

【第72回】大学等におけるオンライン教育とデジタル変革に関するサイバーシンポジウム

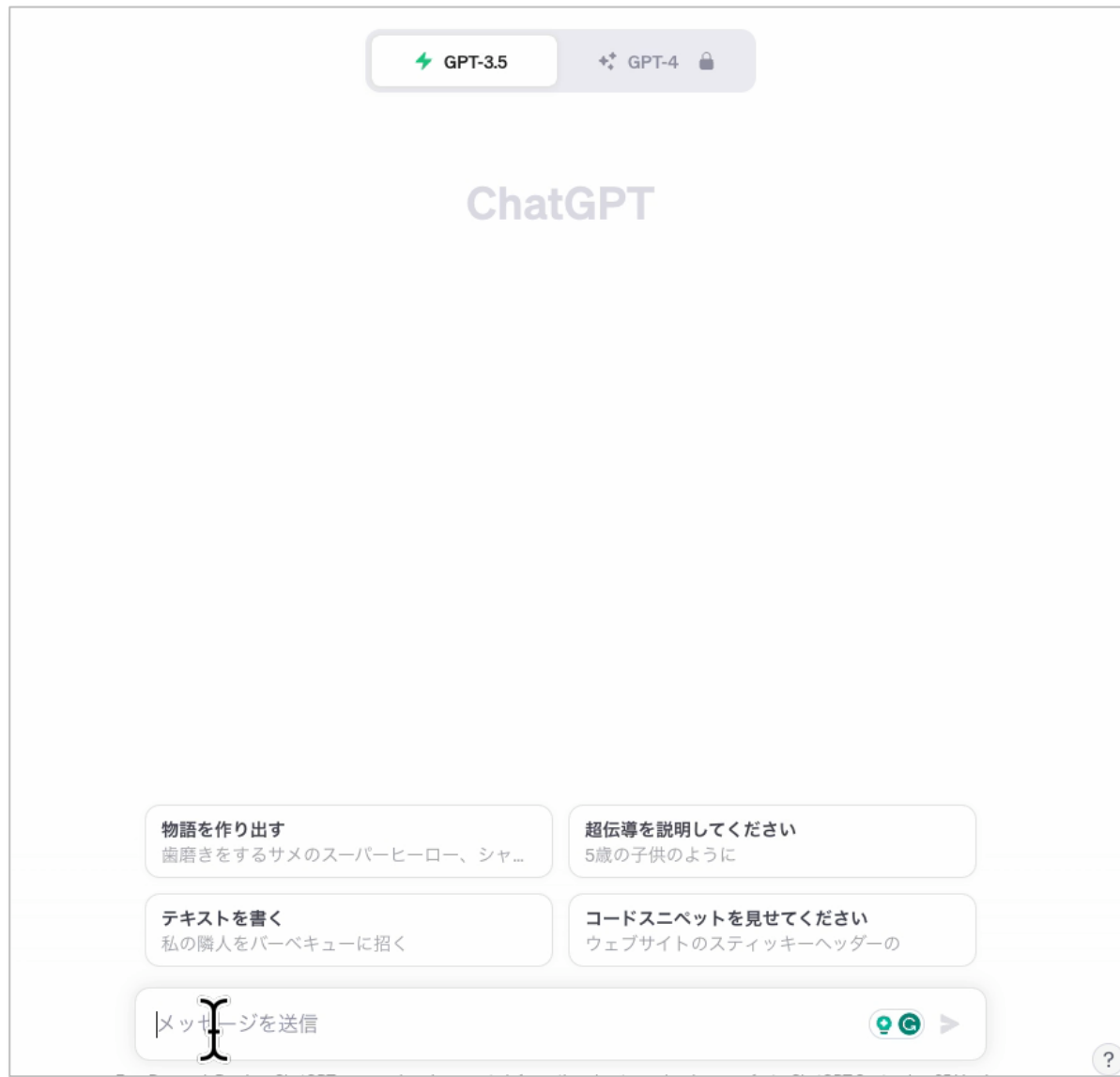
LLM-jpでの活動:言語モデル構築

東北大学 言語AI研究センター
センター長・教授 鈴木潤

-
- 背景：対話型文章生成AI / 大規模言語モデル
(large language model: LLM)

代表例：ChatGPT

<https://chat.openai.com/>



従来の言語モデル：Language Model (LM)

● 文章の出現確率を予測する確率モデル

語彙
(単語の集合)

に
。
これ
です
今日
1
…
仙台
良い
東京
は
…
天気
…

$P(Y = \$BOSS\$ \text{今日は良い天気です。} \$EOS\$)$

確率 高

$P(Y = \$BOSS\$ \text{良いです。天気は今日} \$EOS\$)$

確率 低



「最近の」言語モデル

- 前提：ニューラル言語モデル
 - 文章の確率を深層ニューラルネットワークで近似
 - [文脈] => [正解次単語] を分類問題として学習
 - 前方の単語列 (文脈) $Y_{<j}$ を利用し次の単語 \hat{y}_j を予測 (繰り返す)

$$\hat{y}_j \sim P_{\theta}(y_j | Y_{<j})$$

「最近の」言語モデル

- 前提：ニューラル言語モデル

- 文章の確率を深層ニューラルネット
- [文脈] => [正解次単語] を分類問題
- 前方の単語列 (文脈) $Y_{<j}$ を利用し返す)

$$\hat{y}_j \sim P_{\theta}(y_j | Y_{<j})$$

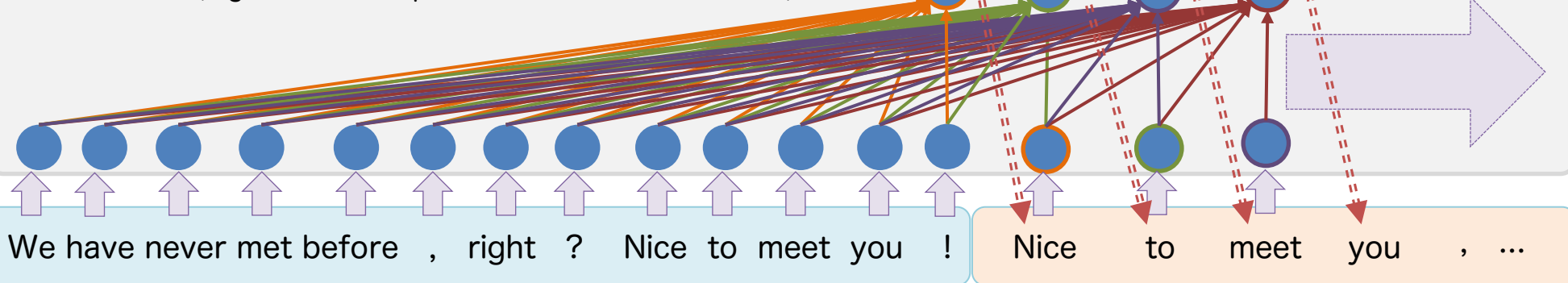
| Vocabulary | Vocabulary | Vocabulary | Vocabulary |
|------------|------------|------------|------------|
| A | A | A | A |
| this | this | this | this |
| that | that | that | that |
| ... | ... | ... | ... |
| meet | meet | meet | meet |
| have | have | have | have |
| you | you | you | you |
| ... | ... | ... | ... |
| Nice | Nice | Nice | Nice |
| ... | ... | ... | ... |
| to | to | to | to |
| ... | ... | ... | ... |
| too | too | too | too |
| , | , | , | , |
| . | . | . | . |

則 (繰り

OUTPUT text~>

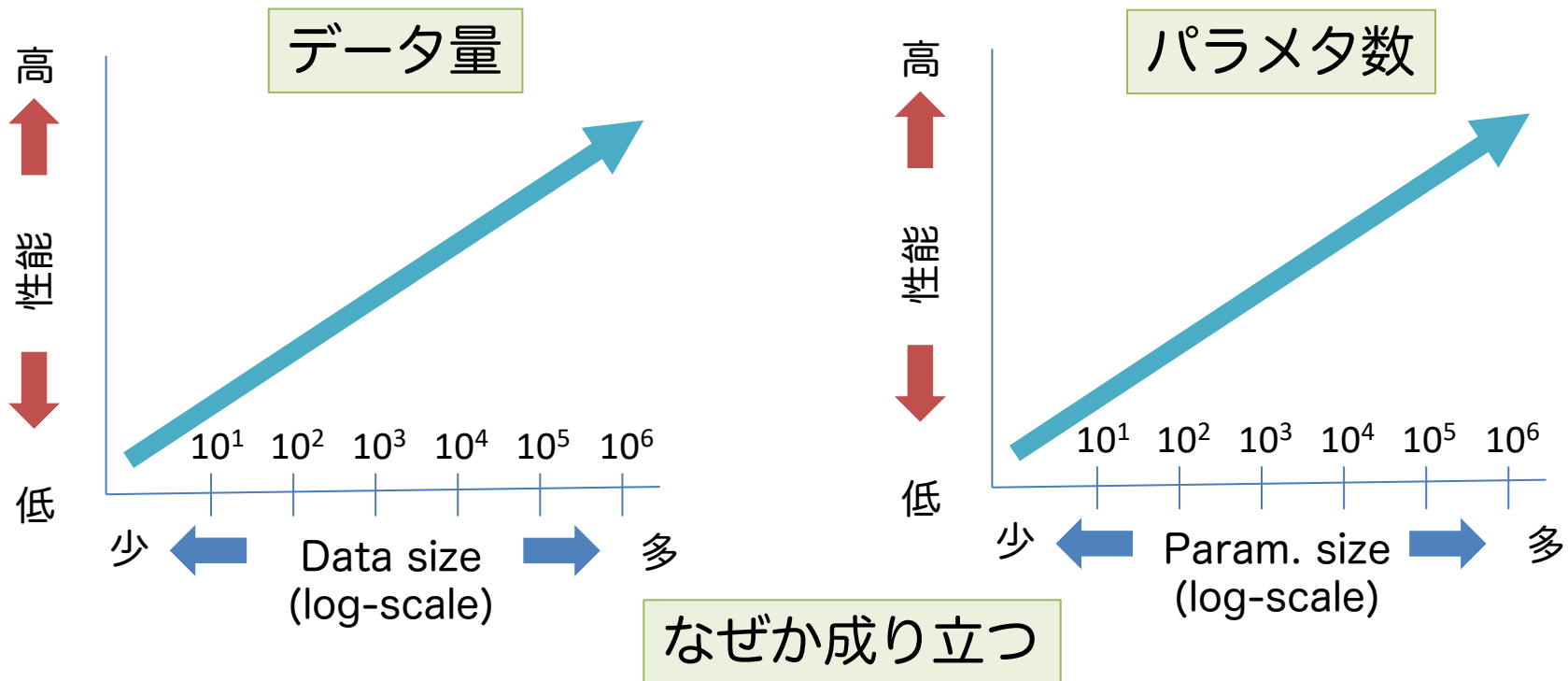
Nice to meet you , too ...

Neural LM (e.g., Generative pre-trained transformer: GPT)



ニューラル言語モデルのスケール則

- 学習データ量/モデルパラメタ数を増やせば性能向上
 - データ量/パラメタ数の対数スケールで線形の関係



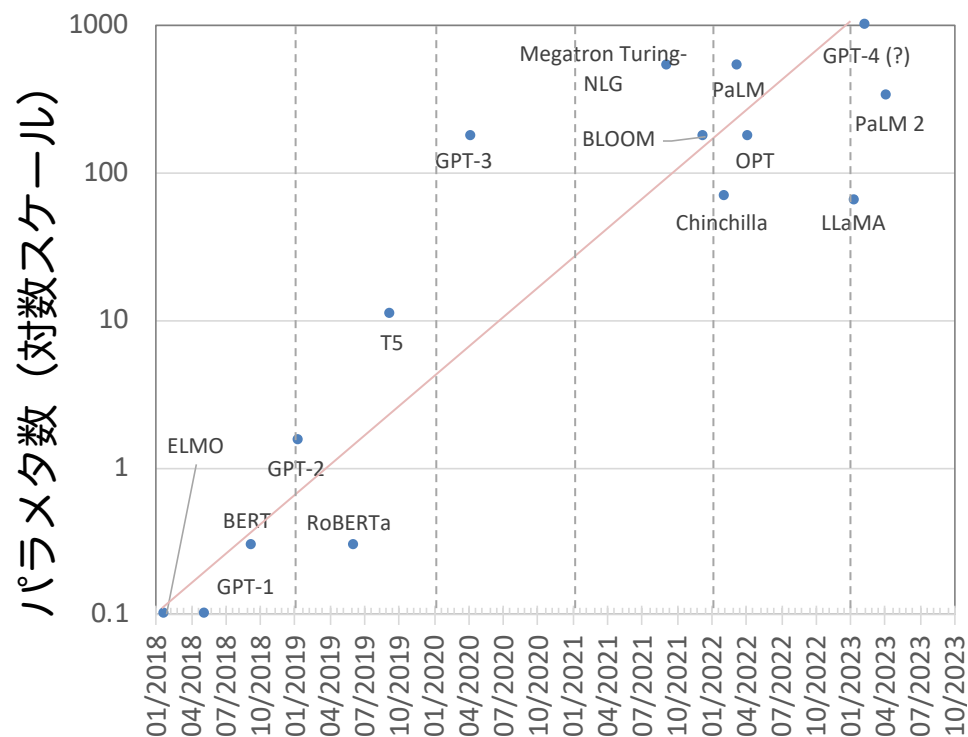
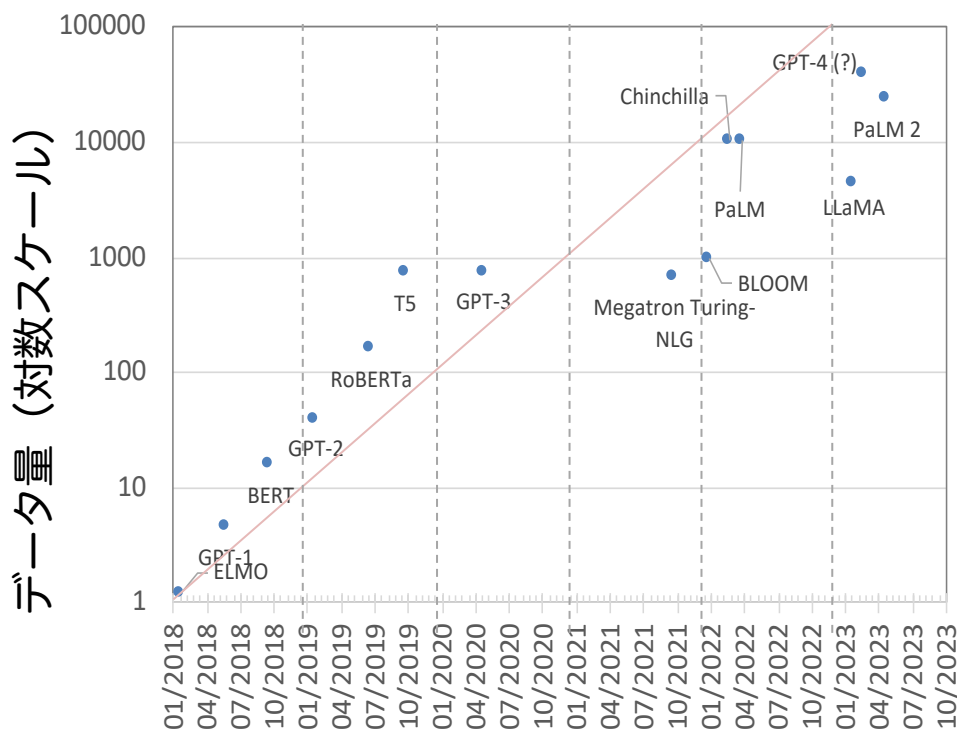
Large language model (LM) => 大規模計算環境が必須に

[参考] 年毎の大規模化

- 学習データ量 / パラメタ数 : 年々順調に増加

| | トークン数 | ファイル容量 |
|--------|----------|----------------|
| GPT-2: | 10 B | 40 GB |
| GPT-3: | 500 B | 750 GB |
| GPT-4: | 13,000 B | 20,000 GB (憶測) |

| | パラメタ数 | 必要メモリ量 |
|--------|---------|------------------|
| GPT-2: | 1.5 B | 6 GB |
| GPT-3: | 175 B | 700 GB (ChatGPT) |
| GPT-4: | 1,800 B | 7,200 GB (憶測) |



-
- LLM-jpの取り組み紹介

言語モデルの状況 => 英語主体

- 学習データの大半が英語データ

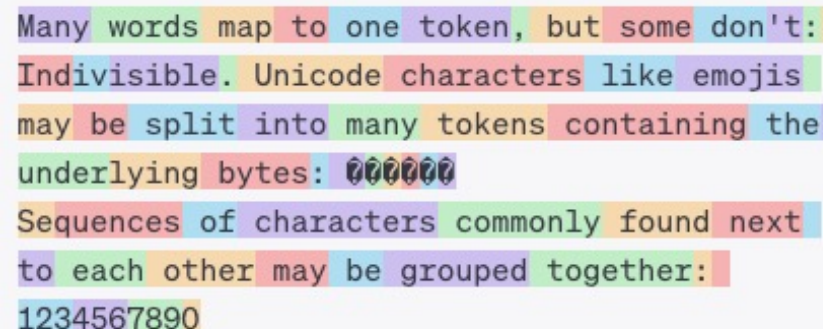
- GPT-3の日本語データはわずか0.1%

| | language | number of words | percentage |
|----|-----------------|-----------------|------------|
| 1 | English (en) | 181,014,683,608 | 92.65% |
| 2 | French (fr) | 3,553,061,536 | 1.82% |
| 3 | Germany (de) | 2,870,869,396 | 1.47% |
| 4 | Spanish (es) | 1,510,070,974 | 0.77% |
| 5 | Italiano (it) | 1,187,784,217 | 0.61% |
| 6 | Portuguese (pt) | 1,025,413,869 | 0.52% |
| 7 | Dutch (nl) | 669,055,061 | 0.34% |
| 8 | Russian (ru) | 368,157,074 | 0.19% |
| 9 | Romanian (ro) | 308,182,352 | 0.16% |
| 10 | Polish (pl) | 303,812,362 | 0.16% |
| 11 | Finnish (fi) | 221,644,679 | 0.11% |
| 12 | Danish (da) | 221,551,540 | 0.11% |
| 13 | Swedish (sv) | 220,920,577 | 0.11% |
| 14 | Japanese (ja) | 217,047,918 | 0.11% |
| 15 | Norwegian (no) | 212,193,299 | 0.11% |
| 16 | Chinese (zh) | 193,517,396 | 0.10% |
| 17 | Czech (cs) | 139,918,438 | 0.07% |
| 18 | Hungarian (hu) | 127,224,375 | 0.07% |

- 処理単位 (トークン) の区切りが細かすぎる

<https://platform.openai.com/tokenizer>

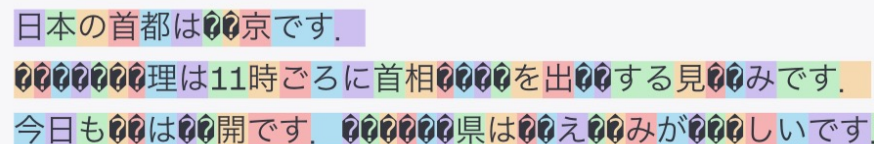
英語



Many words map to one token, but some don't: Indivisible. Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🍌🍌🍌🍌🍌. Sequences of characters commonly found next to each other may be grouped together: 1234567890

日本語

日本の首都は東京です。
岸田総理は11時ごろに首相官邸を出発する見込みです。
今日も桜は満開です。神奈川県は冷え込みが激しいです。



日本の首都は東京です。
岸田総理は11時ごろに首相官邸を出発する見込みです。
今日も桜は満開です。神奈川県は冷え込みが激しいです。

https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

LLMに関する議論

(by NII所長 黒橋先生)

- 何が起きているのか自然言語処理の研究者もわからない
- 大規模言語モデルの研究開発が一部の組織の寡占状態であることは健全とはいえない



- 完全にオープンなモデルを継続的に構築し「LLMの原理解明」多分野展開などの研究開発を進めることが必要
- 日本語の情報を十分にカバーし 使用ルール・入力情報の機密性がコントロール可能 等の要件も経済安全保障的な観点から必須
- 一方で、LLMの研究開発には大規模かつ多様なデータ、強力な計算基盤、多額の費用が必要 => もはやビッグ・サイエンス

- All Japanで産学の自然言語処理と関連分野の800+名規模(2023.12.12) の研究者/技術者が参加 (産学半々程度)
- オープンかつ日本語に強い大規模モデルを構築

- LLMの原理解明
- モデル・データ・ツール・技術資料・試行錯誤・議論の過程等・失敗も含め**全て公開**
=> 技術者の教育・参考資料としても有益

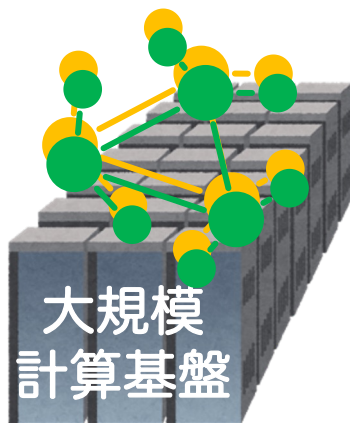


LLM勉強会：多くのWGにより構成

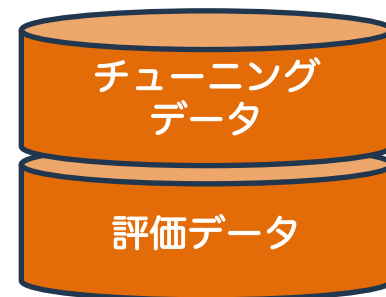
コーパス検討WG



モデル構築検討WG



チューニング検討WG (評価データ含む)



- 新WG

- 安全性検討WG

- サブグループなど

- トークナイザー検討
- コーパス検索
- チューニング実験


その他

- 法律系相談
- wandb-support
- データ流通・標準化
など

目標 (開始時: 2023.05)

- 今年の秋冬頃に妥当な性能をもつ13Bパラメタモデルの事前学習済み日本語言語モデル (Version1) を完成 (学習データ量: 270B tokens 予定)
=> この作成経験を踏まえて次のフェーズを考える

○ [参考] GPT-3 [\[2005.14165\] Language Models are Few-Shot Learners](https://arxiv.org/abs/2005.14165)



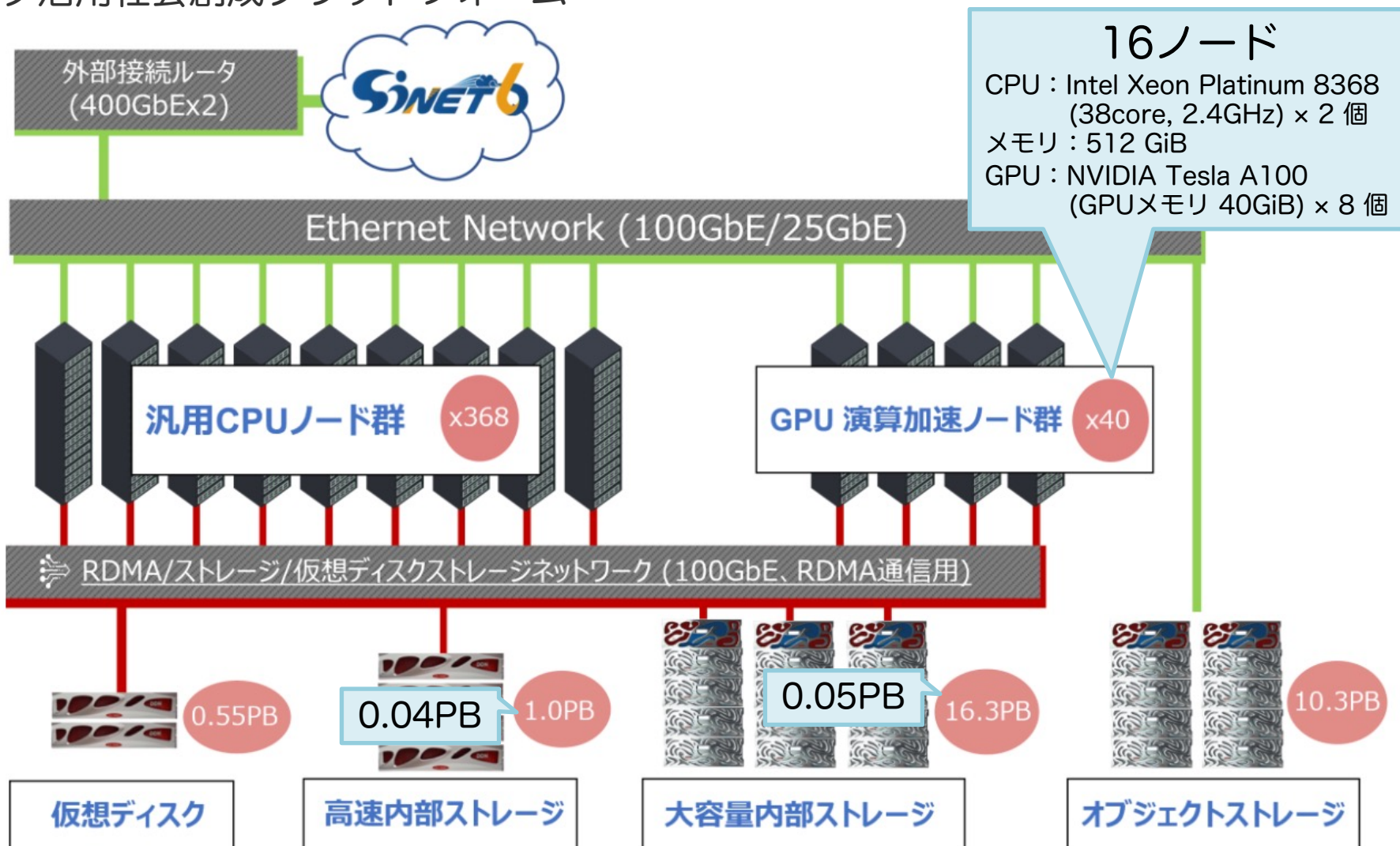
| Model Name | n_{params} | n_{layers} | d_{model} | n_{heads} | d_{head} | Batch Size | Learning Rate |
|-----------------------|---------------------|---------------------|--------------------|--------------------|-------------------|------------|----------------------|
| GPT-3 Small | 125M | 12 |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

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

利用計算機環境：mdx

<https://mdx.jp/about/system>

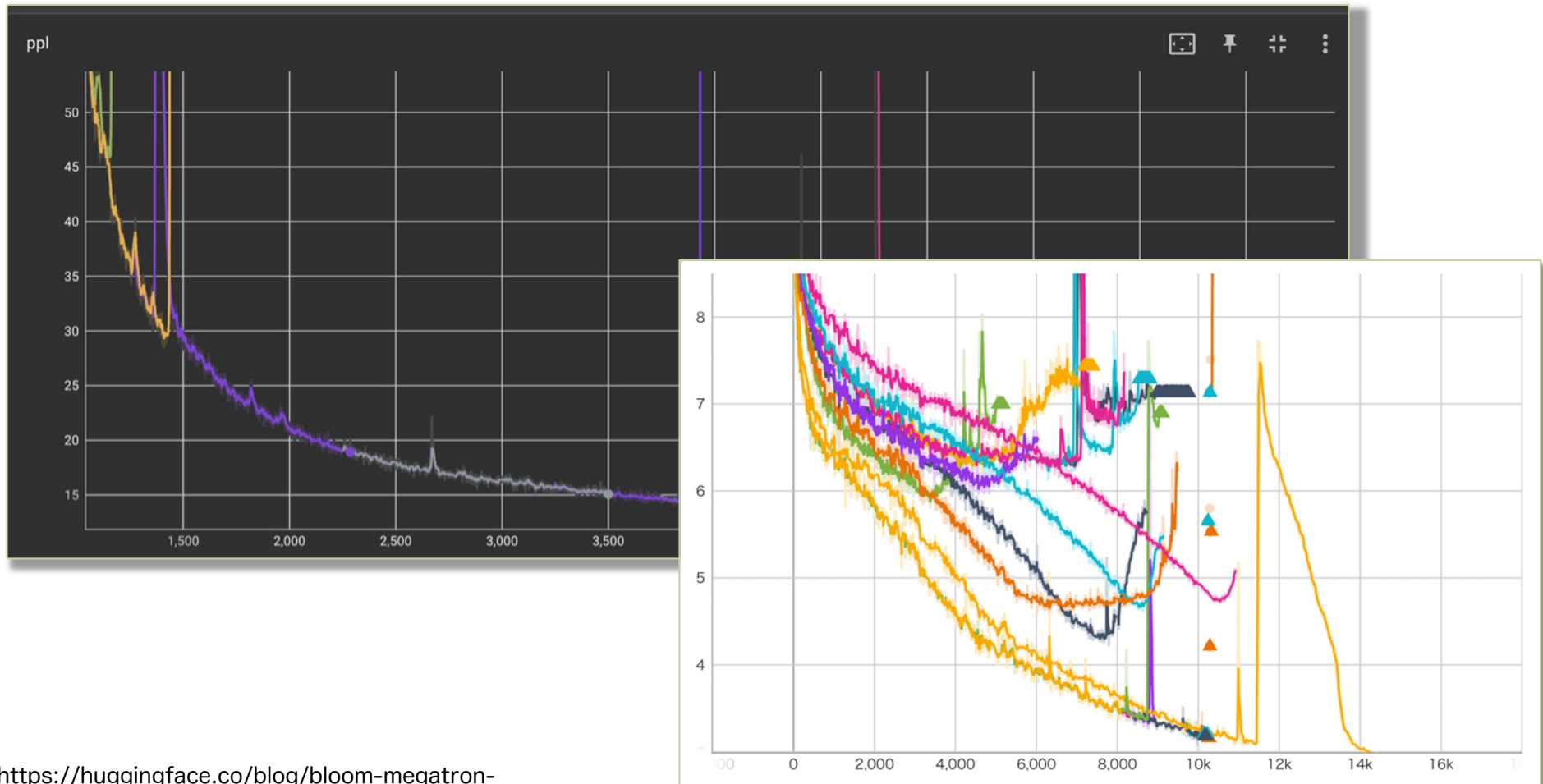
データ活用社会創成プラットフォーム



LLMの事前学習

- LLMの事前学習はそれほど単純ではない

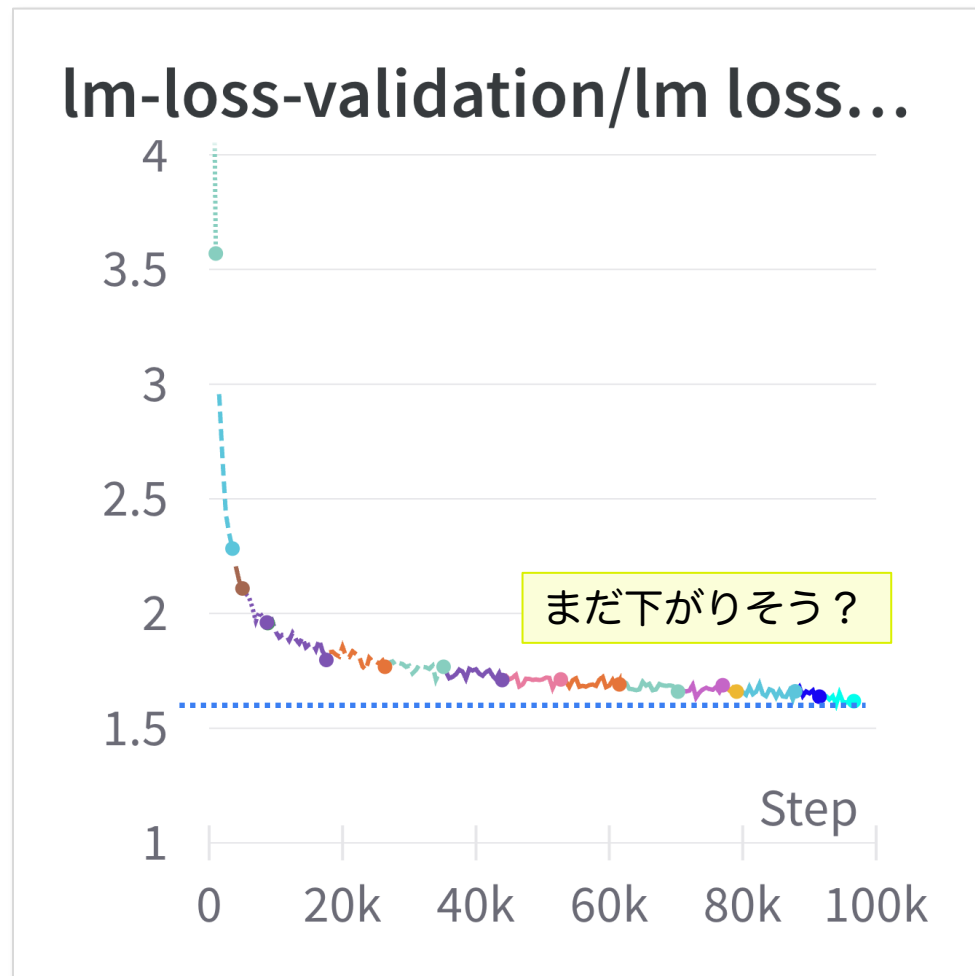
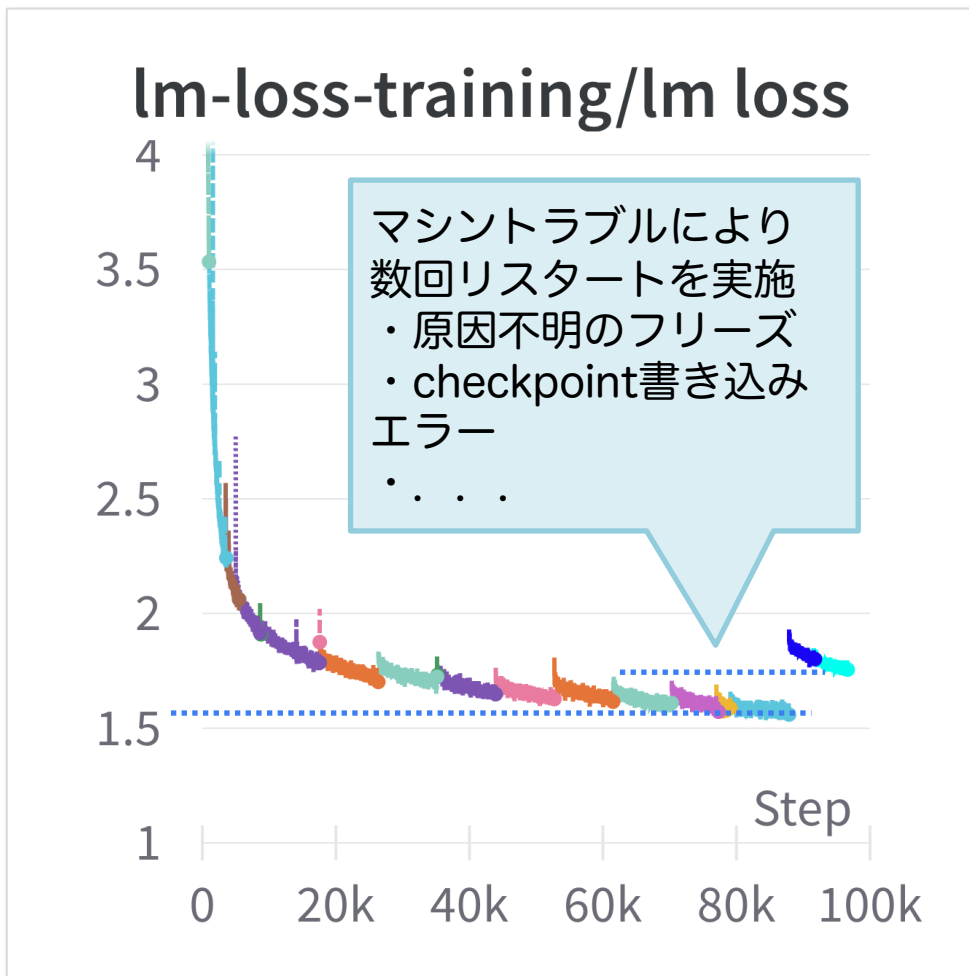
https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/chronicles/10_percent_update.md



<https://huggingface.co/blog/bloom-megatron-deepspeed#bf16optimizer>

モデル事前学習の学習曲線

- 13Bパラメタモデル：(トラブルなしで約18日)



LLM-jp Models Version1公開 (2023.10.20)

<https://huggingface.co/llm-jp>

llm-jp University

Organization Card

This repository provides large language models developed by **LLM-jp**, a collaborative project launched in Japan.

Finetuning済

Model Variant

- [llm-jp-13b-instruct-full-jaster-v1.0](#)
- [llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0](#)
- [llm-jp-13b-instruct-full-dolly-oasst-v1.0](#)
- [llm-jp-13b-instruct-lora-jaster-v1.0](#)
- [llm-jp-13b-instruct-lora-jaster-dolly-oasst-v1.0](#)
- [llm-jp-13b-instruct-lora-dolly-oasst-v1.0](#)

Pre-trained models

- [llm-jp-13b-v1.0](#)
- [llm-jp-1.3b-v1.0](#)

事前学習済

- Model size: 13B params
- Data size: 300B tokens
- Vocab size: 50k tokens
- Tool: Megatron-DeepSpeed


事前学習済みモデルとしては13Bと1.3Bのモデルを公開 (MDS形式も公開)

- [llm-jp/llm-jp-1.3b-v1.0](#)
Text Generation • Updated Oct 20 • ↓ 1.91k • ♥ 4
- [llm-jp/llm-jp-13b-v1.0](#)
Text Generation • Updated Oct 20 • ↓ 2.57k • ♥ 28
- [llm-jp/llm-jp-1.3b-v1.0-mdsfmt](#)
Text Generation • Updated Oct 20
- [llm-jp/llm-jp-13b-v1.0-mdsfmt-it87870](#)
Text Generation • Updated Oct 20
- [llm-jp/llm-jp-13b-v1.0-mdsfmt](#)
Text Generation • Updated Oct 20


実行例

Slack連携

https://drive.google.com/file/d/12BSt5IYPzt245BiYI_g17NcfA0DT7cl/view

 **Hirokazu Kiyomaru (京大黒橋研)** 1 minute ago
@llmjp-13b-finetuned 時間管理能力を向上させるにはどうしたらいいですか？

2 replies

 **llmjp-13b-finetuned** APP 1 minute ago
時間管理能力を向上させるにはいくつかのステップがあります：

- タスクをリスト化する：一日のタスクを書き出してリスト化し、優先順位をつけましょう。それによって、どのタスクに時間を割くべきか、またどのタスクをスキップすべきかを把握できます。
- 記録をつける：タスクの時間、場所、内容をメモしておきましょう。これは、あなたがそのタスクをいつ、どこで、何のために行ったかを明確にするためです。
- 優先順位をつける：タスクをリストアップしたら、時間とエネルギーが最も効率的に使える順に並び替えましょう。
- 休憩を取る：長時間の作業の後は、適度な休憩を取ることが重要です。これにより、エネルギーと集中力を持続させ、効率を上げることができます。
- 計画を立てる：タスクのスケジューリングは、時間管理の重要な部分です。あらかじめスケジューリングを行い、時間を有効に使うことができるようにしましょう。
- 優先順位を見直す：時間管理能力を向上させるためには、定期的にタスクの優先順位を見直すことが重要です。これは、重要度に基づいてタスクを優先順位付けすることにより行います。
- 自己反省する：時間管理を行う上で、自分自身を責める必要はありません。時間管理には時間がかかり、改善には時間を要します。自分自身を理解し、進歩を評価するために、定期的に反省することが重要です。

時間管理能力を向上させるには時間と努力が必要ですが、時間管理を上手に行うことで、より生産的でストレスの少ない生活が送れるようになります。



検索

ベクトル検索と N-gram ベースの検索、それぞれの上位3件の結果を返却

何番目の学習データに出てきたか（出てこなかったか）を確認可能

 **llmjp-search** APP 4 minutes ago
Search engine: Vector Search | Rank: 1 | Source: ja_cc | Iteration: 87,538/96,656

順位をつけて実行しやすくするといった作業が必要になります。こうしたタスクの量や手順を管理する作業がタスク管理です。タスク管理を行う必要性ビジネスパーソンにとってタスク管理は重要な作業です。業務中には多くのタスクが発生しますが、質の高い仕事をするためにはすべてのタスクを漏れなくさばいていかなければなりません。このとき順番でタスクをこなしておくのか決めていなかったり、作業中に他のタスクが割り込んできたりすると作業効率が落ちてしまいます。また、重大な仕事を忘れてしまうということもあります。あらかじめ「どのタスクをどのようにさばくか」という計画を練っておけば、日々の業務に迷いなく取り組めるようになります...

[See more](#)

Search engine: Vector Search | Rank: 2 | Source: ja_cc | Iteration: 79,211/96,656

解する、もしくは仕事の進捗状況やスケジュールを把握するのに欠かせない作業がタスク管理です。タスク管理ができるようになると効率よく業務をこなすことができるようになります。時間に余裕を持って作業できるぶん仕事のクオリティ向上も期待できます。上手にタスク管理を行なうためには、タスクの洗い出し、期限の設定、タスクの具体的な実行計画を立てるという3つのステップを踏むのが一般的です。まず、自分がやるべきタスクをすべて洗い出します。紙などに書いてタスクを視覚化し、それぞれのタスクの中身や仕事の全体量を把握しましょう。リストができたなら、タスクが実行可能な状態まで具体的にになっているかどうか、リストに重複や漏れ...

[See more](#)

Search engine: Vector Search | Rank: 3 | Source: ja_cc | Iteration: 12,385/96,656

「整理」、そして「まとめあげ」です。まずは、1日のうち、何にどれだけの時間を使っているのかを正確に記録しましょう。記録したら、1日の終わりに、自分の時間の使い方

-
- 今後の予定

次の目標 Version2 (10/20 当時)

- 今年度中に175Bパラメタ級モデルの学習を開始
- 来年度前半に175Bパラメタ級モデルを 2T tokens 級データで学習完了

=> 計算機環境の調達が必要 (mdxでは実施困難)

○ [参考] GPT-3 [\[2005.14165\] Language Models are Few-Shot Learners](#)

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

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

まとめ

- 背景：対話型文章生成AI / 大規模言語モデル (LLM)
- LLM-jp の取り組み
 - LLMに関する議論
 - オープンかつ日本語に強い大規模モデルを構築
 - 13Bパラメタモデル (Version1) の構築と公開
 - 2023.10.20
- 今後の予定
 - 175Bパラメタ級モデルの構築