



# GPT-3日本語版プロジェクトの紹介



# Who am I

生さんこと‘坂本俊之’

職業：ITコンサルタント・機械学習エンジニア

職種：フリーランスエンジニア

C&R研究所

『作ってわかる！アンサンブル学習アルゴリズム入門』等  
著者



なにやった人？

- ・ 個人的に、GPT-2日本語版とかOCRプログラムとか、人工知能（AI）プログラムを作ってる人
- ・ 作ったAIプログラムは、GitHub上で公開して、モデルも自由にダウンロード出来るようにしている



<https://github.com/tanreinama/>



# なにやった人かをもう少し

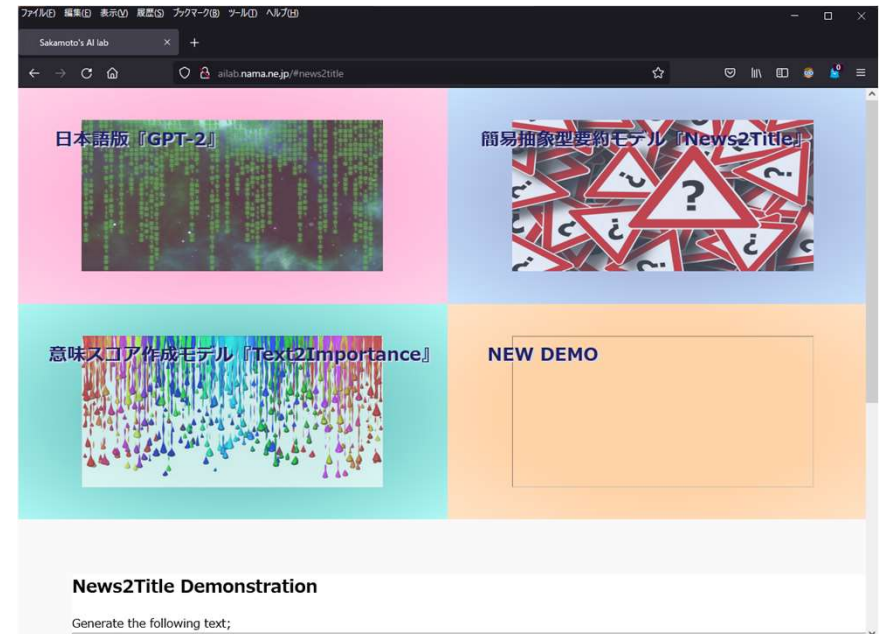
## GPT-2日本語版とは？

- 「あまりに危険」と言われたOpenAI謹製の文章生成AI
  - ・その日本語版を独自に一から作成したもの
  - ・コーパス収集やエンコードアルゴリズムから全て独力で制作
- 2019/12制作開始、2020/8/26クラウドファンディング
  - ・2020年後半あたりから若干バズり始めた気がする
  - ・現在GitHub ☆217 23forks

## その他

- 文章要約AI「News2Title」
- 文章スコアリングAI「Text2Importance」
- 日本語版RoBERTa
- 日本語OCR

などなど



デモサイト：<http://ailab.nama.ne.jp/>



# GPT-2 & GPT-3とは何か？

## A. 米OpenAI社が作成した、巨大な自然言語処理AIモデル

→2019年に、それまでのAIを遙かに超える性能のAIモデルが登場。

あまりにも容易にフェイクニュース等を生成可能なことから、研究者が「危険すぎる」として、公開を見送る事態に。  
(小規模なモデルから段階的に公開され、現在はGPT-2モデルは公開済み)

## GPT-3は、GPT-2をさらに大規模にしたAIモデル

→モデルそのものは公開されていない。(APIを通じて有償で利用するのみ)

英語におけるチャットボット・自動回答AI等のベースとして、様々なサービスにおけるバックエンドとして利用されている。

## 日本語版GPT-2とは？

→坂本俊之（応募者）が公開している、GPT-2の日本語版モデル。

日本語のコーパス（学習データ）収集から初めて、全てを1から学習させたもの。

モデル形式（ニューロンの繋ぎ方）は英語版と共通。



# なぜGPT-3日本語版が『破壊的な挑戦』となるのか？

## A. 様々なAIの基礎となる汎用言語モデルだから

→GPTのような言語モデルを基礎とする、マルチドメインAIがAI研究の最先端。

→ソースコードの自動生成を行う「GitHub Copilot」  
入力された文章から絵を描く「DALLE」

これらは、GPTと  
同じ技術の応用

GitHub Copilot

**Your AI pair programmer**



コメントからソースコード生成

## 今後も、様々なAIが、GPT-2&3をベースに登場してくる

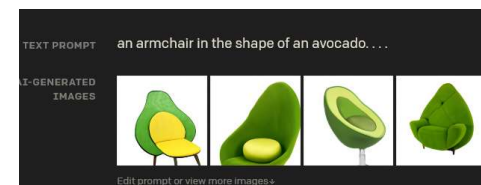
→言語モデルとして、チャット・自動回答AI・機械翻訳等の用途ももちろんだがそれ以上。

→AI≡画像認識、という時代は既に終わり、今後は言語×AIの時代が来る。

→GPT-3のようなAPIとして、ではなく、自由に使える公開モデルがあるかどうかで、応用AIの開発が出来るかどうか、が決まる。

→特に、AI assisted Designにおいては、基礎となる超重要な技術。

DALL-E



『アボガドで出来た椅子』  
をAIに描かせた結果



# なぜGPT-3の日本語版が求められるのか？

## A. 日本のAI競争力に直結するから

→AIの大きさはパラメーター数で測られるが、自然言語処理AIでは大規模化が進む。

→GPT-3は1750億パラメーターだが、中国「悟道2.0」は1兆6000億とさらに巨大。

**既に、米中では、自然言語処理AIのパラメーター数競争が始まっている！！**

単に、チャットボット等会話AIの性能程度の話ではない！

→プログラミング・デザイン・設計など様々な分野で、今後はAI assisted Designによる開発が一般化してゆくと予想。

※ソースコード生成の「GitHub Copilot」、イラスト生成の「DALL-E」などがさらに進化して産業へ応用されてゆく。

→そうなったときに、日本語ネイティブなAIモデルが存在しなければ、日本の産業全体の国際競争力が低下する。

→プログラムを書くなら、まず英語を学ばないとコーディング補助ツールが使えない、

設計図を引くにも、まず中国語を学ばないと製図補助ツールが使えない、

・・・等々という自体になりかねない！！



# 技術を乗り越えて実現したい目標

## A. GPT-3（相当）の日本語版学習済みモデルを作成し、公開したい

- GPT-3はそもそも英語版で作成され、日本語では翻訳などを通じて利用するしかない。
- API等を通じて利用するのではない、誰でも自由に利用できる学習済みモデルを公開したい。
- 日本語AIは、海外の会社が作ったモデルではなく、国内に独自技術を保持しておくべき。

## 技術的困難点

- 自然言語処理AIはパラメーター数の巨大化が進み、通常のGPUでは学習が難しい。
- 既に、ゲーム用GPUで学習可能な上限サイズのことを公開しているが、更に巨大なモデルが求められる。

## 実現の道筋

- モデルの軽量化 + GCP上のTPUサーバーで学習させる事で、巨大モデルを学習する。
- GPT-3は基本的にGPT-2のスケールアップモデルなので、GPT-2の開発経験があれば作成は容易。
- 開発スケジュールは「必要となる資金と開発計画」セクションで後述。

注：GCP=Google（alphabet社）が提供しているクラウドサービス

GPU = グラフィックスアクセラレーター、機械学習（AI作成）や仮想通貨マイニング等の高速計算にも用いられる

TPU = 機械学習用アクセラレーター





## これまでの取り組み





# GPT-2日本語版プロジェクトの取り組み

## プロジェクトページ

公開ページ： <https://github.com/tanreinama/gpt2-japanese/>

デモンストレーション： <https://ailab.nama.ne.jp/>

## 開発の推移

### 最初期・・・2019年11月

GPT-2に日本語版Wikipediaコーパスを学習

### 独自コーパス作成・・・2020年3～7月

スクレイピングプログラムを作成し、AWS上で実行

週末毎に対応サイトを一つ追加しては起動する日々

合計21GB程度の混合コーパスを作成

### experimental V1・・・2020年8月

12heads,12layersのサイズのモデルを作成

一般モデルと小説生成モデルを作成

パラメーター数 = 1億

### 日本語BPEEncoder・・・2020年9月

最も基礎となるエンコード方式から新規作成

オリジナルGPT-2から離れ、独自の進化を開始

### mediumモデル相当モデル・・・2020年11月

オリジナルGPT-2のmediumモデルに相当するサイズ

パラメーター数 = 3億2000万



# クラウドファンディングでの資金調達

## AIの学習は金がかかる

- 巨大モデルの学習は、ゲーム用のGPUではメモリが足りず、AWS等のクラウド上でサーバーを借りる必要がある。
- 学習に時間がかかる（GPT-2 large相当モデルの学習には、1GPUで100日間程度かかる）。

## Campfireでクラウドファンディングを開始

プロジェクトページ：<https://camp-fire.jp/projects/view/320938>

自然言語処理AIモデルの公開に理解があるスポンサーを求めて、クラウドファンディングサイトで資金を募集。

わずか25日間の募集期間にもかかわらず、25人の支援者、30万円以上の支援をGET！



これで、1GPUで学習出来るサイズのものは学習出来た

しかし、largeモデル相当の学習までが精一杯（パラメーター数＝6億4000万）  
米中の巨大モデル（パラメーター数＝1000億以上）には到底及ばない・・・

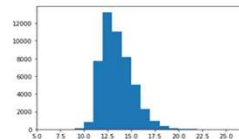
# 技術的詳細

## 独自エンコーダーの実装

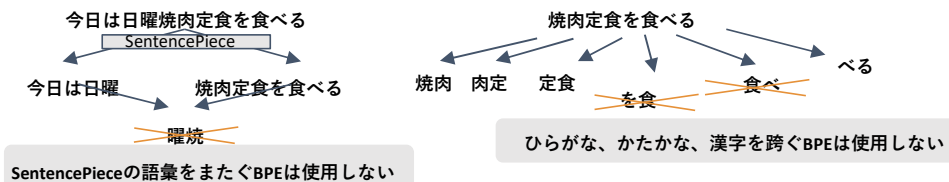
### 日本語BPEEncoderの開発

- 日本語には、英語のような単語間の区切りが無い
  - 予め分かち書きしておく必要がある
  - 標準的な手法は形態素解析を使う
- AI研究のトレンドはサブワード単位の学習に移行
  - 1単語をより細かく分割してエンコードする
  - 形態素解析の使用に疑問が生じる
- 既存のサブワード分割アルゴリズムの性能が低かった
  - SentencePieceを使うと性能が低下する

SentencePiece分かち書き後の  
Token内平均文字数 = 13.02



- 完全オリジナルの日本語エンコーダーを開発



## ゲーム用GPUの限界

### 2020年12月～2021年3月にかけてlargeモデル相当モデルを学習

- RTX3090×1枚で、98日かけて学習
  - ただし、学習OptimizerがAdagrad (lrを手動調整)
  - GPUメモリ不足で理想的なアルゴリズムで学習することが出来ない
- lossがmediumモデル同等まで低下しない
  - 出力文章も悪いように見える（定性的な評価はまだ）
  - モデルに対してコーパスが少ないのか、Optimizerが悪いのか不明

large相当モデル学習曲線→

パラメーター数 = 6億4000万





## 必要となる資金と開発計画



# GPT-3（相当）日本語版モデル作成目標

## 作成モデルのサイズ目標

- オリジナルGPT-3は、1億2500万パラメーターから1750億パラメーターまでバリエーションがある。
- 30億パラメーター程度（2.7B相当）が当面の目標。

1000億パラメータークラスはスーパーコンピュータと数千万円規模の資金が必要

Model Name	$n_{\text{params}}$	$n_{\text{layers}}$	$d_{\text{model}}$	$n_{\text{heads}}$	$d_{\text{head}}$	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	$6.0 \times 10^{-4}$
GPT-3 Medium	350M	24	1024	16	64	0.5M	$3.0 \times 10^{-4}$
GPT-3 Large	760M	24	1536	16	96	0.5M	$2.5 \times 10^{-4}$
GPT-3 XL	1.3B	24	2048	24	128	1M	$2.0 \times 10^{-4}$
GPT-3 2.7B	2.7B	32	2560	32	80	1M	$1.6 \times 10^{-4}$
GPT-3 6.7B	6.7B	32	4096	32	128	2M	$1.2 \times 10^{-4}$
GPT-3 13B	13.0B	40	5140	40	128	2M	$1.0 \times 10^{-4}$
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	$0.6 \times 10^{-4}$

## 現在のモデルのままスケールアップ可能な範囲

- AWS等のクラウドで借りれる8GPUサーバーで、20億パラメーター程度が限界。（学習時間：6ヶ月ほど見込み）  
p4d.24xlarge(東京) → \$44.5 / 時間      $\$44.5 \times 24 \times 180 = \$192,240 \div 2 \text{千万円オーバー} \rightarrow \text{無理}$

## モデルの改良とTPUクラスターの利用

- モデルを改良し消費メモリ削減 & 高速化、GCPのTPUサーバーで学習。（学習時間：3ヶ月ほど見込み）

TPU v3-8(アイオワ) → \$2.4 / 時間

$\$2.4 \times 24 \times 90 = \$5,184 \div 57 \text{万円程度} \rightarrow \text{可能}$



# 日本語版モデル開発計画

## 改良版GPT-3（相当）モデルの作成

- 独自エンコーダーを使用する日本語版モデルに合わせて、モデルの最適化を図る。
- TPUによる学習が出来るように、学習プログラムを改良。
- 活性化関数の適正化による高速化。
- 追加で収集した学習データ（40GB～）で学習。

予定開発期間：

ステージ1：

モデル最適化 ～0.25ヶ月  
学習プログラム改良 ～0.25ヶ月  
実験的学習（学習パラメーター最適化） ～0.5ヶ月  
大規模モデル学習 ～3ヶ月

合計4ヶ月

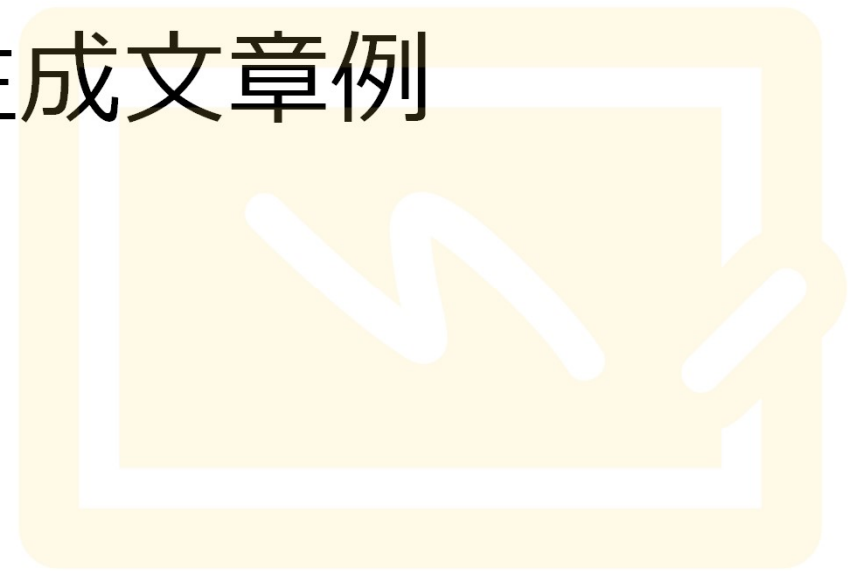
ステージ2：

超大規模モデル学習プログラム作成 ～1ヶ月  
実験的学習（学習パラメーター最適化） ～1ヶ月  
大規模モデル学習 ～6ヶ月





## GPT-2日本語版 生成文章例



# デモンストレーションサイト (ailab.nama.ne.jp) のログより

## Input Text:

個人的な見解は、「日記にえらそうなことを書く人間は、教養がありそうに見えるが、社会や人付き合いとかについてはほとんど無知なガキンチョ。お手上げだ。」です。例外はありますが、処世術に関して他人が口を出すと多くの場合ろくなことにはなりません。成長の足をひっぱるのです。他人の指導でなく、自己責任で生存競争に臨んで、その人間が育つか滅びるか決まるのです。もし人を育てたいと思うのであれば、他人は関与しないということが一番だと思います。これが個人的に言えることであっても、もしあなたがそういう人を育て上げたとき、どう思いますか？

## Output Text:

人に嫌われなくなければ、嫌われないことを努力しなければ、本当に嫌われます。それを教えなさい。無理ですよ。ただそれに対しての説得は自己責任だということは、わかってくれましたか？その時点で、まだ私は人間が信じられなくなっているのかもしれませんが、でも、今から思うに、あなたがそういう人を育てるのが当たり前という前提なら、私があなたの親に、「お母さん、こんなこと言う割には、何で嫌われないんでしょうね？嫌われないのに、どうして育てられないんでしょう？」って、質問するかもしれない。これまで通りでいくなら、それはそれで構わない。でもそれには、あなたの思うような教育は一切期待しない。あなたがどう思っているか知らないけれど、教育されないと言うのなら、あなたがどういう風な育て方をしているか知る義務もないですし、あなたが嫌われなくなければ、その後も育ててもらわなければいけないですよ。ただ言っておくけど、あなたは、あなたなりの価値観を持った、この人たちに、育てられてきたんじゃないの？自分の思うような育ち方しても、相手にとって「嫌われるわ」なんて思うはずがない。





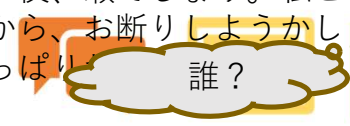
# デモンストレーションサイト (ailab.nama.ne.jp) のログより

## Input Text:

イギリスも氷河期世代に相当する人々がいてね、今の50代なんだけども。本当に仕事がなく海外移民した人もかなりいた。IT業界に流れ着いた人も沢山いて、今の業界のベテランの人がそんな感じ。

## Output Text:

そんな中に、俺みたいに日本で生活する子がたまにいて、日本語が堪能な人間が日本に流れ着いたのよ。俺も日本語堪能だったけど、まあ仕事がないわけではないんで、日本語勉強できたわけだから、結構楽なんだ。でもその子はお昼ごはんを食べて、その後は帰ってしまう。俺が、ちょっと寂しいと思ったのは日本にいたのに、何か別の場所で生活していたと思うようになったからだ。俺が日本に住むことによって、別の世界に来てしまったのは当然といえば当然だ。俺は、子供に興味ないけど、自分の子供である彼女と一緒にいると、子供っぽい感じがする。日本で生活するには、何か教育の授業を受けることだと、俺は思う。その中に彼女が入ったのは、たまたまね。何か聞いてはいけないこととか.....まあ、それは後で聞くことになると思うけどさ。その後、彼女と初めて会ったのが、高校の入学式だった。俺はこの日を待って、彼女の家に向かった。彼女はあっけにとられつつも、俺のことを「こんにちは」って言ってくれた。その時点で俺は気づいたよ。この人、俺の親じゃないの？ って。「こんちはー」という挨拶の後、彼女を部屋に招き入れた。部屋の中で彼女はベッドの上で、漫画本に目を走らせていた。部屋の床に漫画本が転がっているのは、お気に入りの漫画が倒れていないかを確認しているのだろう。彼女も漫画が好きなんだなと思った。そして、彼女はベッドの上の漫画本を見た。「うーん。これは面白そうだね」何を言っているんだろうと、俺は思った。「何を見てるの？」「ああ、漫画だけど」彼女がそう言うと、漫画本のページが開いた。彼女も俺も、何だか読んでいたようだった。しばらくすると、彼女が話を切り出した。「今日の夜、暇でしょう。私と映画を見ないかい？」「別に俺は暇だけど」「私と映画を見たらどうだい？」「俺、映画なんて観たことないから、お断りしようかしら」「なら、見ようよ。どうせ、私も暇なんだから」彼女は笑っているようだった。俺は、彼女の意図がさっぱりとわかった。



# デモンストレーションサイト (ailab.nama.ne.jp) のログより

## Input Text:

むかし、むかし、あるところにおじいさんとおばあさんが住んでいました。ある日、おばあさんは川へ洗濯に行きました。

## Output Text:

綺麗な川で、洗濯物が次々に干された様子を見て、おじいさんはホッとしました。洗濯物を見て、洗濯物を取り替えたおばあさん。今度はおじいさんも洗濯に出かけました。そこへ、おじいさんとおばあさんのお家がありました。お父さんとお母さんは洗濯物を取り替えて、洗濯機で洗濯をしていました。おじいさんとおばあさんは、川の辺で洗濯物を取り換えました。洗濯機で、今度はおじいさんも洗濯物を洗濯しました。今度はおばあさんも、洗濯機で洗濯をしました。おばあさんと、おじいさんは、洗濯機で洗濯物を洗い、洗濯機で洗濯を終えました。洗濯機が終わった後、おばあさんとおじいさんは、洗濯物を、洗濯機まで運び、洗濯機のある場所において、畳んだあと、取り換えに向かいました。おじいさんの、着た服は、おばあさんの、衣類で、おばあさんにとっての、おじいさんになったのです。こうして、川に洗濯物を引いたおじいさんとおばあさんは、川へ洗濯に行きました。川をわたると、少しだけ暗くなりましたが、お二人は、川にたどり着けました。洗濯物を見て、家の中を掃除したあとで、おばあさんに、洗濯物を外に出すように言いました。それから、お二人、洗濯物を川の近くに置きました。洗濯物が引くたびに、おばあさんは、汚れた洋服や、汚れた服を着ていました。それからしばらく、お二人、洗濯物を外で、洗濯しました。時々、川で洗濯をしたお二人を川に上げるのを手伝ったり、お父さんとお母さんと一緒に、川に洗いにいったり、川で洗濯をしたりしていました。次に、川の近くには、何もない所がありました。今度はお母さんのいる川なのでした。お母さんはいくらか汚れた服を着ていて、川の近くにいるおかげで、洗いできました。お母さんがいなくなれば、川での洗濯も終わりなのです。次に、川の近くには、小さな公園がありました。公園の入り口では、洗濯機がかけ、洗濯ができ、お母さんと一緒に川の近くで、洗濯をしました。小さな公園の入り口では、洗濯機がなく、入れないと洗濯ができないみたいでした。そこで、お父さんがいない日は、川の近くに、洗濯物を干しました。川に洗濯物が洗いにいくことはないので、お母さんとお父さんは、洗濯ができないことが、とても悲しいです。無限に洗濯しなければならない地獄か何か？

洗濯物を干したのを見ました。川の近くには、小さい公園が置かれ、川の近くにはなにもなく、大きな広場のようになってました。広場では、多くの大人たちも行きました。沢山の人が、洗濯をしていたのです。川から、洗濯するおじいさんに、川の近くにある小さな公園の中に、洗濯機がある、洗濯物が川に洗ってもらう、洗濯機がある、というのを聞かされ、私は、私の洗濯物が綺麗になればいいと、





# GPT-2日本語版 使用例



# 事例紹介

## 事例① - 坂本俊之（応募者）が受託した案件

→漫画あらすじ生成AIの作成

漫画のストーリーを自動生成するAIを作成したいという相談を受け、案件として受託。  
GPT-2日本語版のファインチューニングと、RoBERTa日本語版による後処理で実装。

## 事例② - 坂本俊之（応募者）が作成したデータセット

→日本語フェイクニュースデータセットの作成

クリエイティブコモンズ（改変可）ライセンスで公開されている、Wikipedia日本語版の記事を元に、  
フェイクニュースを作成。

ニュース記事が、フェイクニュースなのか人間が執筆したものを判定するAIの作成に使用。

<https://github.com/tanreinama/Japanese-Fakenews-Dataset> にて公開中。



# 事例紹介

## 事例③ - ブログで紹介された

→npaka氏のブログにてGPT-2日本語版のファインチューニングが紹介された。  
ゲーム中の台詞を学習させて、ゲームキャラが喋るようなセリフを出力。  
<https://note.com/npaka/n/ne55d063e1ed8>

### →結果

ああっ！もう！分かったから！...つたく、しよーがないなあ、お兄ちゃん  
は...  
=====

そういえばお兄ちゃん。『アプトロニック』って水月先輩の事持ってたよ  
ね？実はね...。☆3（スリー）☆3を歌ってる大空寺羽月は、日本の人の想  
像を絶するよね？実はね...。  
=====

ねえ、お兄ちゃん。お姉ちゃんと水月先輩、どっちが好きなの？  
=====

ふへん...。お兄ちゃん、意外と普通の人なんだね〜...。  
=====

ねえ、お兄ちゃん？お兄ちゃんって、一体、何を意味の分からないこと言  
うんだろ？意味の？意味の？意味の？...ま、まあ、お兄ちゃんは気にしな  
いで！



### gpt2-japaneseの使い方 (2) - GPT-2のファインチューニング

♡ 18

npaka  
2020/12/05 08:18



「gpt2-japanese」の「smallモデル」と「ファインチューニングのコード」が公開されたので、日本語によるGPT-2のファインチューニングを試してみました。

#### ▼ 目次

前回

1. gpt2-japaneseのインストール
2. モデルのダウンロード
3. モデルの動作確認
4. データセットの作成
5. ファインチューニング
6. ファインチューニングしたモデルの動作確認

次回

#### 前回

##### gpt2-japaneseの使い方 (1) - mediumモデルによる日本語テキスト生成

「gpt2-japanese」の「mediumモデル」が公開されたので、AIによる日本語テキスト生成を試してみました。1. gpt2-japaneseのインストール  
npaka/gpt2-japanese-japanese-GPT2-Generation-Model-Code

♡ 21

