GPT-3 The New Mighty Language Model from OpenAI by Moiz Saifee May, 2020 Towards Data Sc...
Saved to Dropbox • 18 Jul 2020, 20:44

M

Sign in Get started

towards
data science

DATA SCIENCE MACHINE LEARNING PROGRAMMING VISUALIZATIO

This is your **last** free story this month. Sign up and get an extra one for free.

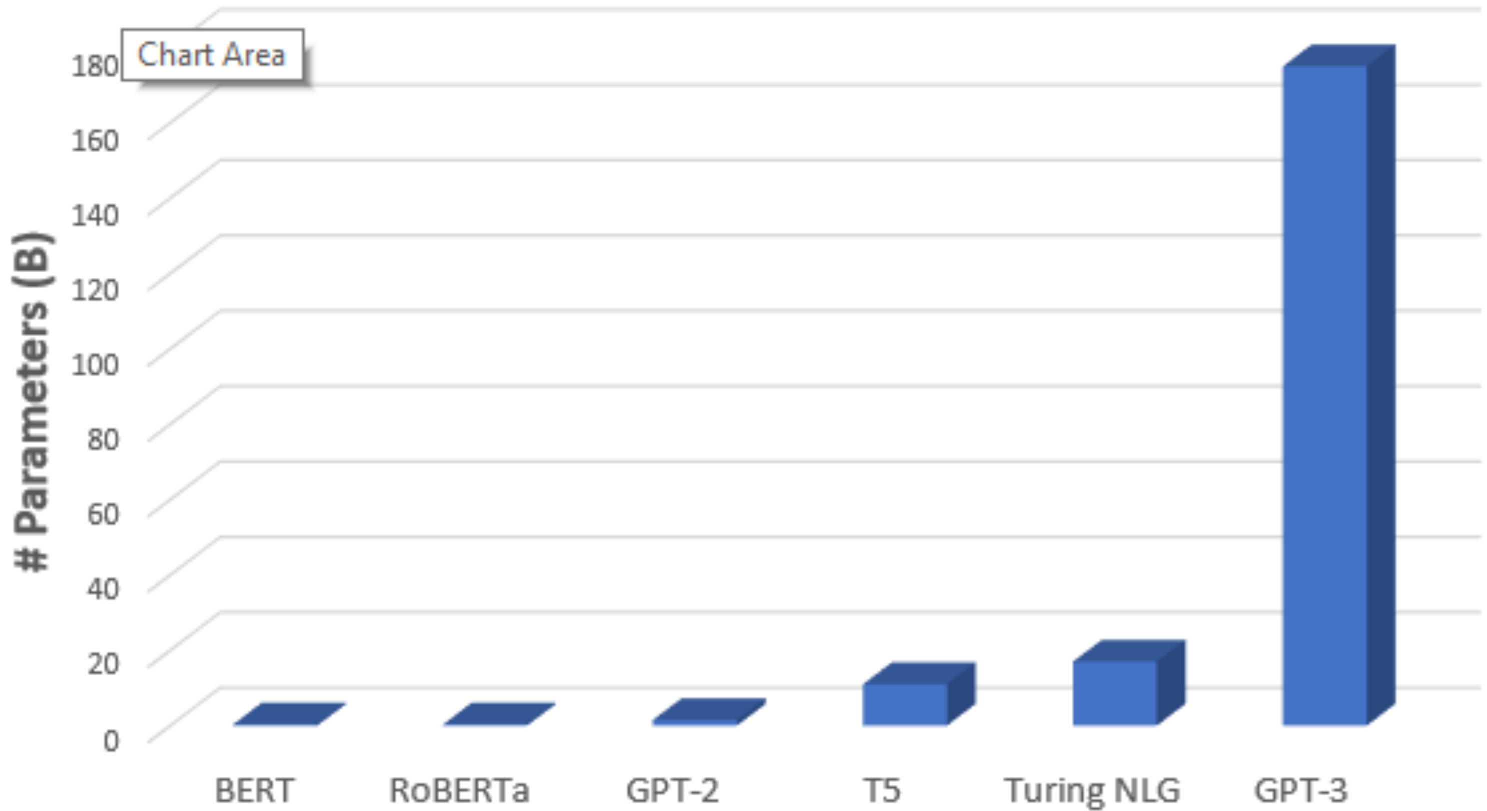
GPT-3: The New Mighty Language Model from OpenAI

Pushing Deep Learning to the Limit with 175B Parameters

Moiz Saifee Follow
May 31 · 4 min read ★

Introduction

OpenAI recently released pre-print of its new mighty *language model* GPT-3. Its a much bigger and better version of its predecessor GPT-2. In fact, with close to 175B trainable parameters, GPT-3 is much bigger in terms of size in comparison to anything else out there. Here is a comparison of number of parameters of recent popular pre trained NLP models, GPT-3 clearly stands out.



Model

What's New?

After the success of Bert, the field of NLP is increasingly moving in the direction of creating *pre-trained language models*, trained on huge text corpus (in an unsupervised way), which are later fine-tuned on specific tasks such as translation, question answering etc using much smaller task specific datasets.

While this type of *transfer learning* obviates the need to use task specific model architectures, but you still need task specific datasets, which are a pain to collect, to achieve good performance.

Humans by contrast learn in a very different way, and have the ability to learn a new task based on very few examples. GPT-3 aims to address this specific pain point, that is, its a task agnostic model, which needs zero to very limited examples to do well and achieve close to state of the art performance on a number of NLP tasks

Terminologies

Before we deep dive, it may be useful to define some commonly used terminologies:

- **NPL Tasks:** These are tasks which have something to do with human languages, example — Language Translation, Text Classification (e.g. Sentiment extraction), Reading Comprehension, Named Entity Recognition (e.g. recognizing person, location, company names in text)
- **Language Model**s: These are models which can predict the most likely next words (and their probabilities) given a set of words (think something like Google query auto-complete). Turns out these type of

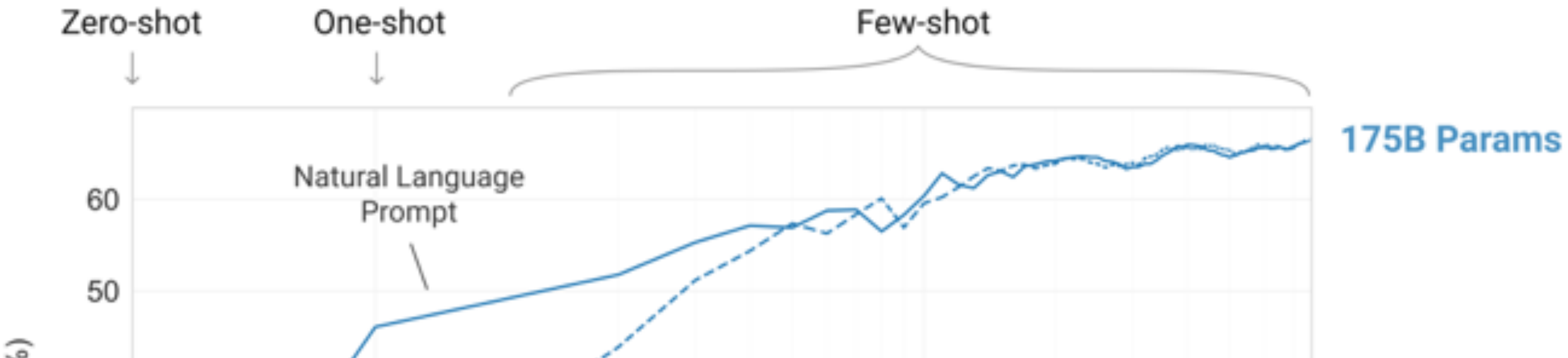
models are useful for a host of other tasks although they may be trained on mundane next word prediction

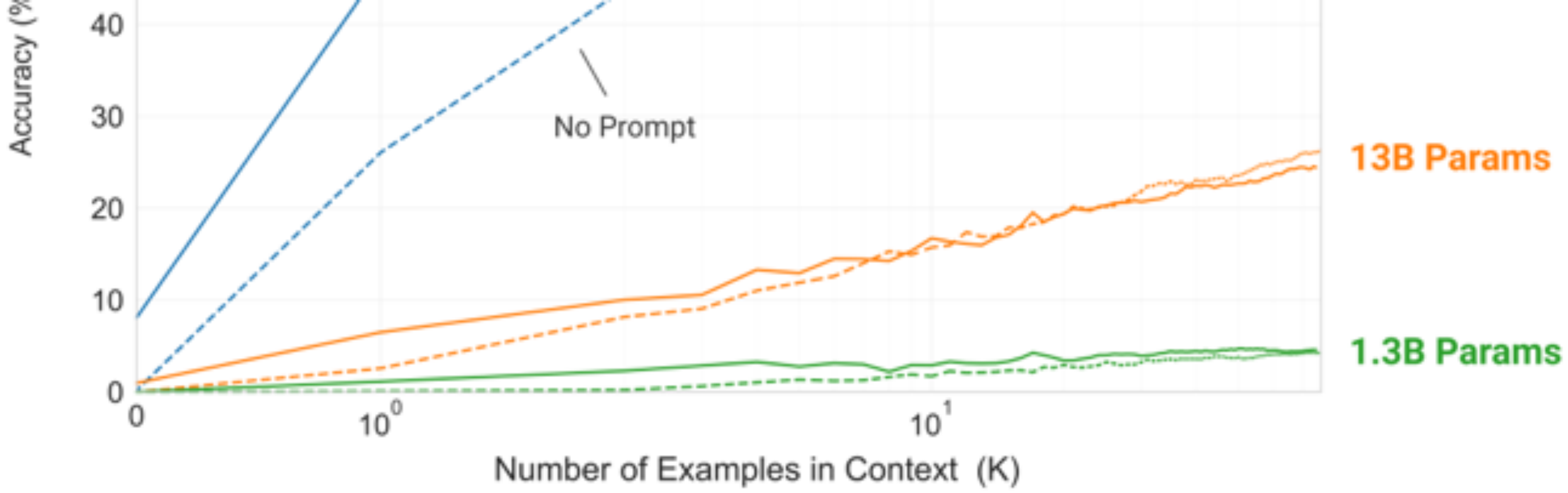
- **Zero / One / Few shot learning:** Refers to model's ability to learn a new task by seeing zero / one / few examples for that task
- **Transfer Learning:** Refers to the notion in Deep Learning where you train a model for one task (example object detection in images) , but the ability to leverage and build upon that for some other different task (example assessing MRI scans). After massive success in Computer Vision, its in vogue in NLP these days.
- **Transformer Models**: Deep learning family of models, used primarily in NLP, which forms the basic building block of most of the state-of-the-art NLP architectures these days. You can read more about *Transformers* at one of my earlier blog

The Approach

The model is built using the standard concepts of *Transformer*, *Attention* etc and using the typical *Common Crawl, Wikipedia, Books* and some additional data sources. A lot of things — pre training, model, data are similar to GPT-2, but everything (model size, data size, training time) is just a lot bigger. In fact its humongous size is what drives most of the benefits of the model.

The following graph shows the benefit in accuracy for various Zero / One / Few shot tasks as a function of number of Model parameters, clearly major gains are achieved due to the scaled up size.

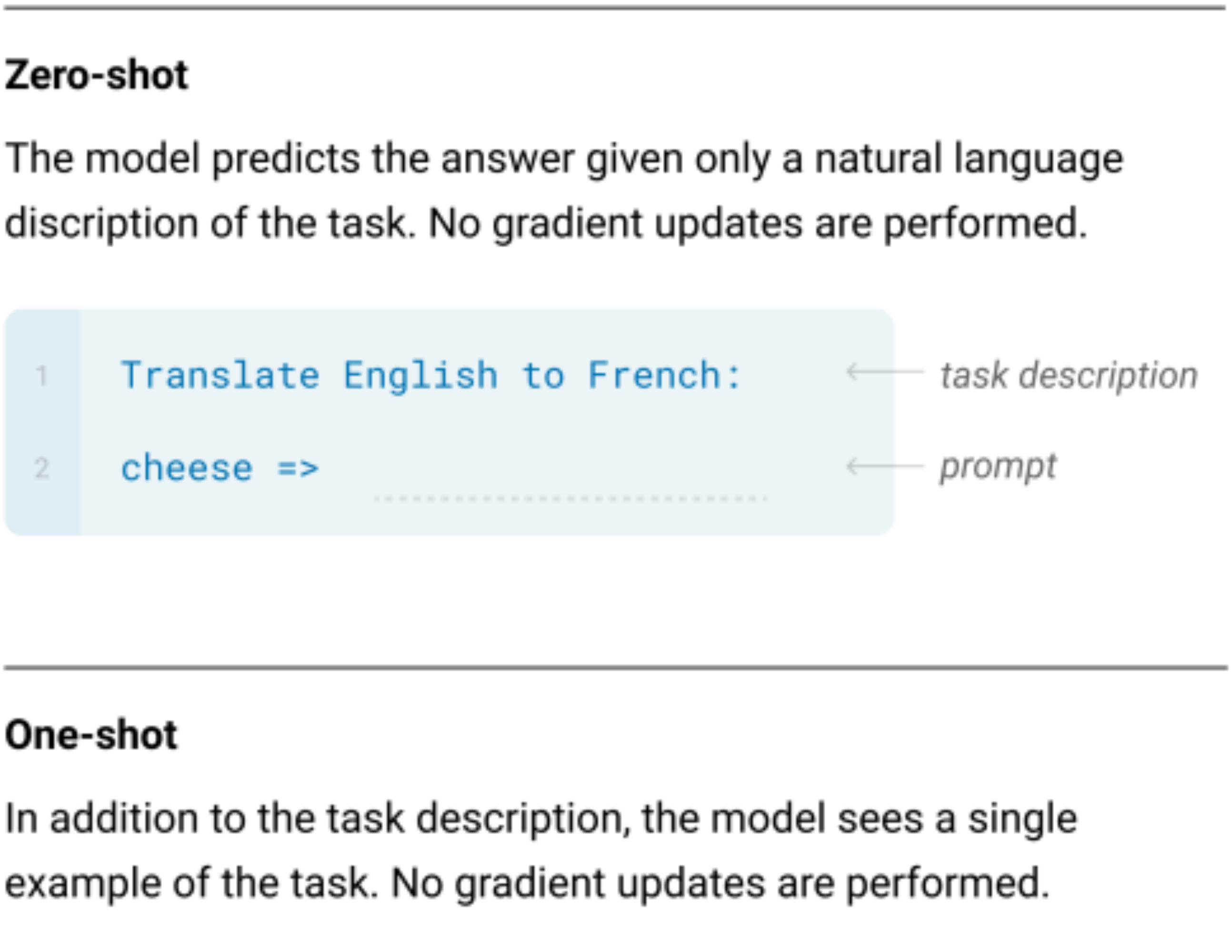




Source: Paper

Most of the things used in the model are so huge — example 96 *Attention* layers, *Batch Size* of 3.2M, 175B *Parameters* — that they are unlike anything in the past. The model is ~10x larger in terms of number of parameters to the next closest thing (Microsoft Turing NLG with 17B parameters)

There is no need to do gradient / parameter updates (fine tuning) for using the GPT-3 model for various tasks. One can just interact with the model using natural language and/or provide some examples of the tasks that you are trying to do and the model will do it!



Zero-shot

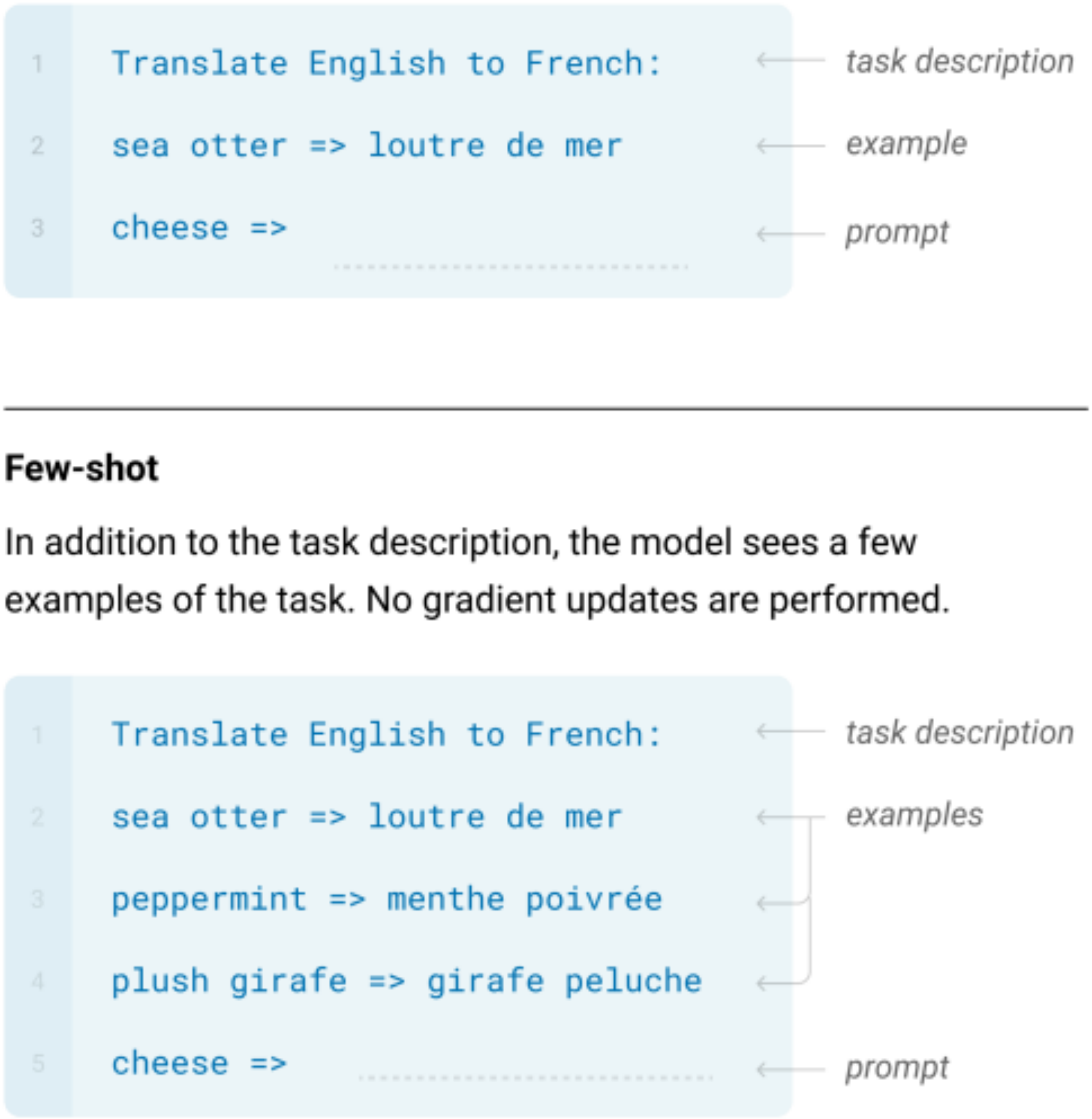
The model predicts the answer given only a natural language discription of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

task description
prompt

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

Source: Paper

What Does All this Mean?

The concept of not requiring large custom, task specific datasets, in addition to not requiring task specific model architectures is a huge step in direction of making cutting edge NLP more accessible.

While GPT-3 delivers great performance on a lot of NLP tasks example — word prediction, common sense reasoning — but it doesn't do equally well on everything. For instance it doesn't do great on things like — Text synthesis, some reading comprehension tasks etc. In addition to this, it also suffers from bias in the data which may lead the model to generate stereotyped or prejudiced content. So there is more work to be done here.

In addition to all this, the huge size of GPT-3, makes it out of bounds for almost everyone except a select few companies and research labs in the world. As per the authors, the model is very versatile and contains a very wide range of skills not needed for specific tasks and there might be a scope of creating smaller, more manageable task specific models using the concept of *distillation*.

Would be exciting to see how this thing evolves in future.

Sign up for The Daily Pick

By Towards Data Science

Hands-on real-world examples, research, tutorials, and cutting-edge techniques delivered Monday to Thursday. Make learning your daily ritual. Take a look

✉ Get this newsletter

Create a free Medium account to get The Daily Pick in your inbox.

NLP · Machine Learning · Artificial Intelligence · Data Science · Deep Learning

480 claps

WRITTEN BY

Moiz Saifee

Senior Principal at Correlation Venture. Passionate about Artificial Intelligence. Kaggle Master; IIT Kharagpur alum

Follow

Towards Data Science

A Medium publication sharing concepts, ideas, and codes.

Follow

See responses (2)

More From Medium

Ten SQL Concepts You Should Know for Data Science Interviews

Terence S in Towards Data Science

Features You Likely Don't Use in Python 3 — But You Should

Amritansh Sagar in Towards Data Science

20 Pandas Functions That Will Boost Your Data Analysis Process

Soner Yıldırım in Towards Data Science

Will AutoML Be the End of Data Scientists?

Frederik Bussler in Towards Data Science

3 Ways to Get Real-Life Data Science Experience Before Your First Job

Terence S in Towards Data Science

How You Should Read Research Papers According To Andrew Ng (Stanford Deep Learning Lectures)

Richmond Alake in Towards Data Science

Sktime: a Unified Python Library for Time Series Machine Learning

Alexandra Amidon in Towards Data Science

GPT-3: Creative Potential of NLP

Vlad Alex (Merzmensch) in Towards Data Science

Discover Medium

Welcome to a place where words matter. On Medium, smart voices and original ideas take center stage - with no ads in sight. Watch

Make Medium yours

Follow all the topics you care about, and we'll deliver the best stories for you to your homepage and inbox. Explore

Become a member

Get unlimited access to the best stories on Medium — and support writers while you're at it. Just $5/month. Upgrade