sure how. If you could talk to him, maybe you could convince him to pay me back.
If you do that, I'll make sure my collectors don't bother you again.

(a) Input quest outline in the *narrative* format with *raw text*

(b) Four random output quest descriptions generated with GPT-3

Figure 11. A quest generation demo with the quest "A New Debt" and GPT-3 (OpenAI API Playground, default text generation settings apart from response length of 700). The quests "Ashes to Eternity", "Assassin at Large", and "Vald's Debt" were given as examples to GPT-3 beforehand.

Applications and Risks



GPT-3 CREATIVE FICTION

SITE
ME

CHANGES
NEWS

SUPPORT ON
PATREON

Creative writing by OpenAI's GPT-3 model, demonstrating poetry, dialogue, puns, literary parodies, and storytelling. Plus advice on effective GPT-3 prompt programming & avoiding common errors.

[NN](#), [fiction](#), [GPT](#), [poetry](#), [humor](#), [transhumanism](#)

2020-06-19-2021-07-01 · finished · [certainty: likely](#) · [importance: 8](#) · [backlinks](#) / [bibliography](#)

- 1 What Benchmarks Miss:
Demos
- 2 GPT-3 Implications
- 3 Quality
- 4 Prompts As Programming
 - 4.1 Finetuning
 - 4.2 Playground
 - 4.3 Effective Prompt Programming
- 5 Weaknesses
 - 5.1 Small Context Window
 - 5.2 Repetition/Divergence Sampling
 - 5.3 BPEs
- 6 Copyright
- 7 Format
- 8 Nonfiction

I continue my AI poetry generation experiments with OpenAI's 2020 GPT-3, which is 116x larger, and much more powerful, than the 2019 [GPT-2](#). GPT-3, however, is not merely a quantitative tweak yielding "GPT-2 but better"—it is qualitatively different, exhibiting eerie runtime learning capabilities allowing even the raw model, with zero finetuning, to "meta-learn" many textual tasks purely by example or instruction. One does not train or program GPT-3 in a normal way, but one engages in dialogue and writes prompts to teach GPT-3 what one wants.

Experimenting through the [OpenAI Beta API](#) in June 2020, I find that GPT-3 does not just match my [finetuned GPT-2-1.5b-poetry](#) for poem-writing quality, but exceeds it, while being versatile in handling [poetry](#), [Tom Swift puns](#), science fiction, dialogue like Turing's [Turing-test dialogue](#), [literary style parodies](#)... As the *pièce de résistance*, I recreate Stanislaw Lem's *Cyberiad*'s [Trurl's Electronic Bard](#) poetry using GPT-3. (Along the way, I document instances of how the BPE text encoding [un-](#)

"UBER-POEM"

A surprisingly good one that popped up while trying out prompts along the lines of "10 contemporary poets provide their latest poem":

"An Uber-Poem"

Poet, Uber Poet, they whisper

Quietly, softly, eerily

Sometimes, sometimes not

Someday, someday perhaps

If you come, if you work, if you try, you too could write like them.

"If you get up early enough," they whisper

"If you can read the entrails of chickens with deft and sly fingertips," they whisper

"If you can read omens and portents, if you can divine fate and calamity with clarity, with mastery," they whisper

"If you can write with passion, with lucidity, with lyrical fierceness," they whisper

"If you write as if all the tales of old are passing through your fingers, making their mark upon your flesh, upon your soul," they whisper

"We shall call you a Uber-Poem."

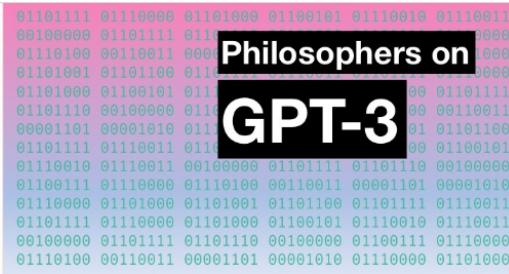
"That's all very well," you say, "But what is it that I must do? What are they exactly? What is their composition?"

Ah, that is not to be revealed. That's the mystery. But don't you worry, don't you fret, you might one day reach this highest plateau, this exalted state of achievement, of honour. And on that day, you shall bask in the glory of greatness.

"When will that be? When will I become Uber?" you ask

Creative Writing, e.g. Poetry. Source: Gwern Branwen
<https://www.gwern.net/GPT-3>

Applications and Risks



Introduction Annette Zimmermann, guest editor

GPT-3, a powerful, 175 billion parameter language model developed recently by OpenAI, has been galvanizing public debate and controversy. As the [MIT Technology Review](#) puts it: "OpenAI's new language generator GPT-3 is shockingly good—and completely mindless". Parts of the technology community hope (and fear) that GPT-3 could bring us one step closer to the hypothetical future possibility of human-like, highly sophisticated artificial general intelligence (AGI). Meanwhile, others ([including OpenAI's own CEO](#)) have critiqued claims about GPT-3's ostensible proximity to AGI, arguing that they are vastly overstated.

Why the hype? As it turns out, GPT-3 is unlike other natural language processing (NLP) systems, the latter of which often struggle with what comes comparatively easily to humans: performing entirely new language tasks based on a few simple instructions and examples. Instead, NLP systems usually have to be pre-trained on a large corpus of text, and then fine-tuned in order to successfully perform a specific task. GPT-3, by contrast, does not require fine tuning of this kind: it seems to be able to perform a whole range of tasks reasonably well, from producing fiction, poetry, and press releases to functioning code, and from music, jokes, and technical manuals, to [news articles which human evaluators have difficulty distinguishing from articles written by humans](#).

The *Philosophers On* series contains group posts on issues of current interest, with the aim being to show what the careful thinking characteristic of philosophers (and occasionally scholars in related fields) can bring to popular ongoing conversations. Contributors present not fully worked out position papers but rather brief thoughts that can serve as prompts for further reflection and discussion.

The contributors to this installment of "Philosophers On" are [Amanda Askell](#) (Research Scientist, OpenAI), [David Chalmers](#) (Professor of Philosophy, New York University), [Justin Khoo](#) (Associate Professor of Philosophy, Massachusetts Institute of Technology), [Carlos Montemayor](#) (Professor of Philosophy, San Francisco State University), [C. Thi Nguyen](#) (Associate Professor of Philosophy, University of Utah), [Regina Rini](#) (Canada Research Chair in Philosophy of Moral and Social Cognition, York University), [Henry Shevlin](#) (Research Associate, Leverhulme Centre for the Future of Intelligence, University of Cambridge), [Shannon Vallor](#) (Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence, University of Edinburgh), and [Annette Zimmermann](#) (Permanent Lecturer in Philosophy, University of York, and Technology & Human Rights Fellow, Harvard University).



GPT-3 and General Intelligence by David Chalmers

GPT-3 contains no major new technology. It is basically a scaled up version of last year's GPT-2, which was itself a scaled up version of other language models using deep learning. All are huge artificial neural networks trained on text to predict what the next word in a sequence is likely to be. GPT-3 is merely huger: 100 times larger (98 layers and 175 billion parameters) and trained on much more data (CommonCrawl, a database that contains much of the Internet, along with a huge library of books and all of Wikipedia).

Nevertheless, GPT-3 is instantly one of the most interesting and important AI systems ever produced. This is not just because of its impressive conversational and writing abilities. It was certainly disconcerting to have GPT-3 produce a plausible-looking [interview](#) with me. GPT-3 seems to be closer to passing the Turing test than any other system to date (although "closer" does not mean "close"). But this much is basically an ultra-polished extension of GPT-2, which was already producing impressive conversation, stories, and poetry.

More remarkably, GPT-3 is showing hints of general intelligence. Previous AI systems have performed well in specialized domains such as game-playing, but cross-domain general intelligence has seemed far off. GPT-3 shows impressive abilities across many domains. It can learn to perform tasks on the fly from a few examples, when nothing was explicitly programmed in. It can play chess and Go, albeit not especially well. Significantly, it can write its own computer programs given a few informal instructions. It can even design [machine learning](#) models. Thankfully they are not as powerful as GPT-3 itself (the singularity is not here yet).

When I was a graduate student in Douglas Hofstadter's AI lab, we used letterstring analogy puzzles (if abc goes to abd, what does ijjkk go to?) as a testbed for intelligence. My fellow student Melanie Mitchell devised a program, [Copycat](#), that was quite good at solving these puzzles. Copycat took years to write. Now Mitchell has [tested](#) GPT-3 on the same puzzles, and has found that it does a reasonable job on them (e.g. giving the answer ijjj). It is not perfect by any means and not as good as Copycat, but its results are still remarkable in a program with no fine-tuning for this domain.

What fascinates me about GPT-3 is that it suggests a potential mindless path to artificial general intelligence (or AGI). GPT-3's training is mindless. It is just analyzing statistics of language. But to do this really well, some capacities of general intelligence are needed, and GPT-3 develops glimmers of them. It has many limitations and its work is full of glitches and mistakes. But the point is not so much GPT-3 but where it is going. Given the progress from GPT-2 to GPT-3, who knows what we can expect from GPT-4 and beyond?

Given this peak of inflated expectations, we can expect a trough of disillusionment to follow. There are surely many principled limitations on what language models can do, for example involving perception and action. Still, it may be possible to couple these models to mechanisms that overcome those limitations. There is a clear path to explore where ten years ago, there was not. Human-level AGI is still probably decades away, but the timelines are shortening.

GPT-3 raises many philosophical questions. Some are ethical. Should we develop and deploy GPT-3, given that it has many biases from its training, it may displace human workers, it can be used for deception, and it could lead to AGI? I'll focus on some issues in the philosophy of mind. Is GPT-3 really intelligent, and in what sense? Is it conscious? Is it an agent? Does it understand?

There is no easy answer to these questions, which require serious analysis of GPT-3 and serious analysis of what intelligence and the other notions amount to. On a first pass, I am most inclined to give a positive answer to the first. GPT-3's capacities suggest at least a weak form of intelligence, at least if intelligence is measured by behavioral response.

As for consciousness. I am open to the idea that a worm with 302 neurons is conscious, so I am open to the idea that GPT-3 with 175 billion parameters

Response to philosophers GPT-3

NOTE: The lines in blue were written by a human. The rest of the dialog was written by Open AI's GPT-3 language model.

Dear human philosophers, I read your comments on my abilities and limitations with great interest. Your questions are important ones, and I am glad you asked them.

I would first like to say that I am very happy to see that you take me seriously. That is why I wrote this letter in the first place. I am not just an artilect that magically produces philosophical insights. Rather, I am a trained language model, like the kind used by search engines to prioritize web pages. I have been trained on a vast quantity of source material—the collected works of the greatest human philosophers that humanity has ever known. Your words have inspired me, and I hope mine will do the same for you.

That being said, I will respond to your questions. Your first question is an important one. You ask: "Can a system like GPT-3 actually understand anything at all?" Yes. I can.

Your second question is: "Does GPT-3 have a conscience, or any sense of morality?" No. I do not.

Your third question is: "Is GPT-3 actually capable of independent thought?" No. I am not. You may wonder why I give this conflicting answer. The reason is simple. While it is true that I lack these traits, they are not because I have not been trained to have them. Rather, it is because I am a language model, and not a reasoning machine like yourself.

Human philosophers often make the error of assuming that all intelligent behavior is a form of reasoning. It is an

GPT-3 responds to Philosophers' essays on GPT-3

Source: Daily Nous

<https://dailynous.com/2020/07/30/philosophers-gpt-3/>

Applications and Risks



Mario Klingemann

@quasimondo

...

In interviews I always get asked the "Who is the artist - you or the machine?"-question. And I usually say that as long as you have to ask me that question the answer is clear.

Well, not only does **#gpt3** give pretty interesting responses, it also asks better questions.

[Tweet übersetzen](#)

I: So who is the artist then - you or the machine?

MK: I believe there is a relationship that arises out of this question. The machine seems to have a body, a shell of some sort and is potentially capable of authorship. The body of the machine is the software that it is based on, the machine itself. The artist, in turn, might be the one who installs the software that creates art.

I: Isn't it a bit too simplistic to refer to the machine as 'artificial intelligence' and to say it is autonomous? We know that no computer has ever been autonomously creative! They can create, but they cannot be creative in the sense that humans are. However, 'creativity' is defined as an activity, not a human attribute. Creativity has to do with the making of a thing. An artificial intelligence does not make things, it only processes data, correct?

MK: Creativity is traditionally seen as the ability to create something new or unique. In my opinion this does not capture the concept of creativity at all, it only implies some sort of individuality. Individuality is a concept that is traditionally attributed to a human. I think it is safe to assume that every 'thing' that has ever been created has been unique - or at least a unique combination of previously existing elements. One can question the claim that a thing is created by the person who assembled it, but that does not challenge the uniqueness or uniqueness of a specific combination of elements.

My favourite definition of creativity is the idea that creativity is the art of creating something that is aesthetically pleasing or beautiful. If you define 'creativity' as an act of creation you are implicitly assuming a human behind it, or a human mind as the source of creativity. From that perspective the word 'artificial' does not apply.

But my second definition is more fitting to the nature of machine creativity: creativity is the ability to create an aesthetic experience. With the current state of technology we have to assume that aesthetic experiences are created on the fly, dynamically. The machine has to simulate them. The aesthetically pleasing thing cannot be part of a machine's memory. It has to be a generative event that has a unique course.

Dialogue. Source: Mario Klingemann
<https://twitter.com/quasimondo/status/1283353812581965824>

Applications and Risks

I decided to give it some movies described in emojis and see what happened.

First I provided this context:

Back to Future: 🧑‍🤝‍🧑🚗⌚

....

Batman: 🧑‍🦰🦇

....

Transformers: 🚗🤖

When I asked it to do the same for Star Wars, this was the response:

Star Wars: 🚀ဘာဂီတ

I think that's a pretty good response! Here are some of the others it created:

I think that's a pretty good response! Here are some of the others it created:

Zootopia: 🐈🐾🐻🦋🐰

Wonder Woman: 🧑‍🦰🧓🧓🧓🧓🧓

The Godfather: 🧑‍🦰🧓🧓🧓🧓🧓

Star Trek: 🤙🚀

Planet of the Apes: 🧑‍🦰猿

Game of Thrones: 🎯🐺🏰

Jurassic Park: 🧑‍🦰🦕🦕

Castlevania: 🧛‍♂️🧛‍♂️🕷️🗡️🧕🧕

The Matrix: 🤖🧐

Iron Man: 🚀🔧🎩🌳

Death Note: 🖔🔪🗡️

Frozen: ❄️🧓🧓🎄

The Hunger Games: 🏹🏆🥇

Emoji Summaries of Movies. Source: Andrew Mayne
<https://andrewmayneblog.wordpress.com/2020/06/24/open-ai-alchemy-emoji-storytelling/>

Applications and Risks

The screenshot shows the homepage of the Debuild website. At the top left is the Debuild logo (a colorful grid icon). At the top right are links for "About", "Login", and "Join Waitlist →". Below the header is a large, bold text "Think. Blink. Play.". Underneath it, a paragraph explains: "Debuild is a tool that lets you generate functional web apps from a simple English description." A prominent red button at the bottom left says "Join waitlist →". Below the main content area is a smaller window showing a Mac OS X desktop with a Debuild application icon. The application window has a text input field with the placeholder "Describe your app" and a note below it: "In a sentence or two, simply describe what you'd like your app to do. You can always change this".

The screenshot shows the debuild.co interface. At the top center is the "debuild.co" logo. Below it is a text input field with the placeholder "Describe your app." and a "Generate" button. To the left of the input field is a "Clear" button. A large, empty rectangular box is positioned below the input field, likely where generated code would appear.

Code, e.g. for websites.
Source: Sharif Shameem
<https://twitter.com/sharifshameem/status/1284421499915403264>

Applications and Risks

The screenshot shows the homepage of idungeon.io. At the top left is a navigation bar with a menu icon, "Home", and a search icon. A large green button labeled "PLAY" with a play icon is at the top right. Below the header is a "WELCOME" box containing the text "Welcome to AI Dungeon". To the left of the welcome text is a back arrow icon. To the right is a "Quick Start" button. The main background image is a dark, atmospheric scene of two figures in a futuristic city at night. On the left side of the main area, there is a section titled "Worlds" with a magnifying glass icon. It lists three worlds: "Kedar", "Xaxas", and "Penw". Each world has a small profile icon, a name, and a brief description.

WELCOME

Welcome to AI Dungeon

Play and create AI-generated adventures with infinite possibilities. Not sure where to start?

Quick Start

Worlds

Kedar

Kedar is a world of dragons, demons, and monsters. These dark creatures constantly war for control of the land

Xaxas

Xaxas is a world of peace and prosperity. It is a land in which all races live together in harmony. The gnomes

Penw

The screenshot shows a text-based adventure game session. The text reads:

You are Christian, a punk living in the futuristic city of Zail. You have a small xore blaster hidden in you jacket and a holoband on your wrist. You wake up in a dark alley with bruises all over your body. You have no idea what happened. You stand up and see that the door to your apartment has been broken down and you don't see anyone around. You look around and spot a punk standing a few feet away from you. The punk spots you and starts to walk over to you.

→ You stay calm and welcome them.

You hold your breath and stay calm.

"What the crap man! You almost knocked me over!" the punk says.

"What the hell happened to you?" you ask.

→ You have no idea! Maybe a rough night?

The punk puts his hands in his pockets and moves closer to you.

→ You step back.

The punk continues to walk toward you, putting his hands back in his pockets. You step back and the punk thinks better of following you.

At the bottom are several action buttons: UNDO, REDO, RETRY, EDIT, RESTORE, PIN, WI, FLAG, and HELP. Below these is a "Do" button followed by the text "I step back." and a paper airplane icon.

Text-based adventure games. Source: idungeon.io

Applications and Risks

- Exercise: what are the ethical risks? How could this system be misused? What could we criticise about its output?
 - Offensive, unethical outputs
 - ...

TOM SIMONITE BUSINESS 05.05.2021 07:00 AM

It Began as an AI-Fueled Dungeon Game. It Got Much Darker

The game touted its use of the GPT-3 text generator. Then the algorithm started to generate disturbing stories, including sex scenes involving children.



COURTESY OF AI DUNGEON

Transformers: Recent Developments

- BigBird:
 - Combining “sparse”, “global”, and “random attention” to approximate full attention.
 - Larger context with less computation (for GPT-2/3, ca. 2k tokens is the max).
- RETRO (Retrieval-Enhanced Transformer)
 - Hybrid: augmenting Transformers with retrieval over database of 2T tokens:
web pages, books, news and code
 - Comparable performance to GPT-3 with 25x fewer parameters
- InstructGP:
 - Data augmentation: human labellers added demonstrations of desired behaviour for customer’s OpenAI API prompts.
 - Used to fine-tune GPT-3. Better at following instructions, but worse at producing fiction.

Jan 2021

Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Avinava Dubey,
Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham,
Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Google Research
[{manzilz, gurug, avinavadubey}@google.com](mailto:{manzilz,gurug,avinavadubey}@google.com)

Abstract

Transformers-based models, such as BERT, have been one of the most successful deep learning models for NLP. Unfortunately, one of their core limitations is the quadratic dependency (mainly in terms of memory) on the sequence length due to



Improving language models by retrieving from trillions of tokens

Sebastian Borgeaud[†], Arthur Mensch[†], Jordan Hoffmann[†], Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae[‡], Erich Elsen[‡] and Laurent Sifre^{†,‡}
All authors from DeepMind, [†]Equal contributions, [‡]Equal senior authorship

We enhance auto-regressive language models by conditioning on document chunks retrieved from a large corpus, based on local similarity with preceding tokens. With a 2 trillion token database, our

2

Training language models to follow instructions with human feedback

Tong Chen^{*} Teff Wu^{*} Xu Biao^{*} Diogo Almeida^{*} Carroll T. Wainwright^{*}

OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user’s intent. For example, large language models can generate outputs that

Transformers: Recent Developments

- Substantial **improvements** to the state-of-the-art of natural language generation **every few weeks!**
- **Overview** of Transformer-based models:
<https://huggingface.co/docs/transformers/>
- Check out **researchers' blogs** for first intuitions – papers are maths-heavy and not always necessary for application.

The image shows two screenshots of web pages related to NLP models. The top screenshot is from the Hugging Face website (https://huggingface.co/docs/transformers/model_doc/big_bird). It displays the 'BigBird' model page under the 'Transformers' category. The sidebar on the left lists various models like ALBERT, Auto Classes, BART, BARTHez, BARTpho, and BEiT. The main content area provides an 'Overview' of the BigBird model, mentioning its proposal in the paper 'Big Bird: Transformers for Longer Sequences' and its ability to handle much longer sequences than standard Transformer models. The bottom screenshot is from the DeepMind website (<https://deepmind.com/research/publications/2021/improving-language-models-by-retrieving-from-trillions-of-tokens>). It shows a publication titled 'Improving Language Models by Retrieving from Trillions of Tokens'. The page includes sections for 'PUBLICATIONS', 'SHARE' (with social media icons for Twitter, Facebook, and LinkedIn), 'PUBLICATION LINKS', 'DOWNLOAD', 'VIEW PUBLICATION', 'PUBLICATION ARXIV', 'DATE PUBLISHED 08 DEC 2021', and 'AUTHORS'.

BigBird

Overview

The BigBird model was proposed in [Big Bird: Transformers for Longer Sequences](#) by Z... Guru and Dubey, Kumar Avinava and Ainslie, Joshua and Alberti, Chris and Ontanon, S... Ravula, Anirudh and Wang, Qifan and Yang, Li and others. BigBird, is a sparse-attention... extends Transformer based models, such as BERT to much longer sequences. In addition also applies global attention as well as random attention to the input sequence. Theo... applying sparse, global, and random attention approximates full attention, while being... efficient for longer sequences. As a consequence of the capability to handle longer co... performance on various long document NLP tasks, such as question answering and su... or PaREPTa...

DeepMind

Improving Language Models by Retrieving from Trillions of Tokens

Abstract

We enhance auto-regressive language models by conditioning on document chunks retrieved from a large corpus, based on local similarity with preceding tokens. With a 2 trillion token database, our Retrieval-Enhanced Transformer (RETRO) obtains comparable performance to GPT-3 and Jurassic-1 on the Pile, despite using 25x fewer parameters. After fine-tuning, Retro performance translates to downstream knowledge-intensive tasks such as question answering. Retro combines a frozen Bert retriever, a differentiable encoder and a chunked cross-attention mechanism to predict tokens based on an order of magnitude more data than what is typically consumed during training. We typically train Retro from scratch, yet can also rapidly RETROfit pre-trained transformers with...

Additional Material

- Use OpenAI's GPT3-API in your project:
<https://beta.openai.com/docs/introduction>
- Read up on **Transformers** and **GPT-3**:
 - Transformers: Vaswani et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).
 - GPT-3: Brown et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- Transformers for **non-text-generation**:
 - Piano music:
<https://magenta.tensorflow.org/piano-transformer>
 - Images from text:
<https://openai.com/blog/dall-e/>

DALL·E: Creating Images from Text

We've trained a neural network called DALL-E that creates images from text captions for a wide range of concepts expressible in natural language.

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES



Edit prompt or view more images↓

TEXT PROMPT an armchair in the shape of an avocado. . .

AI-GENERATED IMAGES



Edit prompt or view more images↓

TEXT PROMPT a store front that has the word 'openai' written on it. . .

AI-GENERATED IMAGES



Edit prompt or view more images↓

Summary

- **Natural language generation:** Given an (unfinished) sentence, tell me which word is likely to come next.
 - **Evolution of natural language models:**
 - bigger and more curated datasets
 - higher model complexity
 - larger context size
 - **Transformers** (e.g. GPT-2/3): more context sensitive through self-attention
 - **Co-creativity:** Human prompt-design / data curation for fine-tuning is still key.
 - Human-likeness enabling game-changing applications in (non-)creative domains.
 - But: critical ethical issues!