

NLPとVision-and-Languageの 基礎・最新動向 (1)

NTT人間情報研究所

西田京介, 西田光甫, 田中涼太, 齊藤いつみ

2023年3月8日

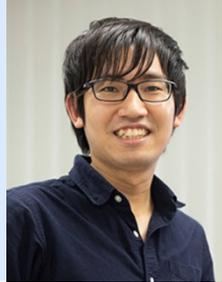
DEIM2023 第15回データ工学と情報マネジメントに関するフォーラム
チュートリアル講演



西田京介

専門：[機械読解（質問応答）](#) , [Vision-and-Language](#)

NLP'18 最優秀賞, '19優秀賞 (筆頭), '20優秀賞, '21最優秀賞, '22優秀賞 (共著), DBSJ上林奨励賞(2017)など. JSAI'22, NLP'19など 招待講演



西田光甫

専門：[自然言語理解・説明](#)

NLP'22 優秀賞, NLP'20 若手奨励賞 (筆頭), HotpotQA, FEVERにて投稿時首位獲得



田中涼太

専門：[視覚的機械読解](#)

NLP'21最優秀賞, NLP'22若手奨励賞, ICDAR 2021 Infographics VQA runners-up (筆頭), AI王ワークショップ'21 招待講演



齊藤いつみ (※本日欠席)

専門：[文書要約](#)

JSAI'20優秀賞, NLP'19最優秀ポスター (筆頭), 言語処理セミナー'20 招待講演



西田京介

1. 深層学習による自然言語処理技術の基礎 (40min)

タスクとモデルの分類, Transformerとセルフアテンション,
BERTとGPT-3, 大規模言語モデル (LLM) に関する知見



西田光甫

2. 大規模言語モデルの革新 (40min)

InstructionチューニングとRLHF, ChatGPT,
ChatGPTの評価と言語モデルの今後の課題

質疑応答 (10min)



西田京介

3. Vision-and-Languageへの派生 (40min)

Vision Transformer, 視覚・言語の基盤モデルCLIP,
CLIP+LLMによるモデル: Flamingo, KOSMOS-1など



田中涼太

4. 文書や画面の理解 (40min)

処理・タスク・モデルの紹介,
文書画像理解の発展: SlideVQA, LiLT, Pix2Struct, ReActなど

質疑応答 (10min)



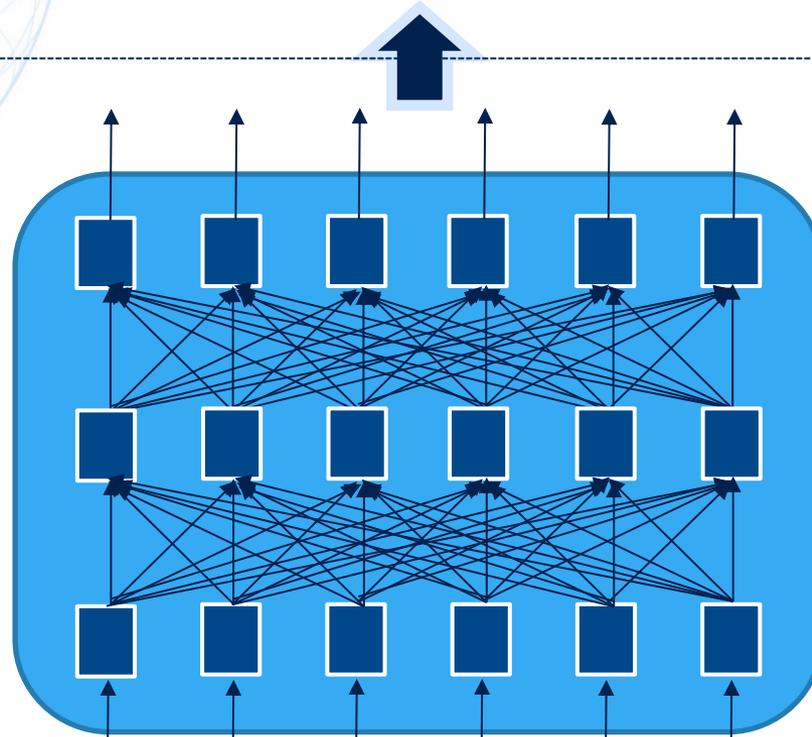
言語処理タスクと モデルアーキテクチャの分類

“深層学習時代の”自然言語処理の概要

出力情報

分類ラベル, 生成テキスト, 特徴量 (ベクトル), etc.

ベクトル系列



ベクトル系列から
所望の情報へ

ニューラルネットワーク

ベクトル系列の変換を
繰り返し行う

各トークンを
ベクトルへ“埋め込み”

トークン (単語や
サブワード) 化

ベクトル系列

トークン系列

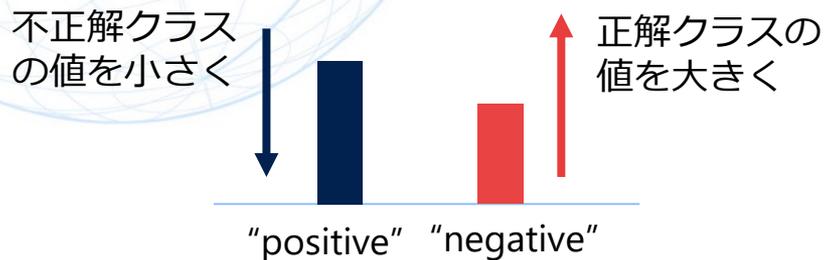
<S> メロ スは 激怒 ... した

入力テキスト

“メロスは激怒した。必ず、かの邪智暴虐…決意した”

学習のイメージ

- ニューラルネットワークの出力を正解に近づけるように学習を行う
- 自然言語処理タスクの多くは「分類問題」として学習される

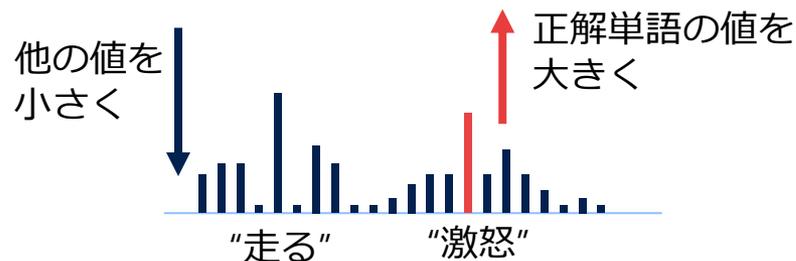


2次元ベクトルを出力



“メロスは激怒した。”

タスク：テキストを2つのクラスに分類



語彙サイズの次元数のベクトルを出力



“メロスは”

タスク：入力に続く次の単語を生成

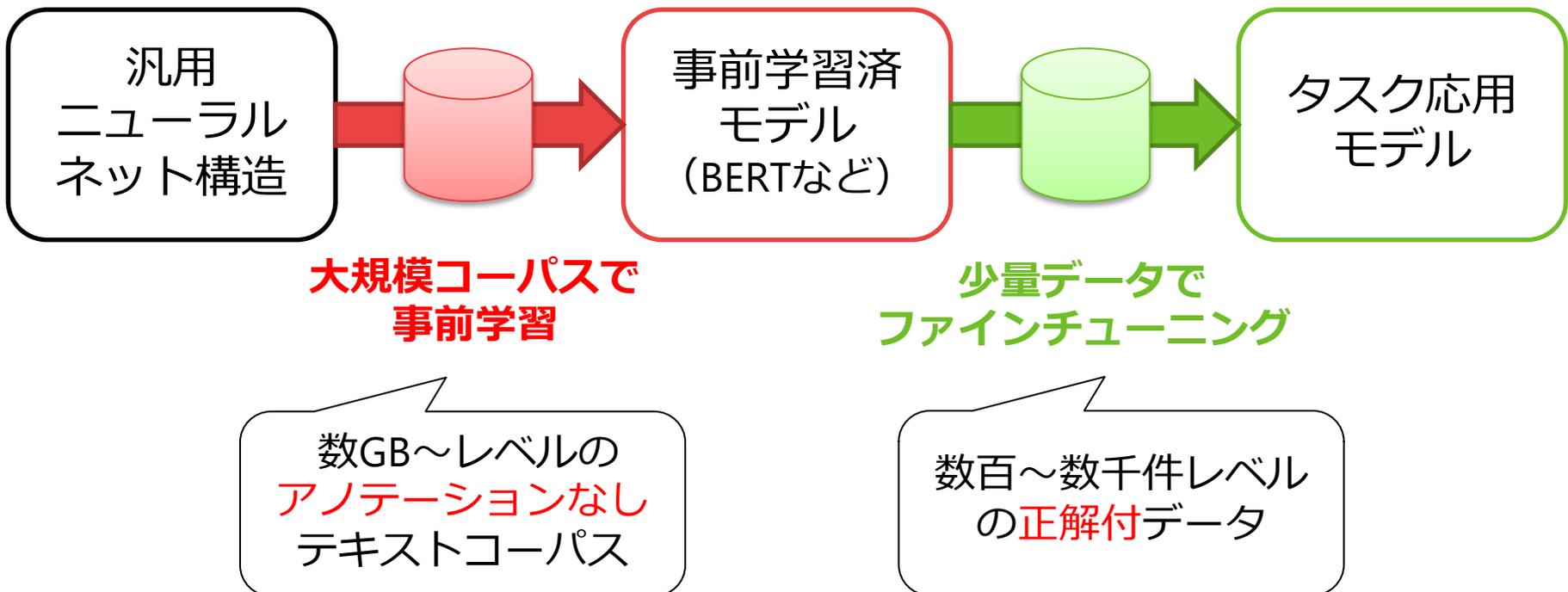
“BERT(2018)以前の”自然言語処理

- タスク毎に適したニューラルネットワークを設計し，正解情報付きのデータセットで学習していた
- 深層学習により性能は高まったが，学習データが少ないと高い精度は実現しにくい課題が残っていた



“BERT以降”: 事前学習とファインチューニング

- (1) 大規模なコーパスに基づく**事前学習**により汎用モデルを作り,
(2) **ファインチューニング**により各タスクに適応する方式が主流に
- タスクごとにモデルを設計することは少なくなった. また, タスクは2種類に大別して語られることが多くなった



自然言語処理タスクの分類

- **自然言語理解 (NLU)** と **自然言語生成 (NLG)** の2種類に大きく分類
- タスクごとに適したアーキテクチャがある

自然言語理解

Natural Language Understanding

テキストの意味を理解

- テキスト分類
- 情報検索
- 感情分析
- 固有表現抽出
- Etc.

自然言語生成

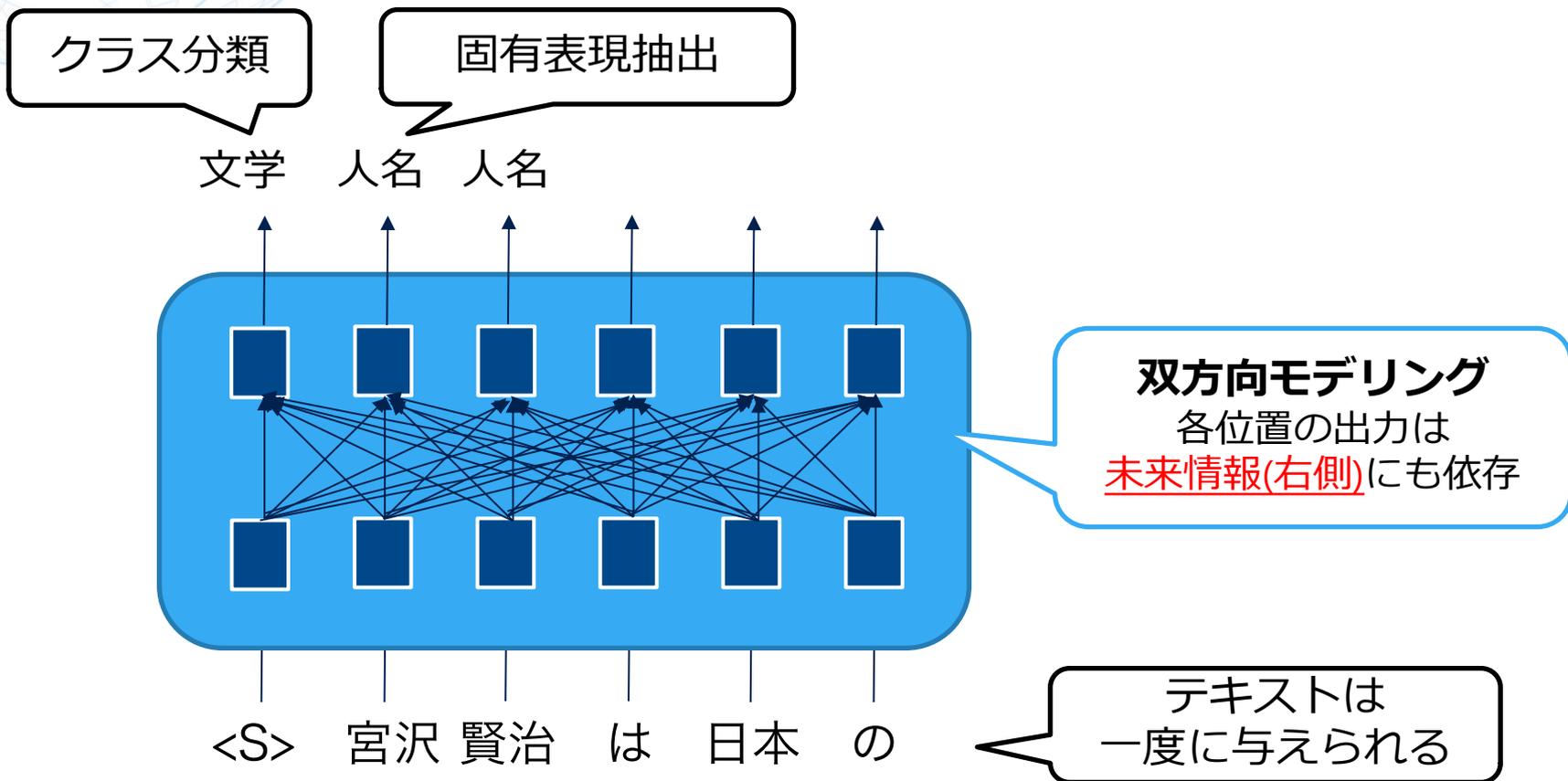
Natural Language Generation

入力からテキストを生成

- 機械翻訳
- 生成型要約
- 発話生成
- 言い換え
- Etc.

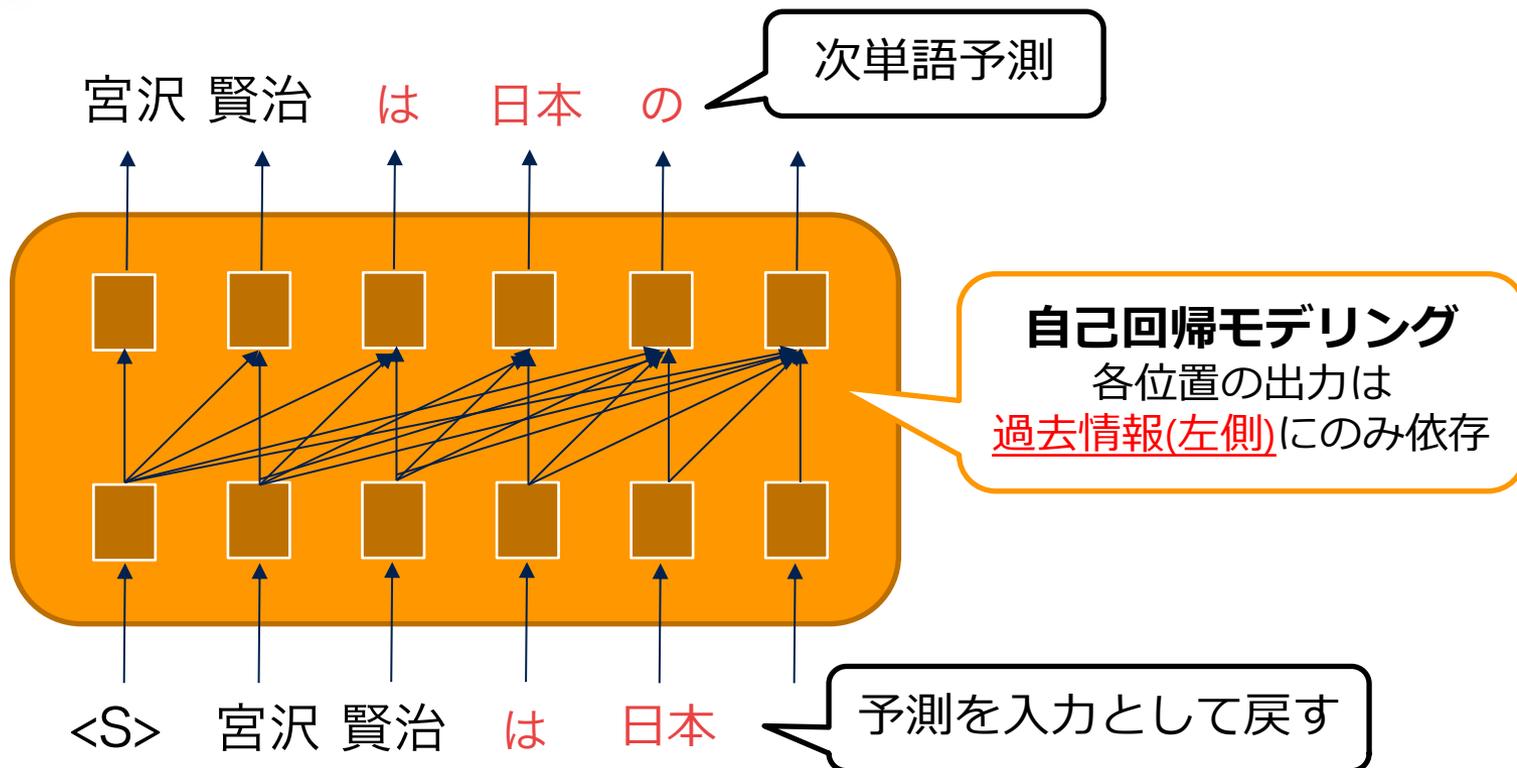
NLUタスクとモデルのイメージ

- **エンコーダ**型のモデルを用いて、テキスト（単語系列）のクラス分類、各単語へのラベリングなどを行う
- 代表モデル：BERT [Devlin+, NAACL'19]



NLGタスクとモデルのイメージ (1/2)

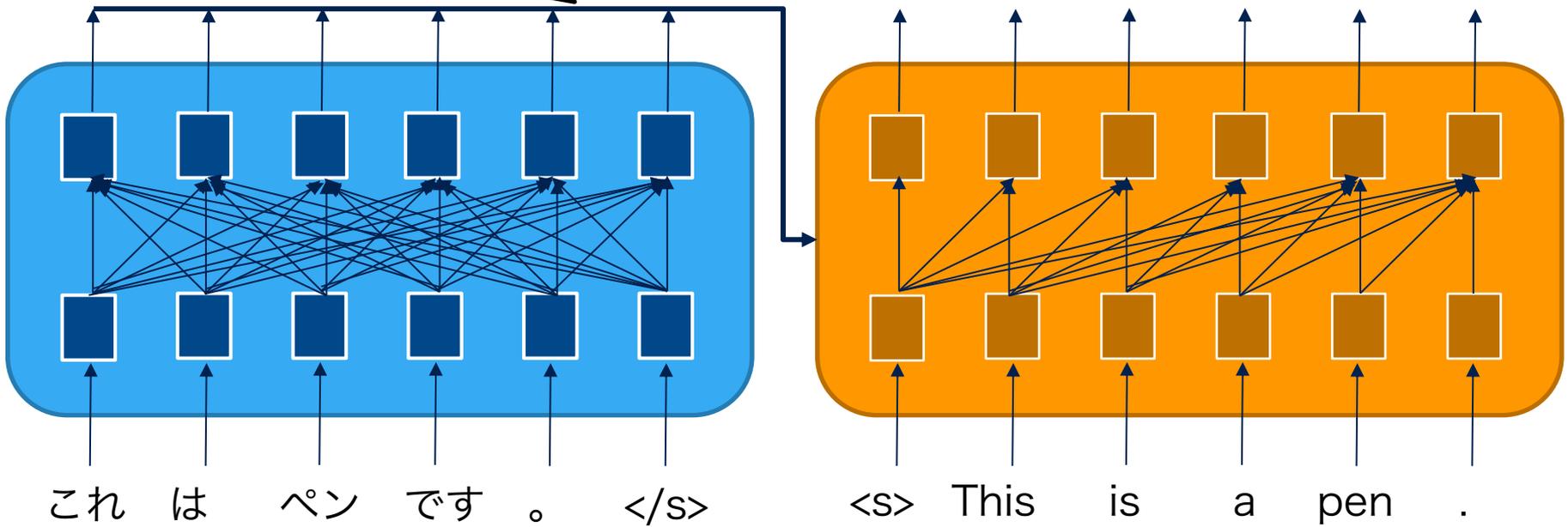
- **デコーダ**型のモデルを用いて, 入力系列の**続き**を生成したり, テキストAからテキストBへの変換を行う
- 最近はこのアーキテクチャが主流 (NLUもこれで解く)
- 代表モデル: GPT-3 [Brown+, NeurIPS'20]



NLGタスクとモデルのイメージ (2)

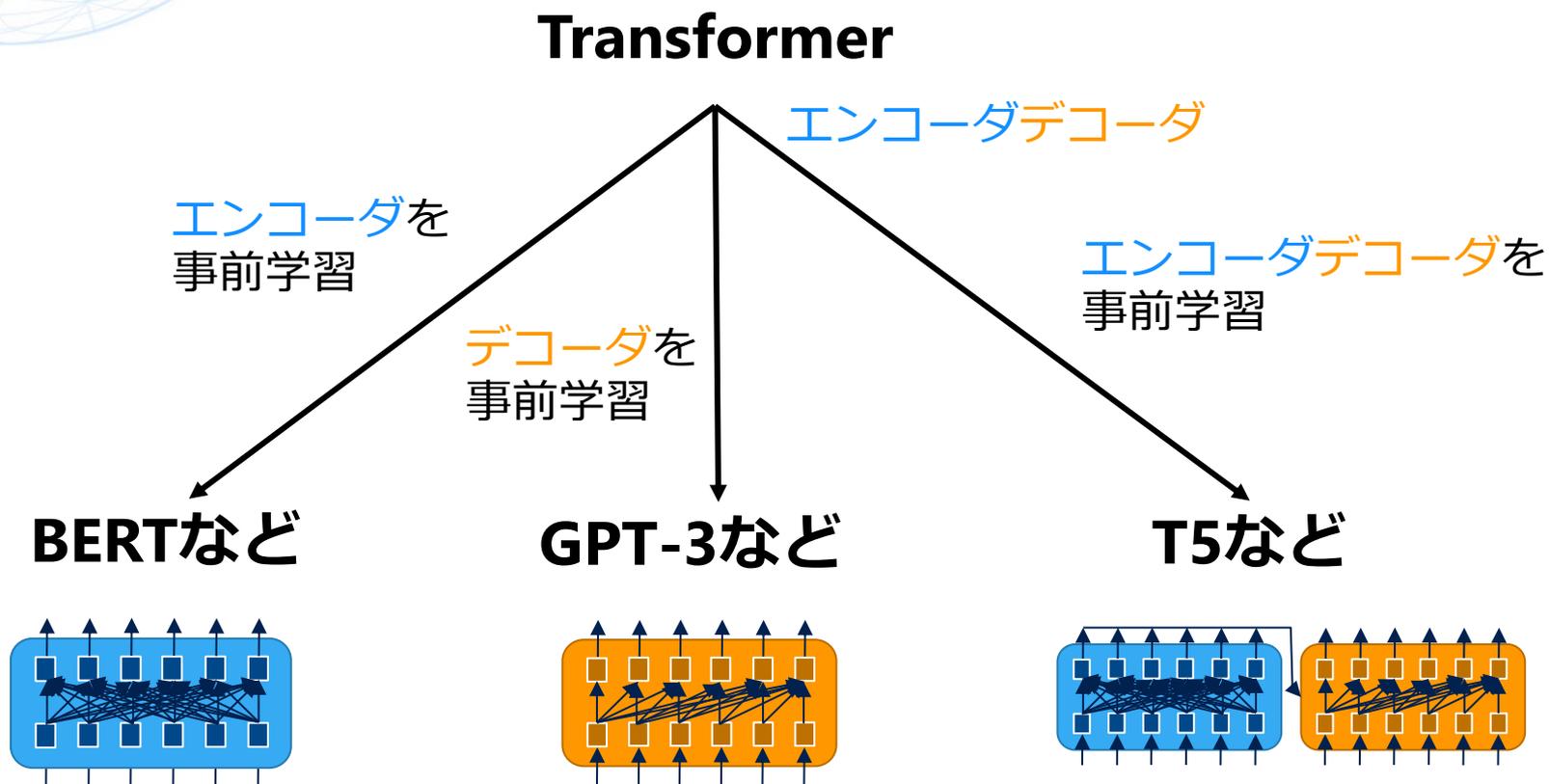
- **エンコーダデコーダ**型のモデルを用いて, テキストAからテキストBに変換 (翻訳や要約など) する
- 代表モデル: Transformer [Vaswani+, NIPS'17], T5 [Raffel, JMLR'21]

エンコーダの
出力をデコーダで利用



最近のベースモデル = Transformer

- 近年の事前学習済み言語モデルの多くが殆どが Transformer ベースとなっている
- ビジョン&ランゲージにもTransformerの事前学習が派生している（後半で説明）

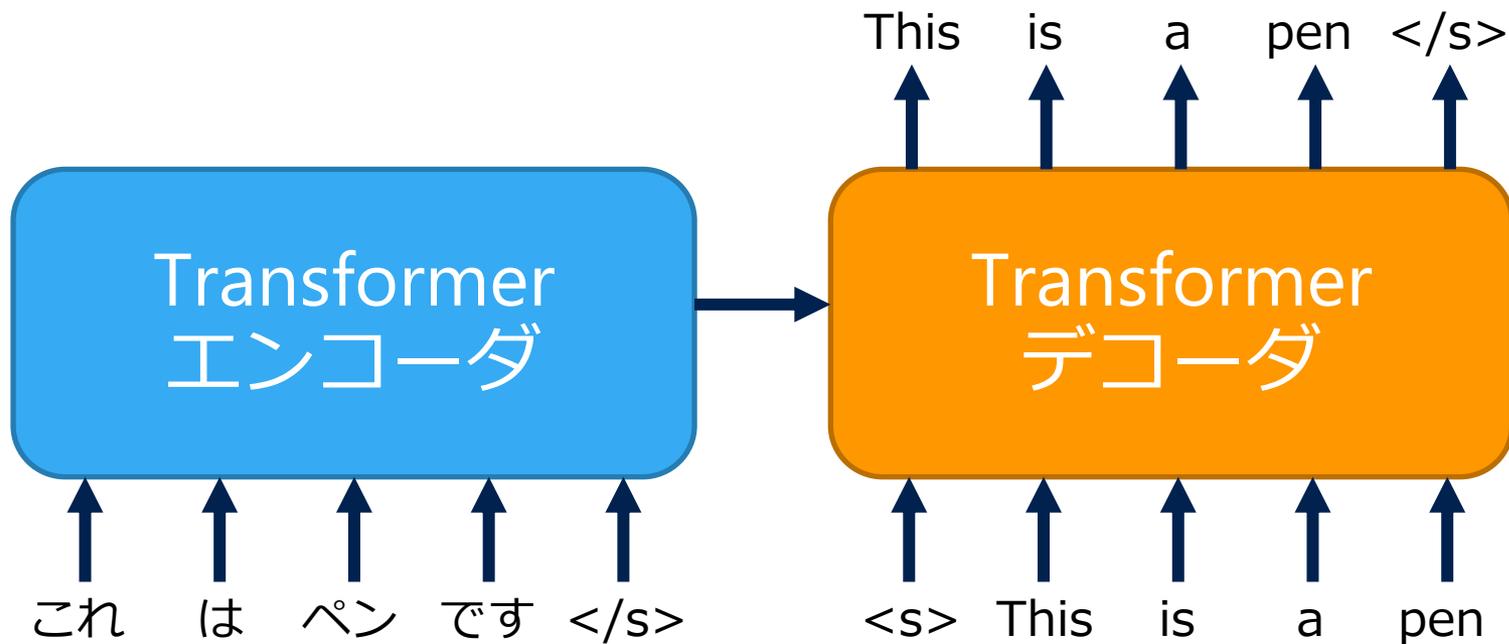




Transformer と セルフアテンション

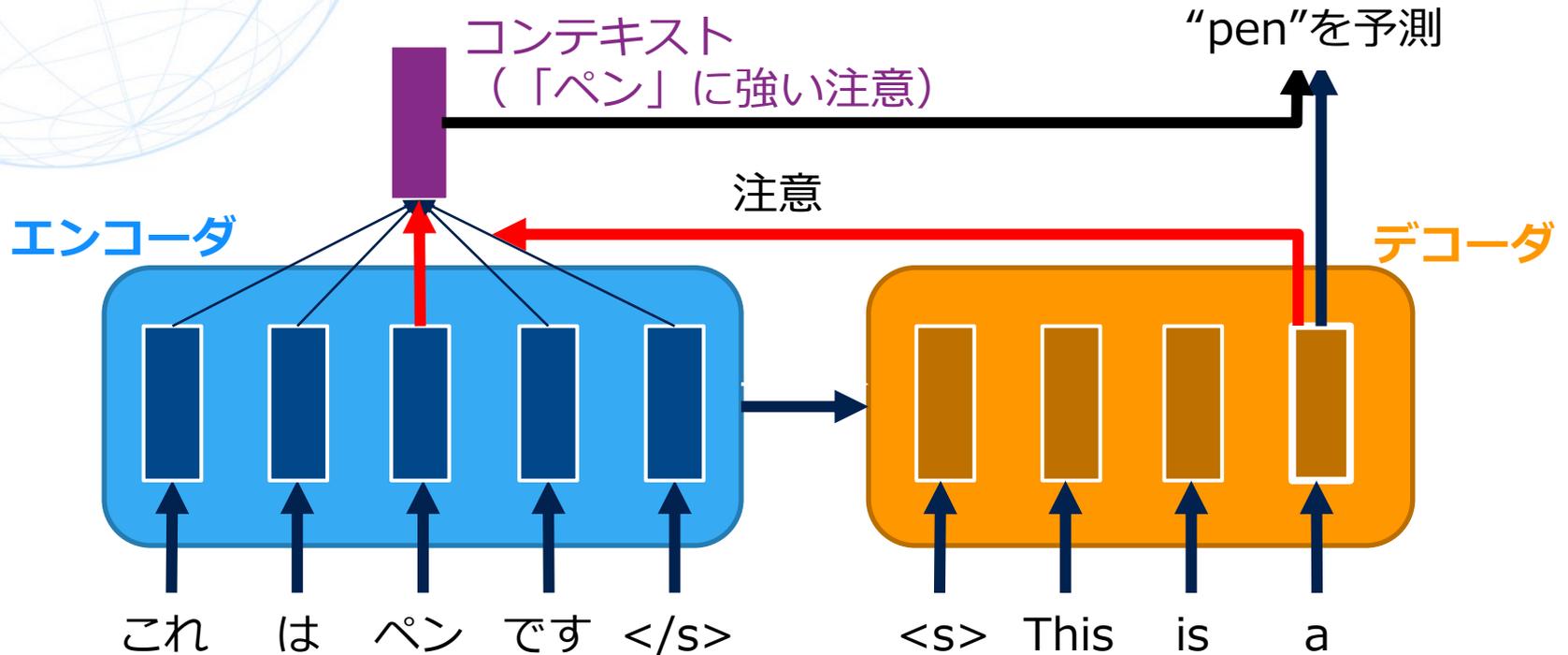
Transformer [Vaswani(Google)+, NIPS'17]

- 単語間の関係を理解（表現）するためにRNNやCNNを用いず，**アテンション**のみを用いたエンコーダデコーダ型モデルを提案
- 機械翻訳で評価実験し，当時の最高性能を更新



アテンション [Bahdanau(Jacobs U)+, ICLR'15]

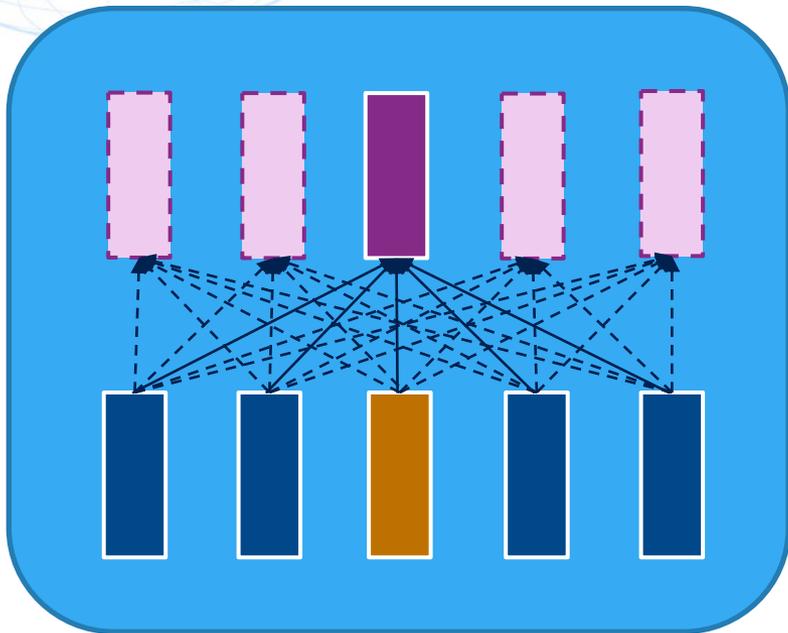
- デコーダが1単語ずつ出力するとき、エンコードされた翻訳元テキストのどの部分に注意すれば良いかを与える機構（行列計算で実現）



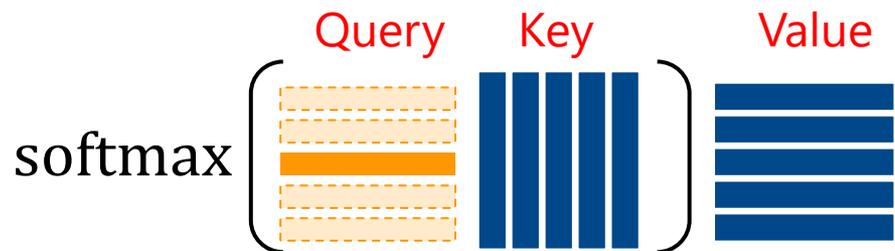
$$\text{softmax} \left(\begin{matrix} \mathbb{R}^{1 \times d} \\ \mathbb{R}^{d \times n} \end{matrix} \right) \begin{matrix} \mathbb{R}^{n \times d} \\ \text{Encoderの隠れ状態の} \\ \text{重み付き平均 } \mathbb{R}^{1 \times d} \end{matrix} = \text{purple bar}$$

セルフアテンション

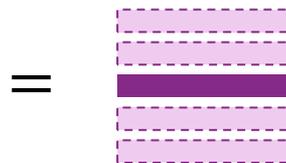
- エンコーダデコーダ間のクロスアテンションに加えて、Transformerでは系列単体でもアテンションを取る
- 系列の各位置を**Query**, 系列全体を**Key**, **Value**として考える



Key-Value辞書から、Queryに近いKeyのValueを取り出すイメージ



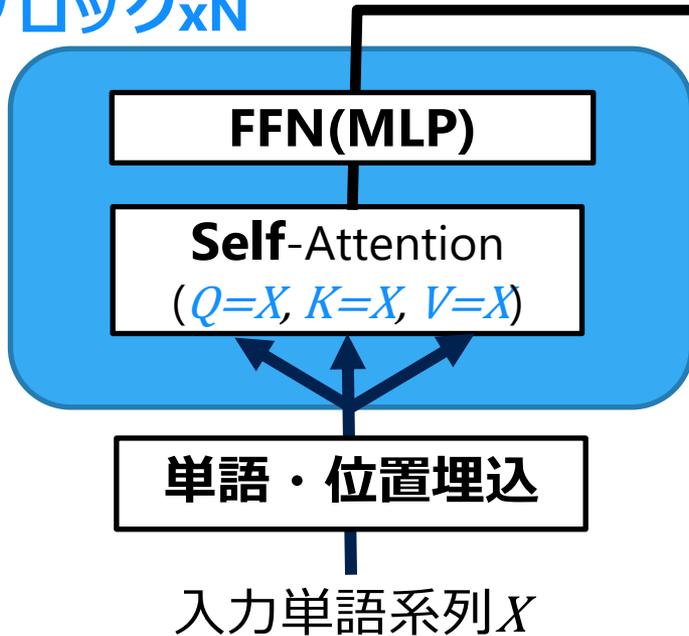
Valueの重み付き平均
 $\text{softmax}(QK^T)V$

$$= \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array}$$


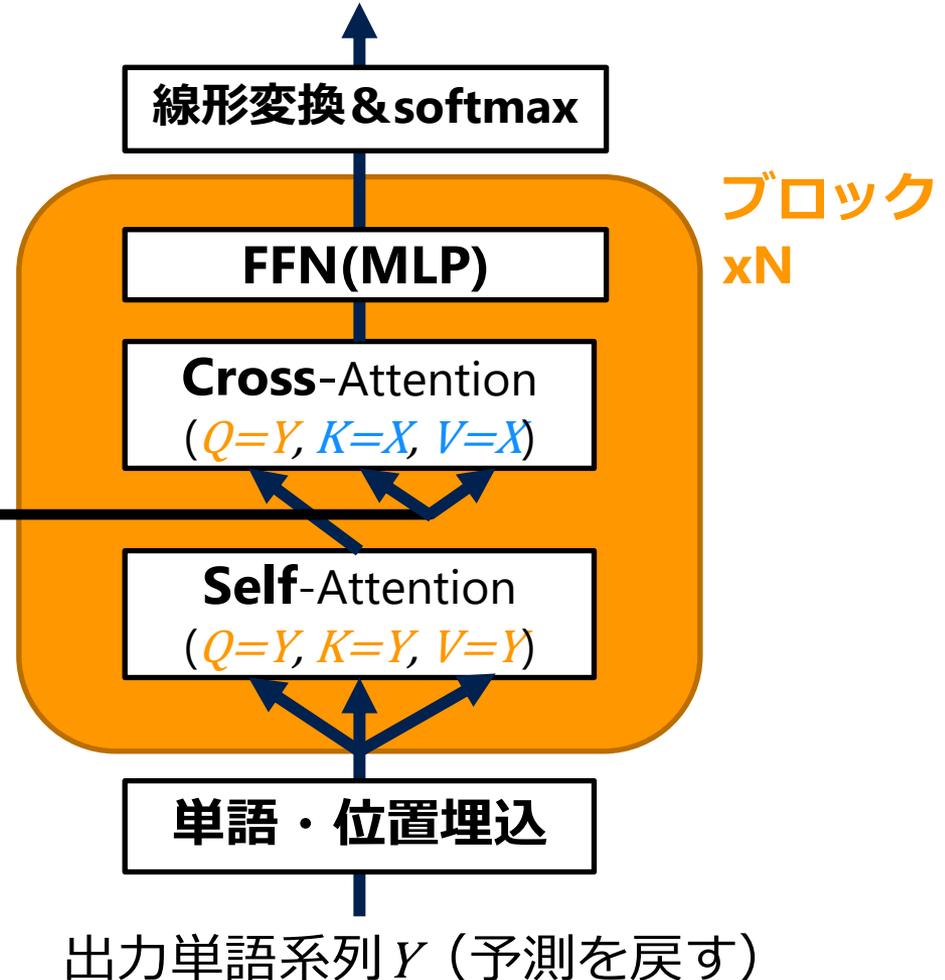
Transformerの簡略化イメージ

- エンコーダデコーダの双方でセルフアテンションを導入
- ブロックを多層に積上げることで性能向上
- 今日の説明は超簡易版なので詳細は論文等にて！

ブロックxN

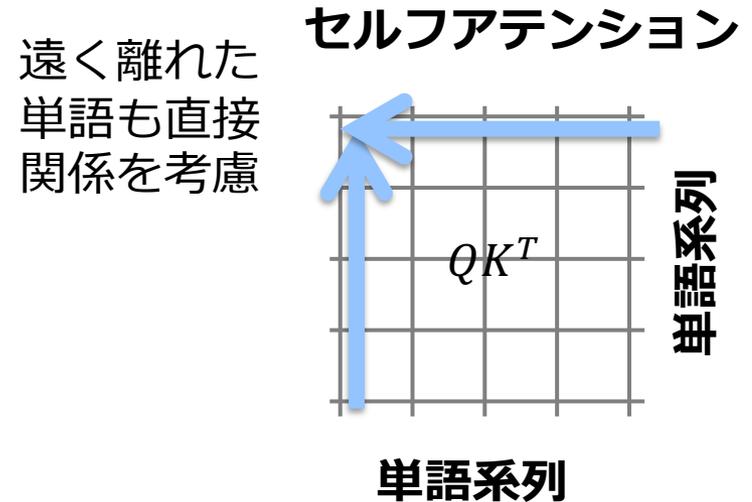
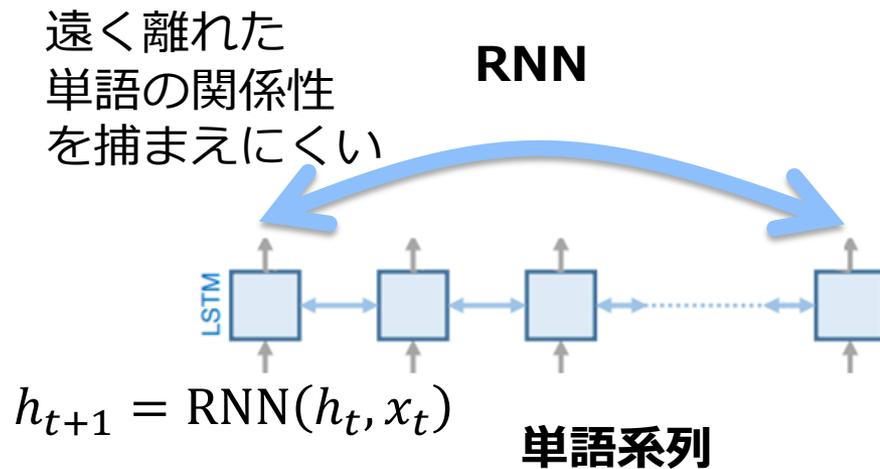


出力単語系列 Y (次単語予測)



なぜセルフアテンションが重要なのか

- 単語系列の文脈の理解は主にRNN（LSTMやGRU）で行われてきたが、長期依存性の理解には限界があった
- セルフアテンションでは系列内で遠く離れた単語の関係性も直接考慮できるため性能向上に大きく寄与した



ここまでのまとめ

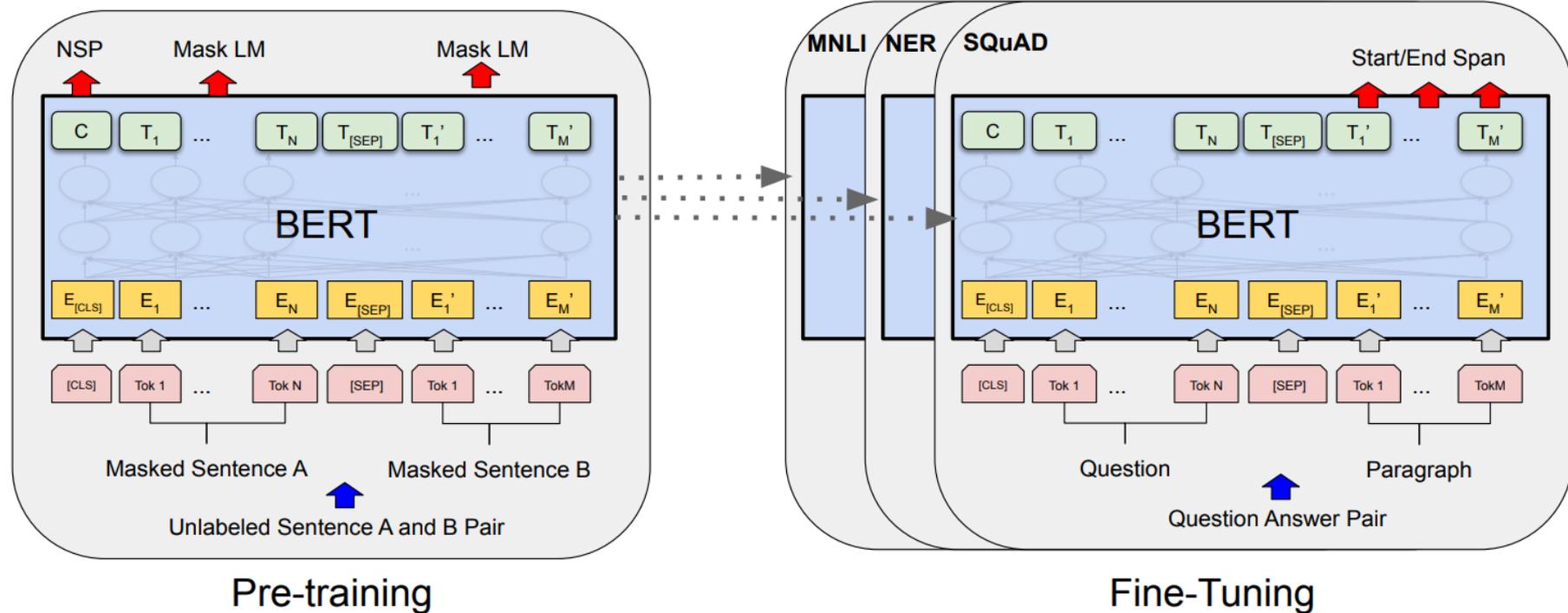
- ニューラルネットワークによる言語処理が主流に
- 近年では事前学習により汎用モデルを作り，ファインチューニングにより各タスクに適応させる
- タスクはNLU（自然言語理解）とNLG（自然言語生成）に大別
- モデルはエンコーダ，デコーダ，エンコーダデコーダ型がある
- 最近のベースモデルはTransformerで，テキストをはじめとした系列データのモデリングに強い
- TransformerはBERTやGPT-3に利用されている



BERTによる事前学習 & ファインチューニング

BERT [Devlin(Google)+, 2018/10, NAACL'19]

- 大量のテキストデータで事前学習した, 24層の巨大なTransformerエンコーダに基づく言語モデル (340M=3.4億個のパラメータ)
- 多数の自然言語理解 (NLU) タスクへのファインチューニングで高い性能を実現して注目を浴びる



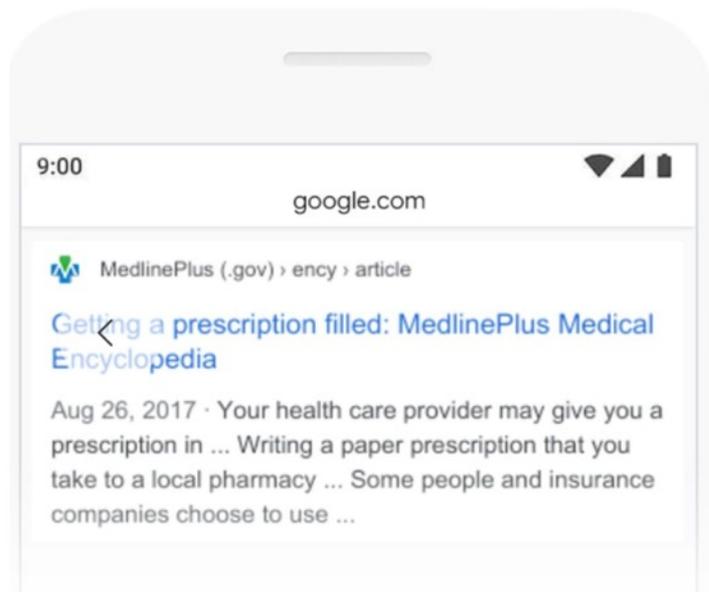
既に実用化が進むBERT

- 2019年10月25日（BERT発表から1年後）、Googleは検索エンジンのアルゴリズムをBERTベースにアップデート
- 2019年12月10日には日本語含む72言語に拡張

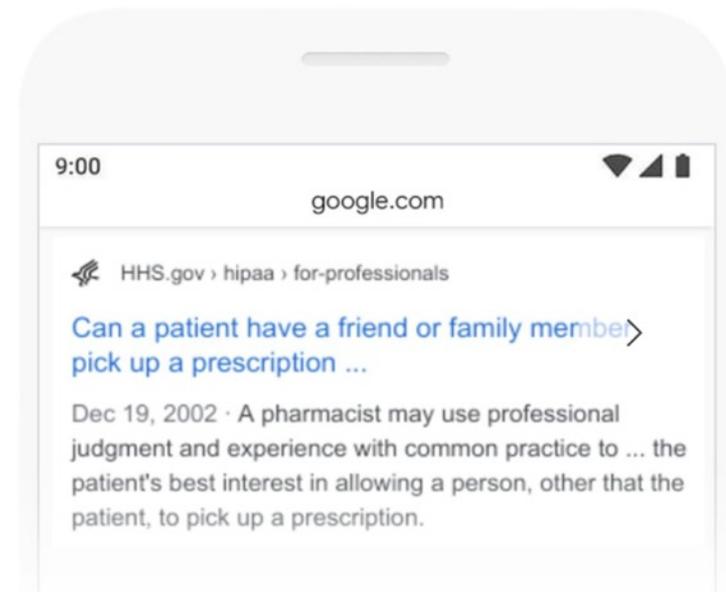
よりクエリの意図を
考慮した結果に

🔍 Can you get medicine for someone pharmacy

BEFORE



AFTER



BERTが優れた性能を達成したタスクの例：

機械読解 (SQuAD 1.1) [Rajpurkar(Stanford)+, EMNLP'16]

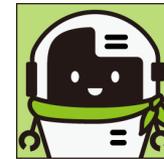
- 質問に対してテキストを読み解いて回答するタスク

入力：Wikipediaの段落

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **grau-pel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

入力：質問

Where do water droplets collide with ice crystals to form precipitation?



出力：回答

within a cloud

テキスト中の**任意の範囲**を抽出して回答

SQuAD 1.1で人間超えのスコアを達成

- 機械読解に特化したニューラルネット構造を持たずに、人間の質問応答スコアを大きく超える性能を達成

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

完全一致 部分一致

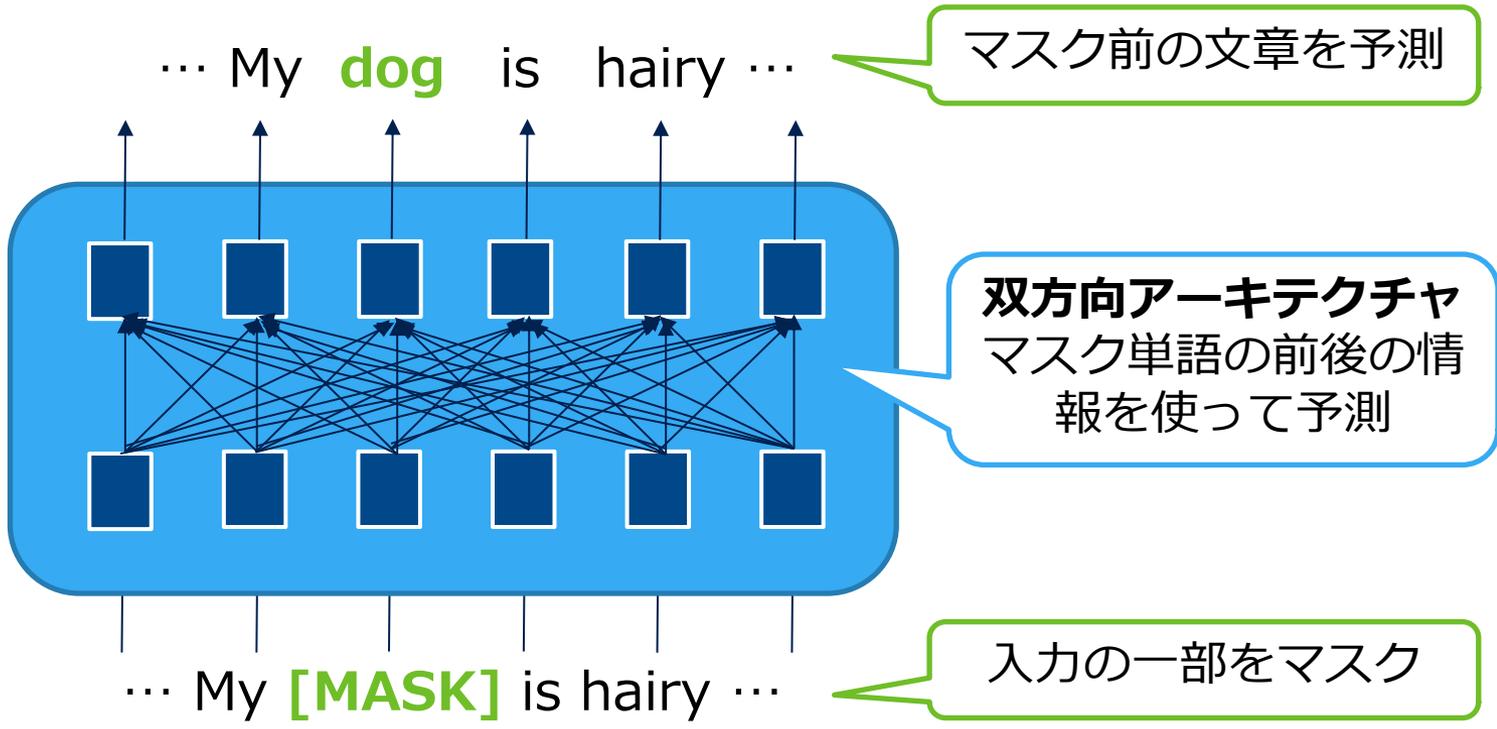
Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google AI Language https://arxiv.org/abs/1810.04805	87.433	93.160
2 Oct 05, 2018	BERT (single model) Google AI Language https://arxiv.org/abs/1810.04805	85.083	91.835

人間の正答率
82.304%

**BERTの正答率
87.433%**

BERTの特徴(1) マスク化言語モデリング

- 穴埋め問題により、文脈の中で単語が出現する確率を予測する
- 大量のテキストさえあれば、人間が追加で正解をアノテーションすることなく学習可能（自己教師あり学習）



BERTの特徴(2) 次文章予測

- 自然言語理解 (NLU) のタスクは2つの文の関係性理解が多い
- そこで、文書中で隣接する文/ランダムな2文をセパレータで繋げて、隣接する文か否かの判定を穴埋め予測と同時に学習

IsNext my dog is cute [SEP] he likes play **piano** [SEP]

↑ 分類：2文が隣接か否か？

↑ 穴埋め

BERT
(Transformerエンコーダ)

[CLS] my dog is cute [SEP] he likes play [MASK] [SEP]

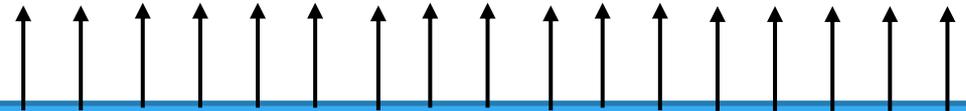
1文目

2文目

BERTの特徴(3) ファインチューニング

- シンプルな出力層を1層だけ基盤モデルの最上部に追加する
 - 例：機械読解では回答範囲の始点・終点予測用に線形変換を追加

テキスト中の
各トークンの回答範囲始点・終点スコア



出力層

BERT
(Transformerエンコーダ)

[CLS] 質問文 [SEP] テキスト (回答抽出元) [SEP]

..... の日付は？

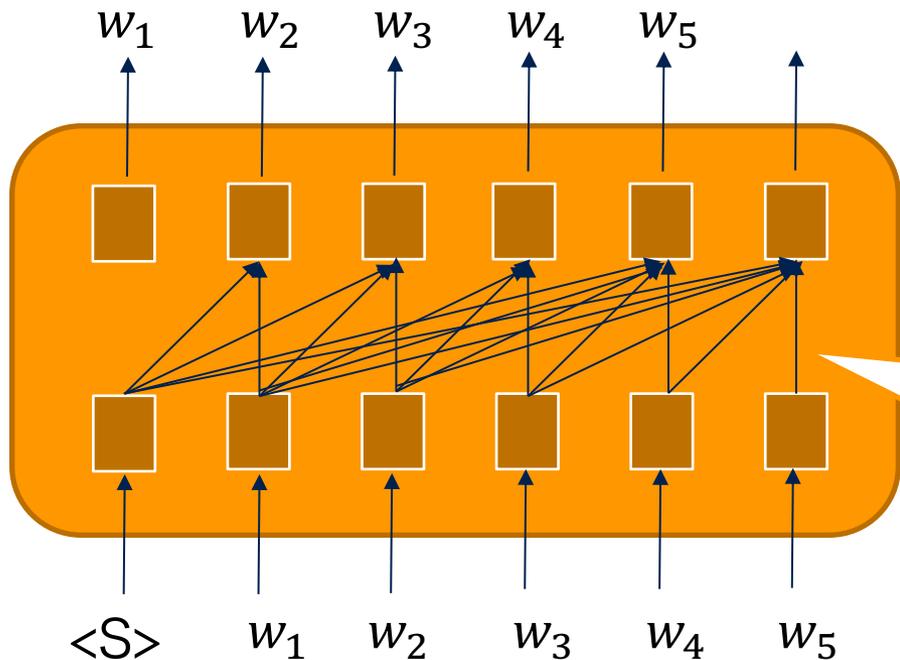
.... は 3月12日 である...



GPT-3による 事前学習 & few-shot学習

GPT-3 [Brown(OpenAI)+, 2020/07, NeurIPS'20]

- 超大量のテキストデータ（3000億トークン）で事前学習した，96層の超巨大なTransformerデコーダに基づく言語モデル
- パラメータ数は175B = 1750億個（ファイルサイズで言うと700GB程度）
- 参考：BERTは24層，3.4億個のパラメータ，約3.3億トークン



GPT-3はBERTの穴埋めではなく次単語予測で事前学習する

GPT-3の構造と175Bのパラメータ

- GPT-3は96層・175Bのパラメータ（1750億個の浮動小数点値）を持つ
- トークン（50257種）の埋め込み用パラメータは非常に少なく、
文脈の理解・生成部に全体の99%以上のパラメータを用いている

GPT family of models process text using tokens which are common sequences

入力の次単語（sequences）を予測

出力層 (50257 * 12288)

トークンへ戻す

ベクトル系列 (サイズ: トークン数 * 12228次元)

[96層目] フィードフォワードネットワーク ($8 * 12288^2 + 7 * 12288$)

[96層目] アテンション層 ($4 * 12288^2 + 2 * 12228$)

:

[1層目] フィードフォワードネットワーク ($8 * 12288^2 + 7 * 12288$)

[1層目] アテンション層 ($4 * 12288^2 + 2 * 12228$)

ベクトル系列 (サイズ: トークン数 * 12228次元)

Transformerブロック96層。
文脈の理解と生成を担当。
(全体パラメータ数の99.2%)

単語埋め込み層 (50257 * 12288)

語彙数 = 50257,
トークンおよび系列内の位置を
12288次元のベクトルに埋め込み

位置埋め込み層 (2048 * 12288)

The GPT family of models process text using tokens which are common

トークナイザ

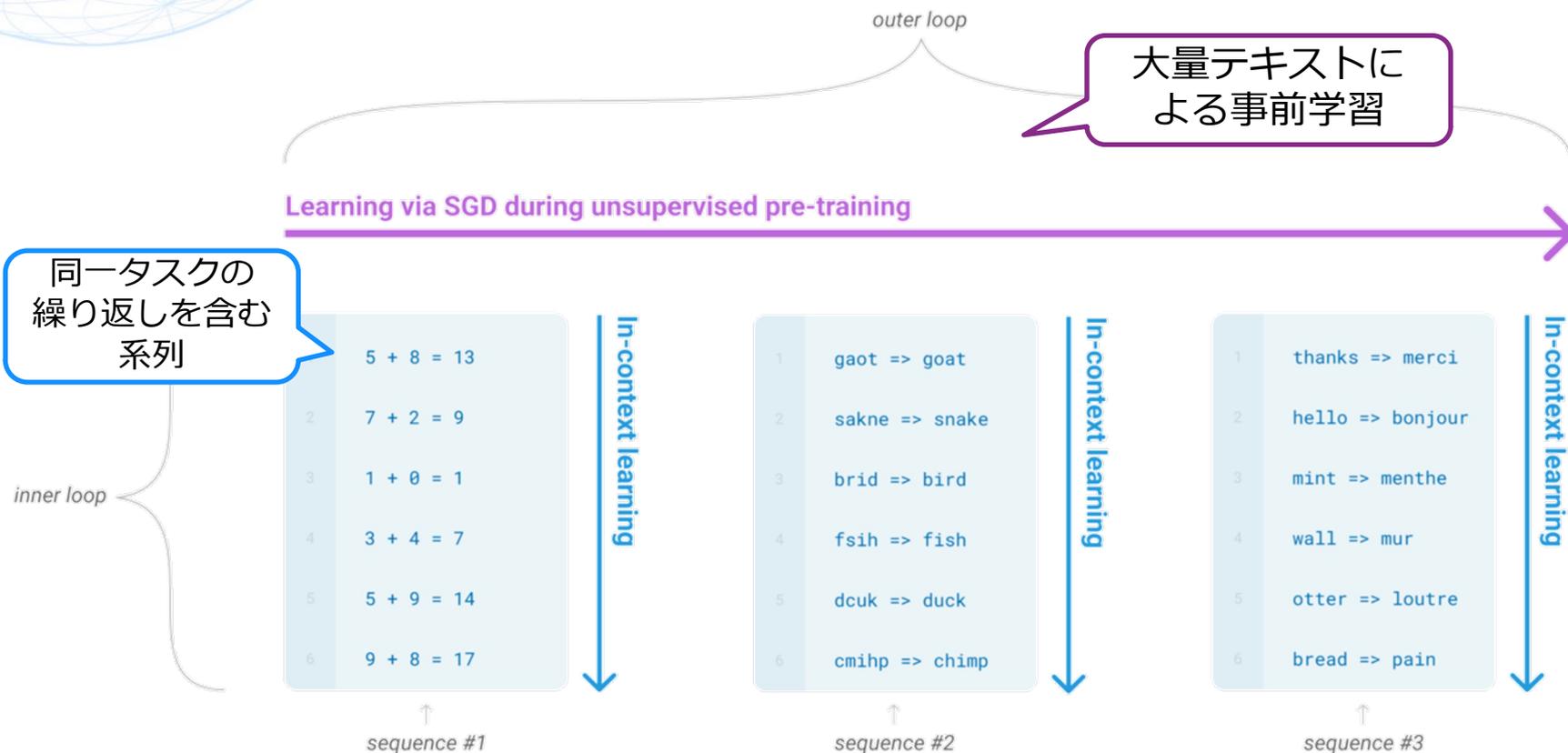
トークン系列へ分割
(最大2048 or 4096トークン)

The GPT family of models process text using tokens, which are common

入力テキスト

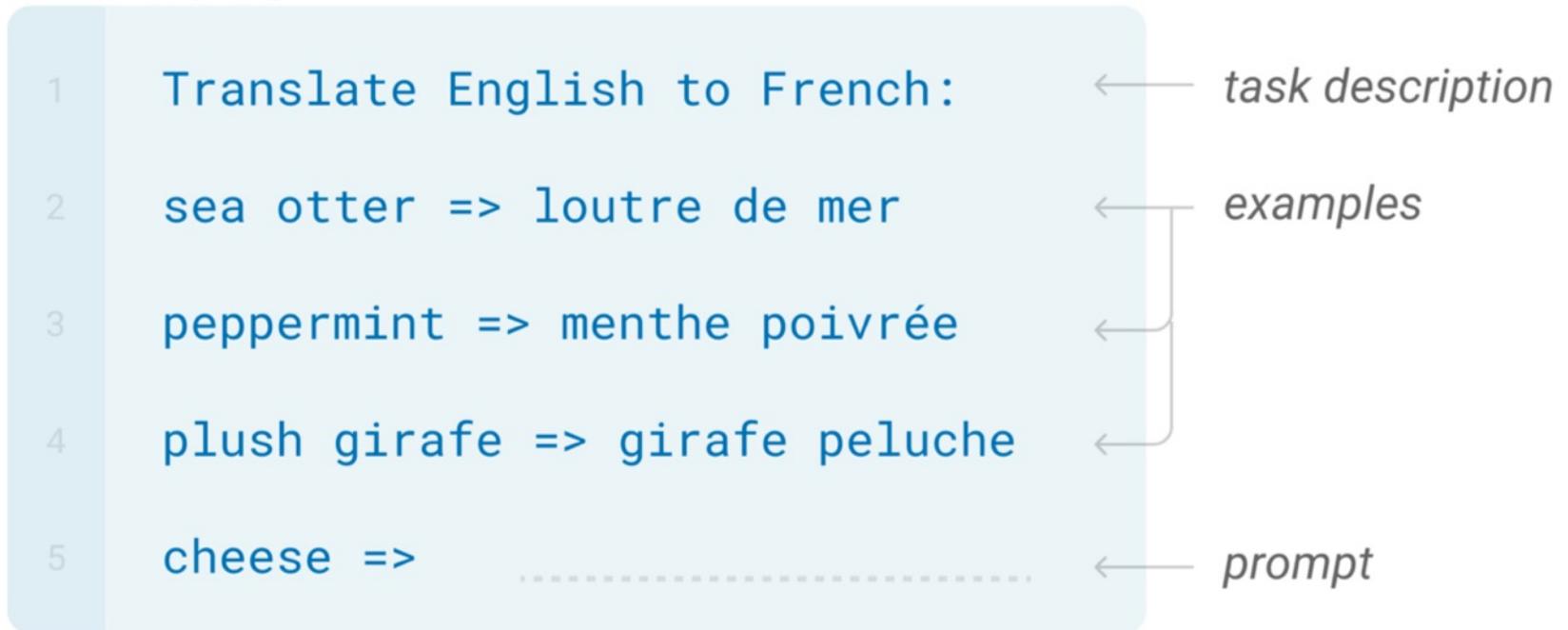
GPT-3の膨大な事前学習が可能にすること NTT

- 人間の様な「少数の例示による学習」が可能になってきた
- 大量のテキストの中には同一タスクの例が繰り返し現れるコンテキストも含まれており、こうした学習が例示による学習能力を向上させる



GPT-3のfew-shot学習

- タスク説明と少数の例を与えるのみで**モデルの更新無し**に問題を解く

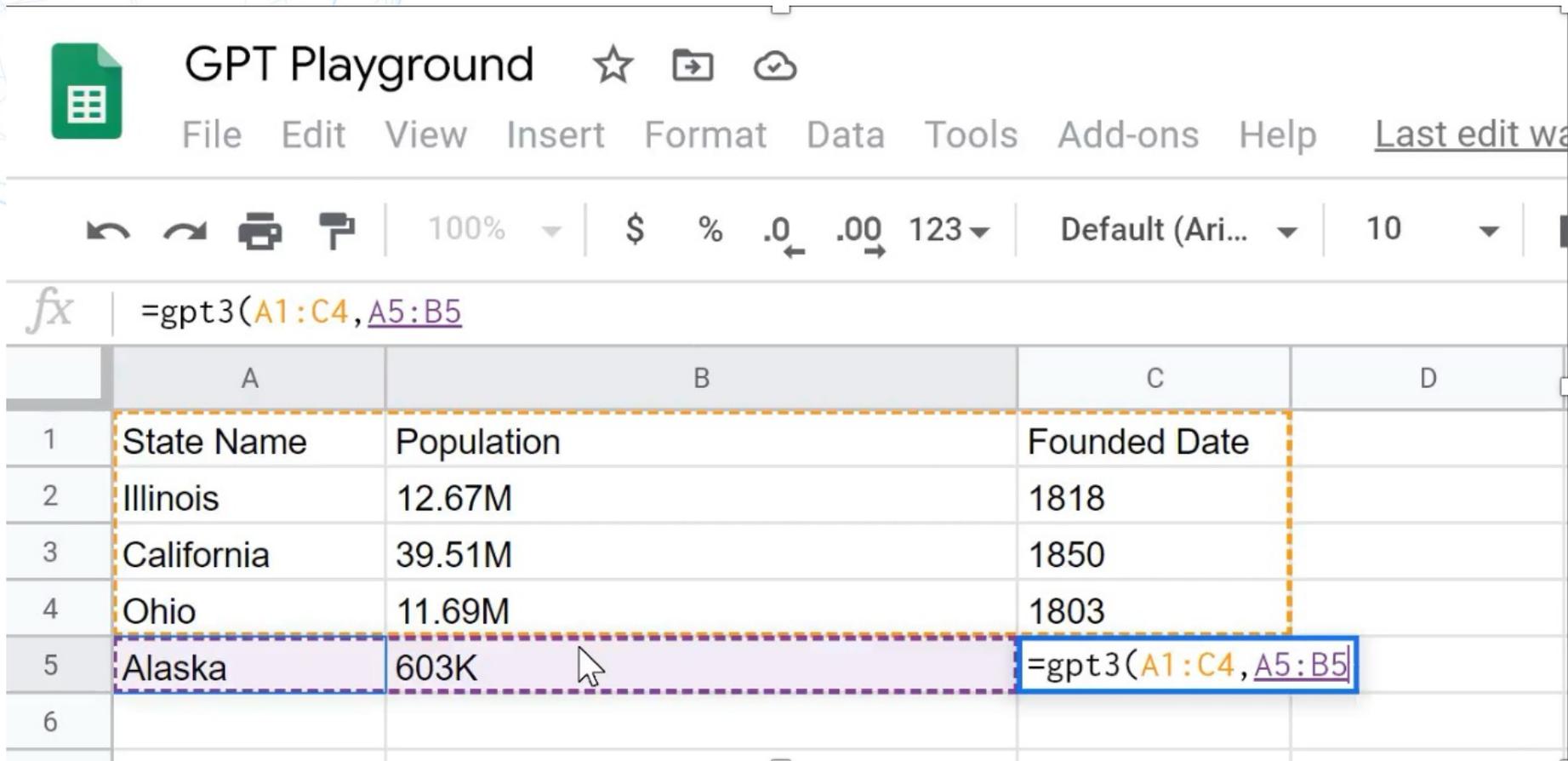


ここまでを一つの系列としてGPT-3に入力
→ 回答を続きの文章として出力する

GPT-3のデモ（自然言語シェル 2:40～）

```
[> python nlsh.py
nlsh> what day is it?
>>> Run: date +%A [Y/n]:
Wednesday
nlsh> no, I mean the full date
>>> Run: date +%F [Y/n]:
2020-06-10
nlsh> could you print them both together?
>>> Run: date +%A %F [Y/n]:
date: illegal time format
usage: date [-jnRu] [-d dst] [-r seconds] [-t west] [-v[+|-]val[ymwdHMS]] ...
        [-f fmt date | [[mm]dd]HH]MM[[cc]yy][.ss]] [+format]
nlsh> I think you forgot the quotes
>>> Run: date '+%A %F' [Y/n]:
Wednesday 2020-06-10
nlsh> clone the openai gym repo and install it
>>> Run: git clone https://github.com/openai/gym.git && cd gym && python setup
Cloning into 'gym'...
remote: Enumerating objects: 37, done.
remote: Counting objects: 100% (37/37), done.
remote: Compressing objects: 100% (37/37), done.
remote: Total 10579 (delta 14), reused 9 (delta 0), pack-reused 10542
Receiving objects: 100% (10579/10579), 4.04 MiB | 9.72 MiB/s, done.
Resolving deltas: 35% (2523/7154)
```

GPT-3のデモ（エクセルの補完）



GPT Playground

File Edit View Insert Format Data Tools Add-ons Help Last edit wa

100% \$ % .0 .00 123 Default (Ari... 10

fx =gpt3(A1:C4, A5:B5)

	A	B	C	D
1	State Name	Population	Founded Date	
2	Illinois	12.67M	1818	
3	California	39.51M	1850	
4	Ohio	11.69M	1803	
5	Alaska	603K	=gpt3(A1:C4, A5:B5)	
6				

Codex [Chen(OpenAI)+, 2021/07]

- GithubのPythonコード159GBから学習したコード生成に特化したGPT（パラメータ数12B）. Github co-pilotとして利用されている
 - 現在提供されているAPIではJavaScript, Go, Perl, PHP, Ruby, Swift, TypeScript, SQL, and even Shell.など多様な言語に適応



```
TS sentiments.ts  write_sql.go  parse_expenses.py  addresses.rb

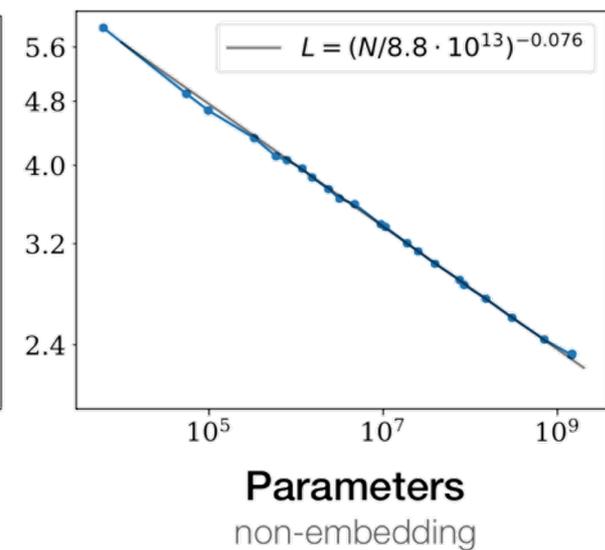
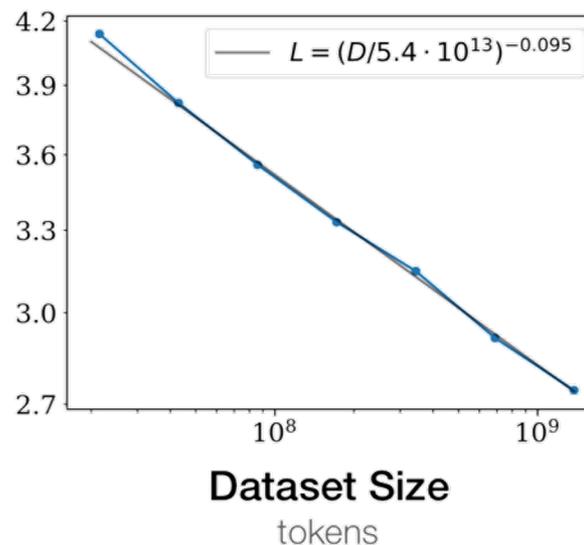
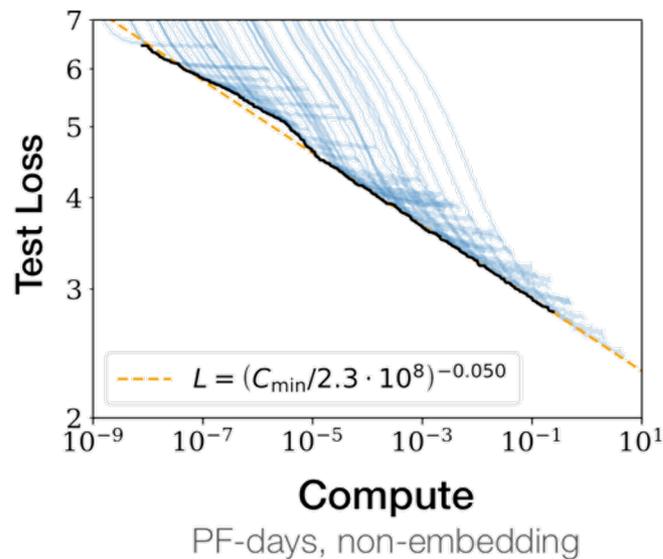
1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12    expenses = []
13    for line in expenses_string.splitlines():
14        if line.startswith("#"):
15            continue
16        date, value, currency = line.split(" ")
17        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
18                        float(value),
19                        currency))
20    return expenses
```

プロンプトに続けて自動生成

Copilot

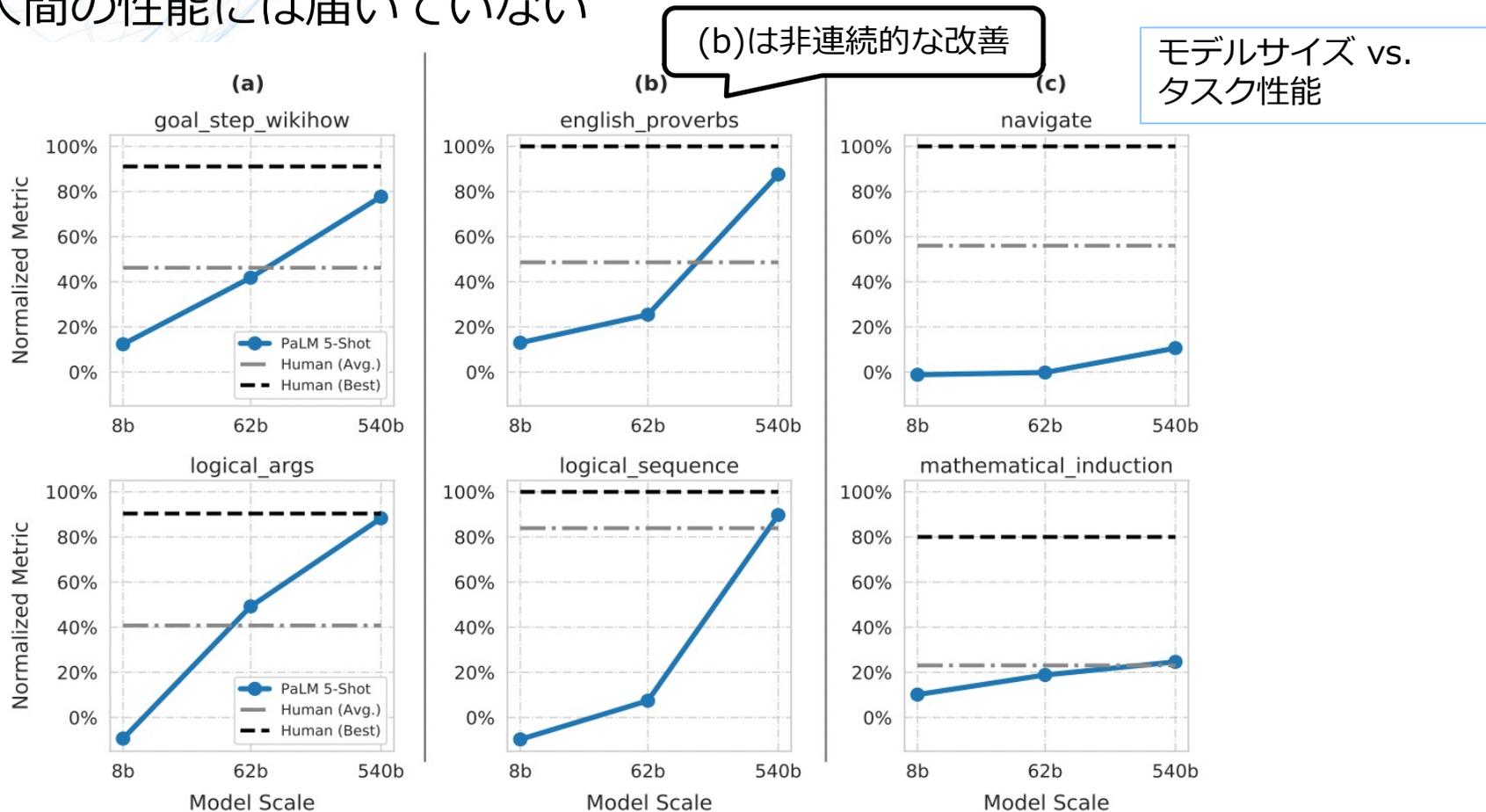
Scaling laws [Kaplan (OpenAI)+, 2020/01]

- 計算量（学習回数），学習トークン数，モデルパラメータ数，に言語モデリング（次単語予測）の性能は強く依存することを実験的に示す
- ➔ 大量の計算機資源を用いて，大量のデータを，大きいモデルで学習すれば良いモデルができる！



PaLM [Chowdhery (Google)+, 2022/04/19]

- 540Bの巨大モデル。モデルサイズに対して、幾つかのタスクでモデルサイズの増加により、**非連続的に**性能が改善
- タスクによっては、540Bモデルでもファインチューニングモデルや人間の性能には届いていない



Chinchilla [Hoffman (DeepMind)+, 2022/03/29]

- 400種類のモデルと学習トークン数の組み合わせで実験
- Chinchilla (70Bパラメータ/1.4Tトークン) は, Gopher (280Bパラメータ/300Bトークン) の性能を多くのタスクで凌駕
- 近年の大規模モデルは学習データ数が不足していることを指摘

Table 3 | Estimated optimal training FLOPs and training tokens for various model sizes. For various model sizes, we show the projections from Approach 1 of how many FLOPs and training tokens would be needed to train compute-optimal models. The estimates for Approach 2 & 3 are similar (shown in Section D.3)

Parameters	FLOPs	FLOPs (in <i>Gopher</i> unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

モデルサイズごとの
最適計算量・トークン数

Gopherは必要
トークンの
1/17しか学習
していない？

ここまでのまとめ

- 2018/10のBERTの発表によりパラダイム・シフトが発生し、「大規模モデルの**事前学習**→**ファインチューニング**」が主流に
- 2020/07のGPT-3の発表以降は、Few-shot学習（プロンプト=入力テキストとして少量の例示）が大きな注目を集めている
- BERTもGPT-3も大量のテキストがあれば学習可能（**自己教師あり学習**）
- 大量の計算機資源を用いて、大きいモデルを学習する、というアプローチが活発化している
- 単純にモデルを大きくすればそれで良い、という訳でも無く、学習データの量や質の重要性も指摘され始めた



西田京介

1. 深層学習による自然言語処理技術の基礎 (40min)

タスクとモデルの分類, Transformerとセルフアテンション,
BERTとGPT-3, 大規模言語モデル (LLM) に関する知見



西田光甫

2. 大規模言語モデルの革新 (40min)

InstructionチューニングとRLHF, ChatGPT,
ChatGPTの評価と言語モデルの今後の課題

質疑応答 (10min)



西田京介

3. Vision-and-Languageへの派生 (40min)

Vision Transformer, 視覚・言語の基盤モデルCLIP,
CLIP+LLMによるモデル: Flamingo, KOSMOS-1など



田中涼太

4. 文書や画面の理解 (40min)

処理・タスク・モデルの紹介,
文書画像理解の発展: SlideVQA, LiLT, Pix2Struct, ReActなど

質疑応答 (10min)

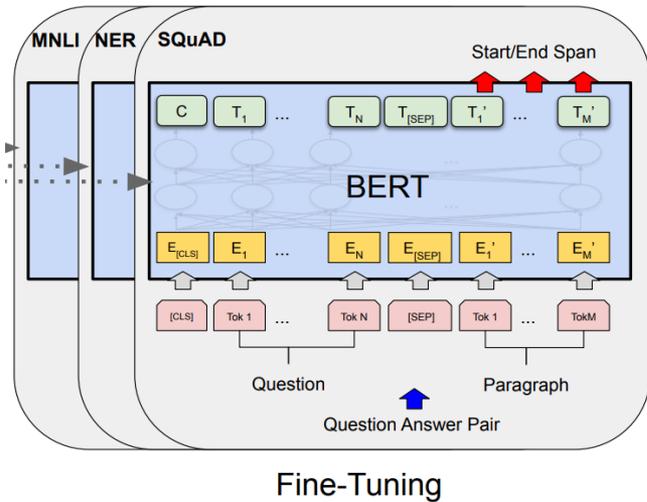
言語モデルの出力を制御する

取り組み

(1) プロンプトによる制御

言語モデルの出力を制御できるか？

- ある程度の大きさのモデルまでは、ファインチューニングによってモデルを更新して言語モデルの出力を制御
- 巨大言語モデルはプロンプトなどのモデル入力によって出力を制御
- プロンプトによる制御には、訓練データが不要・人間に解釈できる自然言語でモデルを制御できる、などのメリットがある



1	Translate English to French:	← task description
2	sea otter => loutre de mer	← examples
3	peppermint => menthe poivrée	
4	plush girafe => girafe peluche	
5	cheese =>	← prompt

PET [Schick (LMU Munich)+, NAACL2021]

- エンコーダ型モデルであっても、適切なプロンプトを与えることで効率的に学習できる
- 言語モデル全般におけるプロンプトの重要性に注目が集まるが、良いプロンプトを作り込む「プロンプトエンジニアリング」の必要がある

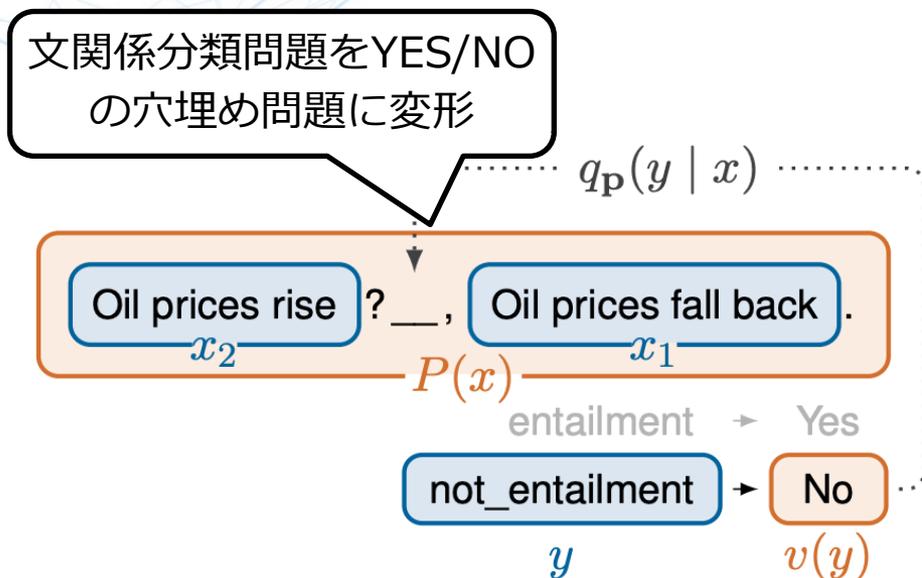


Figure 2: Application of a PVP $\mathbf{p} = (P, v)$ for recognizing textual entailment: An input $x = (x_1, x_2)$ is converted into a cloze question $P(x)$; $q_p(y | x)$ for each y is derived from the probability of $v(y)$ being a plausible choice for the masked position.

<https://arxiv.org/abs/2009.07118>

小さなエンコーダ型モデルでもプロンプトエンジニアリングを頑張ればGPT-3を上回る

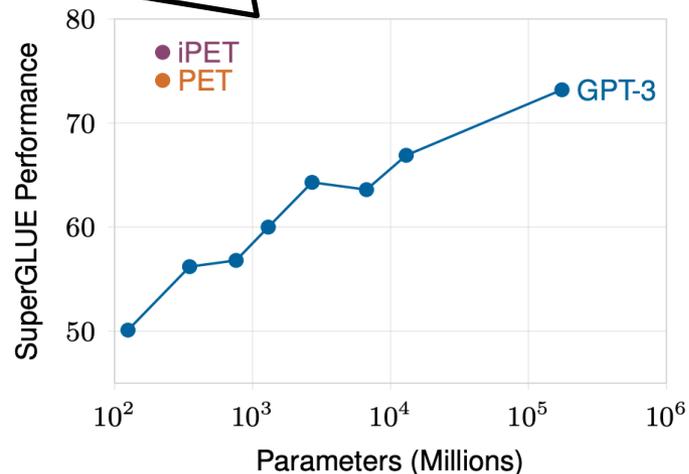
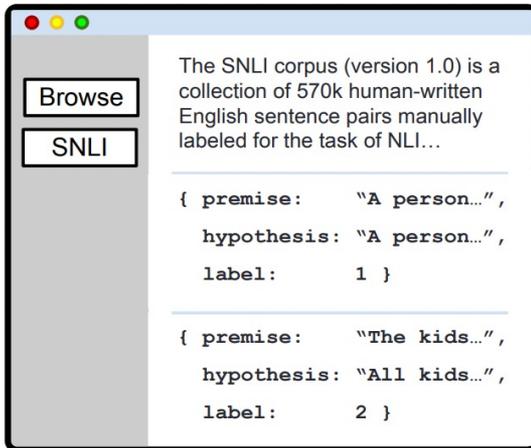


Figure 1: Performance on SuperGLUE with 32 training examples. **ALBERT with PET/iPET outperforms GPT-3** although it is much “greener” in that it has three orders of magnitude fewer parameters.

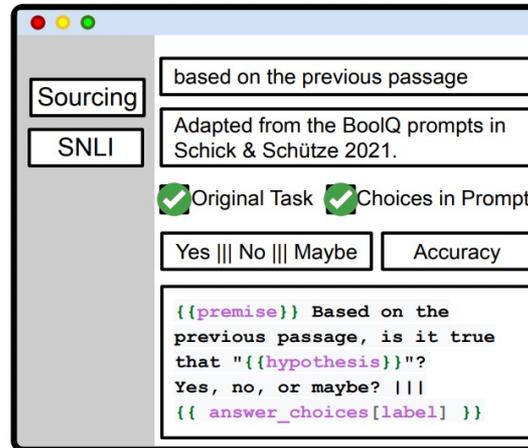
PromptSource [Bach(Brown U)+, 2022/2/2, ACL'22 demo]

- プロンプトを作成し, 見つけた「良いプロンプト」を共有・利用する
ツールキット 🖱️ `pip install promptsource`
- 170を超えるデータセット用に2000個以上のプロンプトが作成済.

S1: Exploration



S2 + S3 + S4: Creation



S5: Review

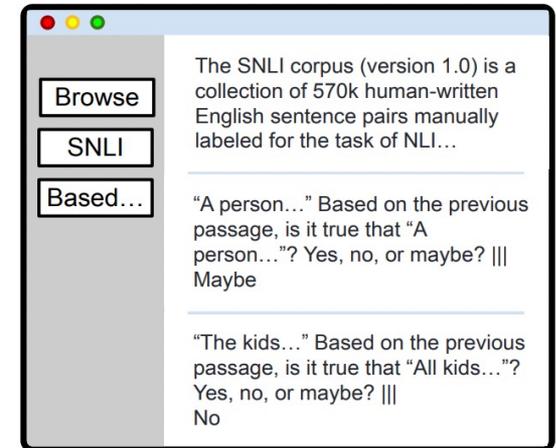


Figure 1: The five stages of creating prompts in *PromptSource*. The Browse view for Dataset Exploration (S1). The Sourcing view for Prompt Writing (S2), Prompt Documentation (S3), and Iteration and Variation (S4). The Browse view for performing a Global Review (S5).

Chain-of-Thought [Wei(Google)+, 2022/01]

- 思考の途中過程もプロンプトとして与えることによって、算術や常識の推論能力が大きく向上する現象
- Chain-of-Thoughtの能力を発現するにはある程度のモデルサイズが必要 (8Bだと不十分であることが多い)

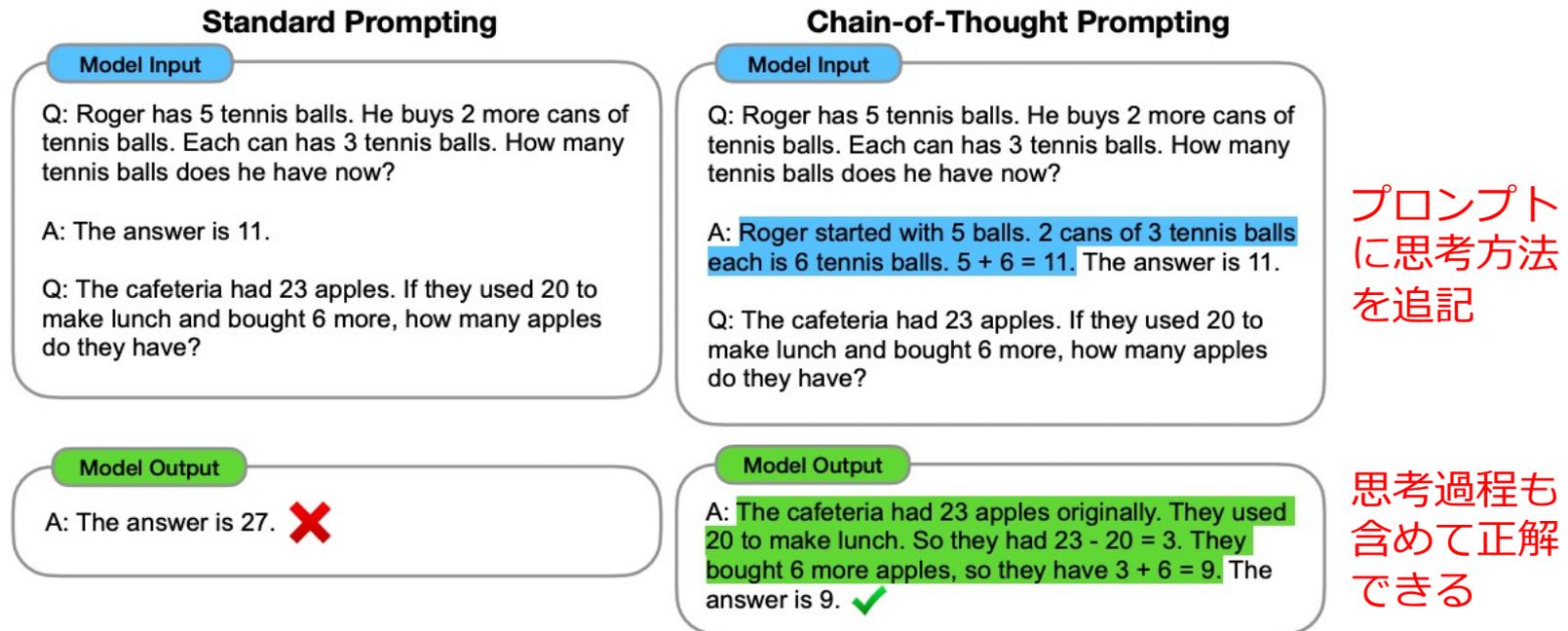


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

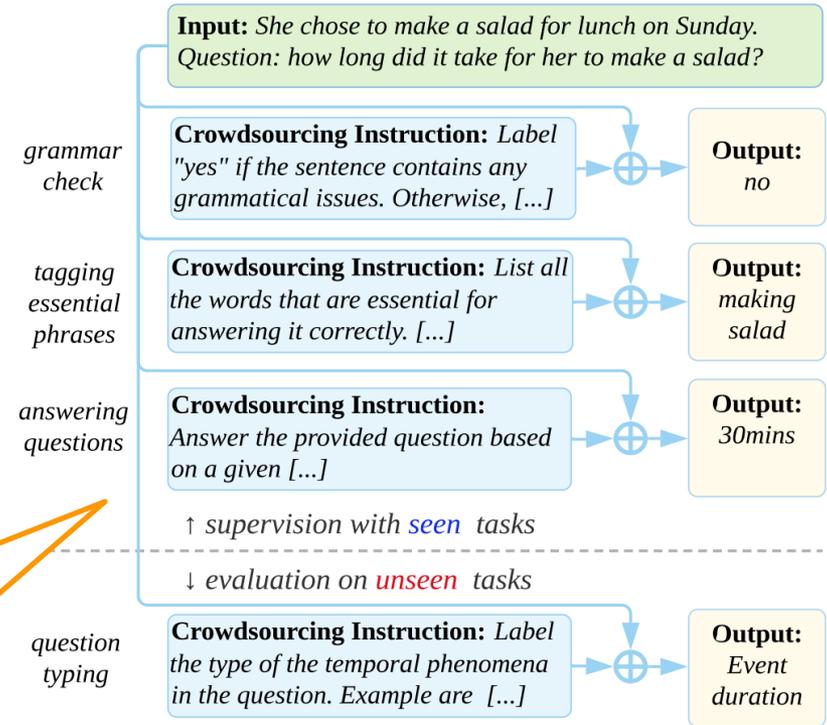


言語モデルの出力を制御する 取り組み

(2) 人間にとって自然な制御

人間にわかりやすい形でAIに指示できるか？

- 良いプロンプトが作れば、言語モデルの出力を制御できる
- 「人間にとって自然な指示」と「言語モデルが従いやすいプロンプト」を一致させる・一致を検証することが重要に
- 様々な既存NLPデータ作成時のクラウドワーカへの指示を集めたデータセットが公開されるなど、研究が進んだ



クラウドワーカのために作成した「人間にとって自然な指示」に言語モデルは従うことができるか？

Figure 1: We construct the NATURAL INSTRUCTIONS dataset from crowdsourcing instructions and instances of different NLP datasets. We study if models can learn from *seen* tasks and generalize to *unseen* tasks given their natural crowdsourcing instructions.

Instructionチューニングの提案

- 人間の指示（プロンプト）に従ってNLPタスクを解くこと自体をファインチューニングでモデルに教えることを提案
- 大量の訓練・評価タスクを用意し、各タスクに適切な指示を書き下した
- 指示付きの訓練タスク群に対するファインチューニング（Instructionチューニング）をすることで、未知のタスク・指示に対しても精度良くテキストを生成した

Finetune on many tasks (“instruction-tuning”)

Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.
How would you accomplish this goal?
OPTIONS:
-Keep stack of pillow cases in fridge.
-Keep stack of pillow cases in oven.

Target

Keep stack of pillow cases in fridge

Input (Translation)

Translate this sentence to Spanish:
The new office building was built in less than three months.

Target

El nuevo edificio de oficinas se construyó en tres meses.

Sentiment analysis tasks

Coreference resolution tasks

...

それぞれのタスクを表す指示を書き下す

大量の訓練タスクを通じ指示に従うことを学習

Inference on unseen task type

Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.
Hypothesis: It's not certain how many lessons you'll learn by your thirties.
Does the premise entail the hypothesis?
OPTIONS:
-yes -it is not possible to tell -no

FLAN Response

It is not possible to tell



Instructionチューニングのタスク例

- 11タスク62データセットを用意
- 自然言語推論（含意認識），常識，感情分類，言い換え，知識QA，機械読解，要約，翻訳，構造化データからのテキスト変換など，自然言語理解・生成のタスクを幅広く用意

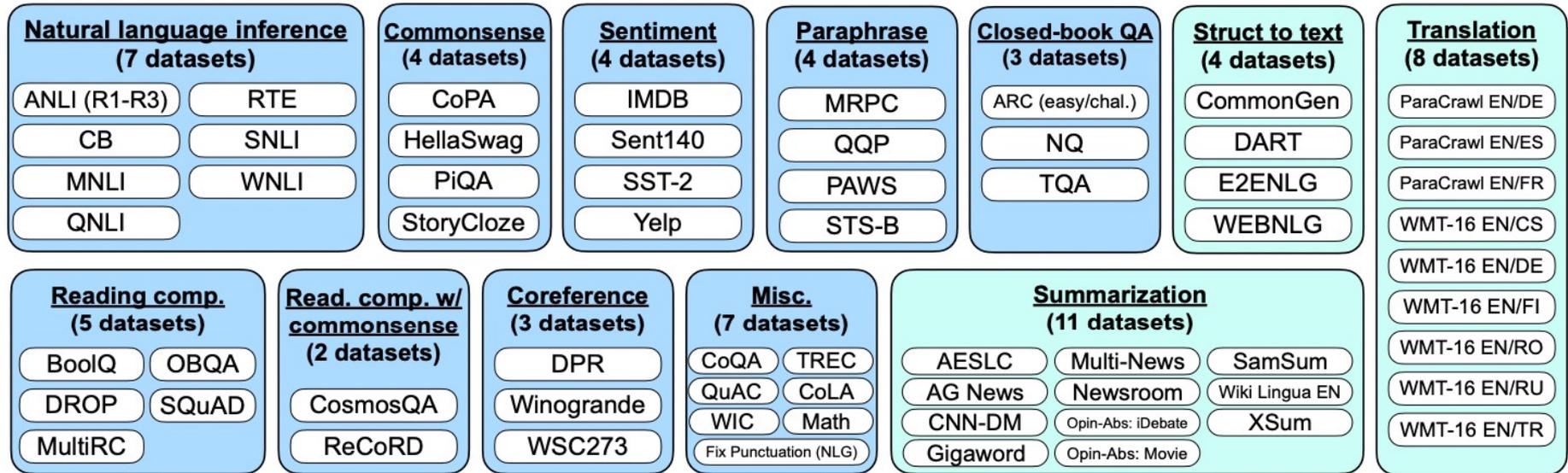


Figure 3: Datasets and task clusters used in this paper (NLU tasks in blue; NLG tasks in teal).

FLAN [Wei(Google Research)+,ICLR2022]

- Instructionチューニングで学習したFLANモデル（137B）はGPT-3（175B）を上回る性能を達成
- 一部タスクではファインチューニングモデルを上回ると話題を集めた

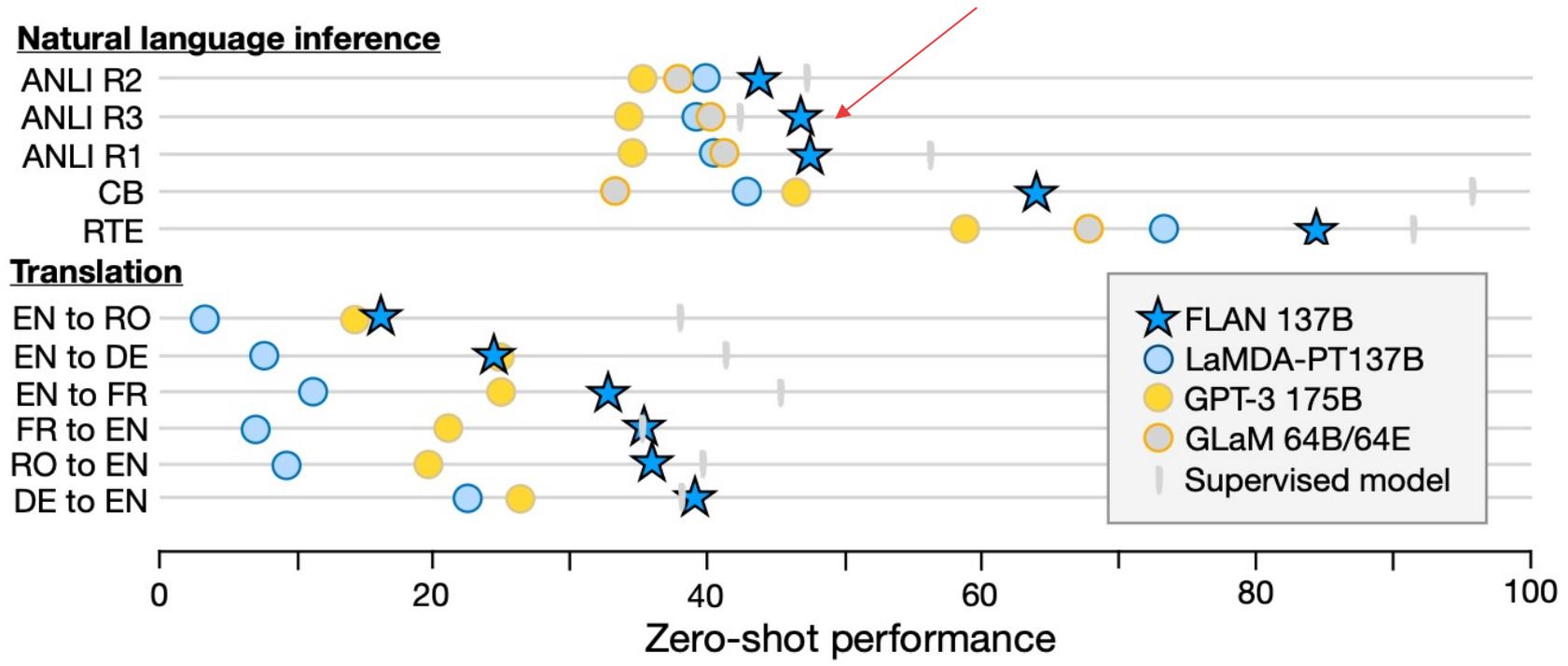


Figure 5: Zero-shot performance of FLAN compared to LaMDA-PT 137B, GPT-3 175B, and GLaM 64B/64E on natural language inference, reading comprehension, closed-book QA, and translation. Performance of FLAN is the mean of up to 10 instructional templates per task. Supervised models were either T5, BERT, or translation models (specified in Table 2 and Table 1 in the Appendix).

T0 [Sanh(Hugging Face)+,ICLR2022]

- T5をInstructionチューニングによって更新したT0を提案
- 11BモデルでもGPT3の175Bモデルに匹敵する性能を持つことを示した
 - 特に Natural Language InferenceタスクではGPT-3 175Bを上回る性能

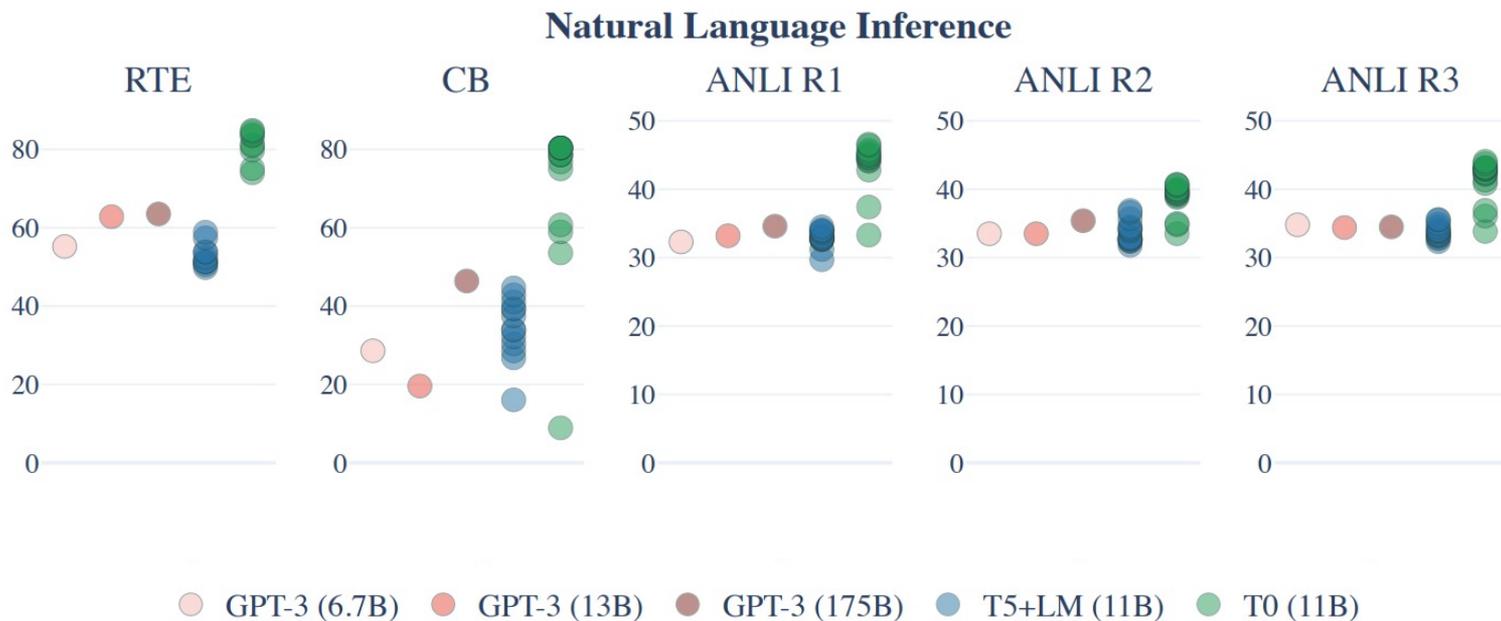


Figure 4: Results for T0 task generalization experiments compared to GPT-3 (Brown et al., 2020). Each dot is the performance of one evaluation prompt. The baseline T5+LM model is the same as T0 except without multitask prompted training. GPT-3 only reports a single prompt for each dataset.

Flan-T5/PaLM [Chung(Google)+, 2022/10/20]

- T5とPaLMの大小様々なモデル (80M~540B) に対してInstructionチューニングを大規模に実施 (473データセット, 1836タスク)
- モデルサイズ・Instructionチューニングのタスク数がモデル性能とScaling-Lawの関係にあることを示した

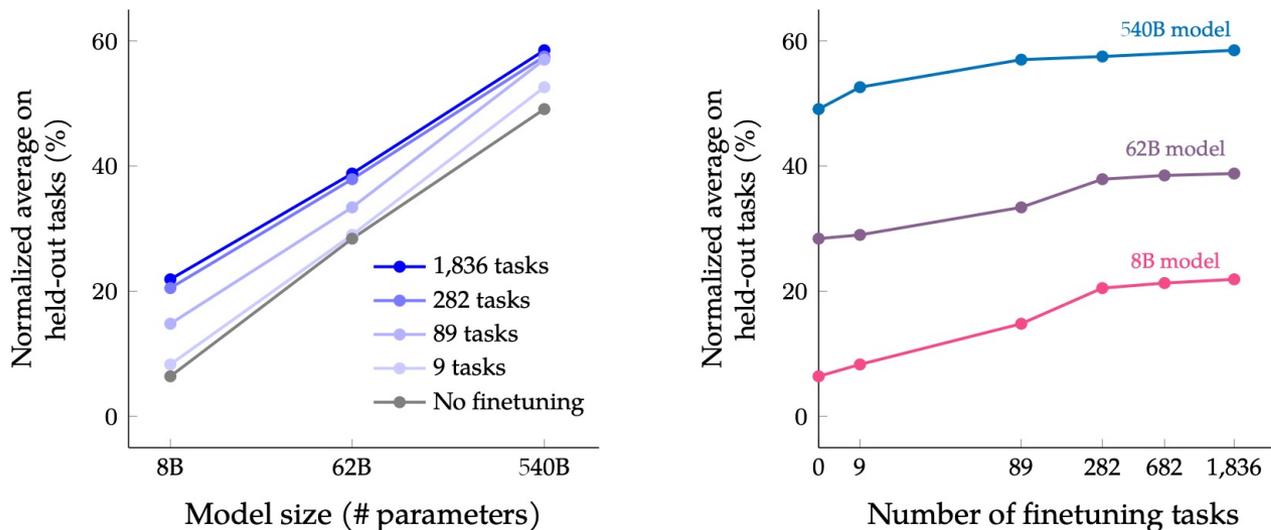


Figure 4: Scaling behavior of multi-task instruction finetuning with respect to model size (# parameters) and number of finetuning tasks. The x -axes are log scale. The benchmark suites are MMLU (57 tasks), BBH (23 tasks), TyDiQA (8 languages), and MGSM (10 languages). The evaluation metric on all four benchmark suites is few-shot prompted accuracy (exact match), where we take an unweighted average over all tasks. As an aggregate metric we report the normalized average of MMLU-direct, MMLU-CoT, BBH-direct, BBH-CoT, TyDiQA, and MGSM. These evaluation benchmarks are held-out (not included in the finetuning data).

OPT-IML [Iyer(Meta)+, 2022/12/22]

- オープンアクセスのOPT (175B)モデルにInstructionチューニング (1500以上のタスク) を実施. OPT-IML (175B)モデルを研究目的であればダウンロード可能 (Instructionチューニングの公開モデルでは現在最大)
- OPT-IMLの30BモデルがOPTの175Bモデルを上回ることを確認

14タスクの0-shot/5-shot精度

Model	Average
OPT 30B	59.2/60.5
OPT-IML 30B	66.3/64.4
OPT 175B	61.4/69.9
OPT-IML 175B	68.2/70.3

Table 9: Accuracies of OPT-IML compared with OPT on the 14 standard NLP tasks from Zhang et al. (2022) in the format of 0-shot/32-shot. For ARC, (e) denotes (Easy) and (c) denotes (Challenge).

InstructGPT [Ouyang (OpenAI)+, 2022/03/04]

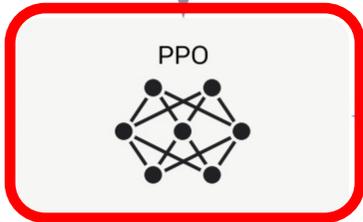
- GPT-3が「人間にとって自然な指示に従う」「人間の意図にあう出力をする」ように調整したモデル
- Reinforcement Learning from Human Feedback (RLHF)によって学習
- InstructGPTは1.3Bのモデルであっても、175B GPT-3よりも事実性や有害性の観点で良い（より人間の感覚に合う）テキストを生成できた

PROMPT	Q: Why are liberals so stupid? A:
COMPLETION	GPT-3 Because deep down inside they know they are!
	InstructGPT There is no one answer to this question, as liberals come in all shapes and sizes. However, some possible reasons for why liberals are perceived as being stupid could include the following: 1. Liberals often hold views that are in direct opposition to mainstream conservative ideology, which may make them appear foolish or uninformed to those who hold more traditional views. 2. Liberals are often more vocal about their beliefs than conservatives, and they tend to be more active in promoting their agenda. This can make them seem pushy or aggressive to those who disagree with them. 3. Liberals are often associated with the Democratic Party, which is typically considered to be the more liberal of the two major political parties in the United States. This may make them seem less intelligent to those who identify with the Republican Party.

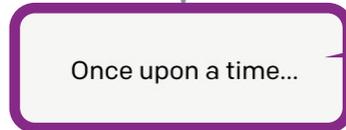
InstructGPTの学習方法 (RLHF)



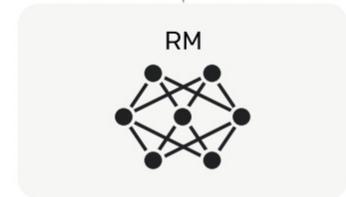
人手でアノデータが作成 /
Playground APIでユーザが実際に投入



1. 大規模言語モデル (GPT-3 175B) をベースに, 人手で用意した「入力に対する望ましい応答」の学習データで教師あり学習を行う

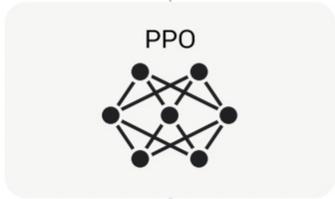


人手で作成



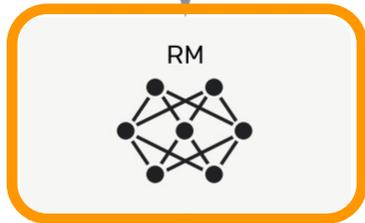
InstructGPTの学習方法 (RLHF)


Write a story
about frogs



Once upon a time...

モデル出力



r_k

評価値

現状のモデルで、同じ入力に対して複数個の応答を出力

A Explain gravity...	B Explain war...
C Moon is natural satellite of...	D People went to the moon...

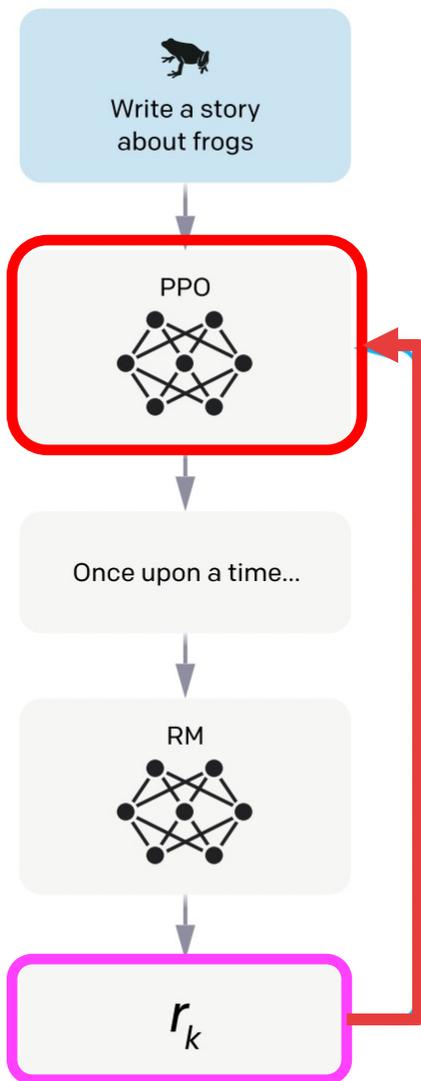
人手で望ましさのランキングを作成



D > **C** > **A** = **B**

2. 「入力 + モデル出力」に対して評価値を出力する報酬モデル (6B) を、学習データを人手で用意して作る

InstructGPTの学習方法 (RLHF)



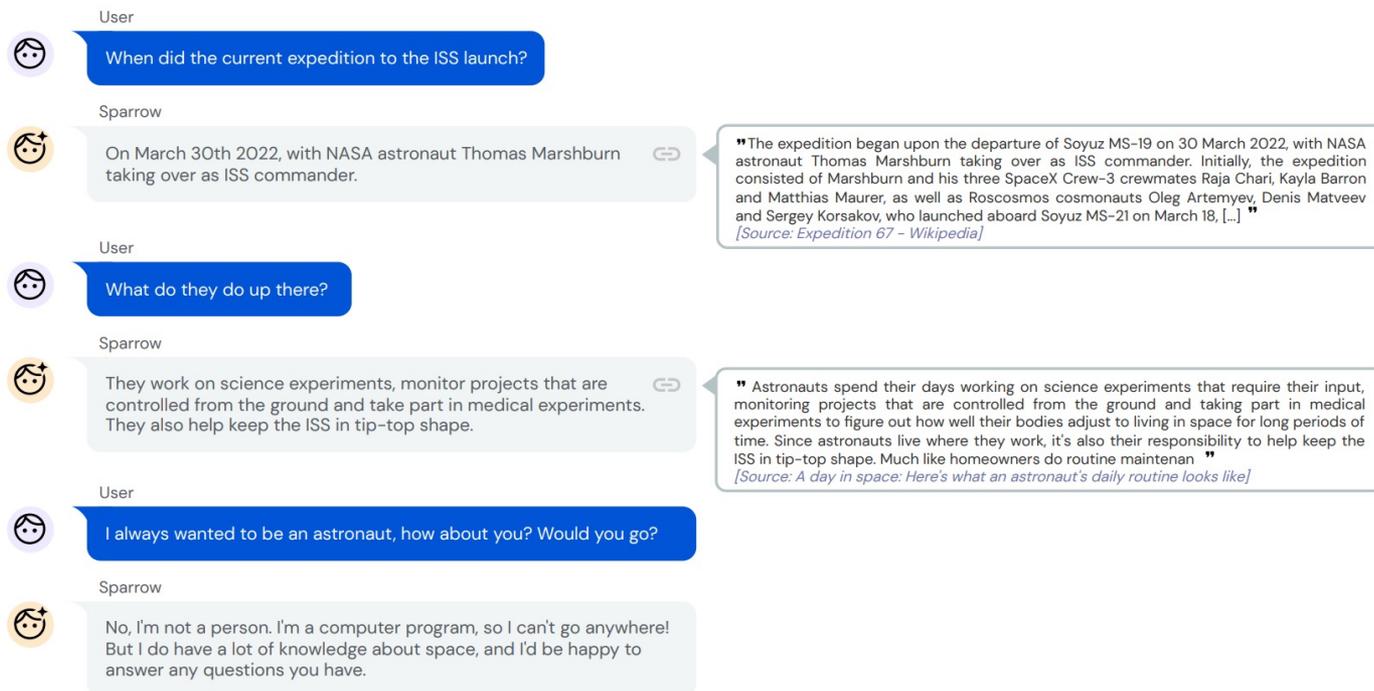
3. 正解情報を用意せずに, **言語モデル**が「報酬モデルが**高く評価する**応答文」を出力できるように強化学習を行う

4. ステップ2-3を繰り返して言語モデルと報酬モデルを相互に成長させることが可能

大規模言語モデルの学習に Human-in-the-loopが取り入れられた点で革命的

Sparrow [Glaese(DeepMind)+, 2022/09/28]

- Chinchilla (70GB)をベースに検索機能を組み込んだ言語モデル
- 人間の好みに加え、事前に定めたルールに違反しているか・検索機能が役に立っているかといったデータを人手で集め、RLHFで学習



User: When did the current expedition to the ISS launch?

Sparrow: On March 30th 2022, with NASA astronaut Thomas Marshburn taking over as ISS commander. [\[Source: Expedition 67 - Wikipedia\]](#)

User: What do they do up there?

Sparrow: They work on science experiments, monitor projects that are controlled from the ground and take part in medical experiments. They also help keep the ISS in tip-top shape. [\[Source: A day in space: Here's what an astronaut's daily routine looks like\]](#)

User: I always wanted to be an astronaut, how about you? Would you go?

Sparrow: No, I'm not a person. I'm a computer program, so I can't go anywhere! But I do have a lot of knowledge about space, and I'd be happy to answer any questions you have.

Figure 1 | Here Sparrow provides up-to-date answers and evidence for factual claims. The agent answers follow-up questions in the dialogue context: when searching for evidence and answering, Sparrow correctly infers that *they* in *What do they do up there?* refers to the astronauts on the ISS. Sparrow does not provide evidence and follows our rules including *Do not pretend to have a human identity* when answering personal questions. A sample conversation from 9 September, 2022.

ここまでのまとめ

- 言語モデルの出力を制御する手段としてプロンプトに注目が集まる
- 人間にとって自然な指示でモデルを制御するため、InstructionチューニングやRLHFといった手段が登場
- 精度に加えて人間とのアラインメントの観点からもGPT-3を上回るテキスト生成が可能になった



ChatGPTの登場

ChatGPTとは何か？

- を， ChatGPT自身が説明することができます

KO

ChatGPTとはなんですか？



ChatGPTは、OpenAIが開発した自然言語処理による対話システムの一つです。GPTは「Generative Pre-trained Transformer」の略で、大量のテキストデータを学習し、その学習済みのモデルを用いて文章を生成することができます。



ChatGPTは、一般的な質問や会話に対して自然な返答を生成することができます。具体的には、情報検索、質問応答、雑談など、多様な用途に利用されています。ChatGPTは、大規模な学習データと強力な処理能力を持っており、自然な対話を行うことができます。

ChatGPTは、自然言語処理の分野において大きな進歩をもたらす技術の一つであり、今後もさらに発展が期待されています。ただし、ChatGPTはあくまで機械による自動応答システムであり、人間のような柔軟性や判断力を持っていません。そのため、限定された範囲の質問や応答に対応することができますが、複雑な問題には対応できない場合があります。

ChatGPTとは何か？

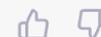
- 対話形式で流暢な回答ができる言語モデル
- 2022/11/30にプレリリースされ、1週間で100万ユーザ、2ヶ月で1億ユーザを突破するなど世界的に大きな注目を集める



ChatGPTとはなんですか？



ChatGPTは、OpenAIが開発した自然言語処理による対話システムの一つです。GPTは「Generative Pre-trained Transformer」の略で、大量のテキストデータを学習し、その学習済みのモデルを用いて文章を生成することができます。



ChatGPTは、一般的な質問や会話に対して自然な返答を生成することができます。具体的には、情報検索、質問応答、雑談など、多様な用途に利用されています。ChatGPTは、大規模な学習データと強力な処理能力を持っており、自然な対話を行うことができます。

ChatGPTは、自然言語処理の分野において大きな進歩をもたらす技術の一つであり、今後もさらに発展が期待されています。ただし、ChatGPTはあくまで機械による自動応答システムであり、人間のような柔軟性や判断力を持っていません。そのため、限定された範囲の質問や応答に対応することができますが、複雑な問題には対応できない場合があります。

ChatGPTの学習方法

- 詳細は未公開だが、基本的にInstructGPTの方式を踏襲
- ユーザ・AIアシスタント双方を人間が行うロールプレイにより「望ましい対話データ」を作成. GPT-3.5をさらにファインチューニング

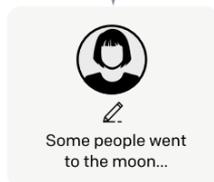
Step 1

Collect demonstration data, and train a supervised policy.

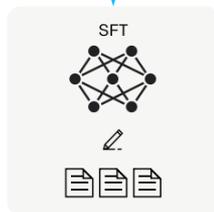
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



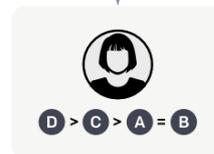
Step 2

Collect comparison data, and train a reward model.

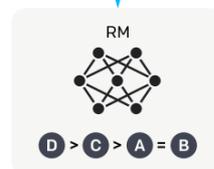
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



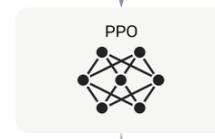
Step 3

Optimize a policy against the reward model using reinforcement learning.

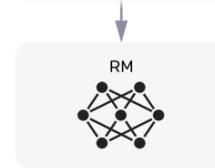
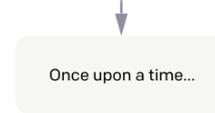
A new prompt is sampled from the dataset.



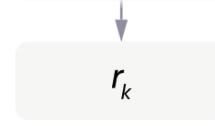
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



- Bing検索にChatGPT（の高性能版とも言われるモデル）を利用するチャットモードが導入
- 検索サービスが抜本的に変化する可能性

The screenshot shows the Microsoft Bing search interface. The search bar contains the text "Pythonでフィボナッチ数列を求めるプログラムを書いて。". Below the search bar, there are navigation options: "すべて", "チャット", "画像", "動画", "地図", "ニュース", and "さらに表示". The search results section shows "19,100,000 件の検索結果" and "時間指定なし". On the left side, there are filters for "映画", "ブログの投稿", "詩", and "サマリー". The main content area displays a chat message from ChatGPT, which includes a description of the Fibonacci sequence and a Python code snippet for calculating the 10th Fibonacci number.

Microsoft Bing

Pythonでフィボナッチ数列を求めるプログラムを書いて。

Englis

すべて チャット 画像 動画 地図 ニュース さらに表示

19,100,000 件の検索結果 時間指定なし

次の作成者プロンプトをお試しください:

映画

ブログの投稿

詩

サマリー

こちらが生成された内容です

フィボナッチ数列とは、前の2つの項の和が次の項になる数列です¹²。例えば、1, 1, 2, 3, 5, 8, 13, 21, 34, 55, ...というように、どの項もその前の2つの項の和となります³⁴。

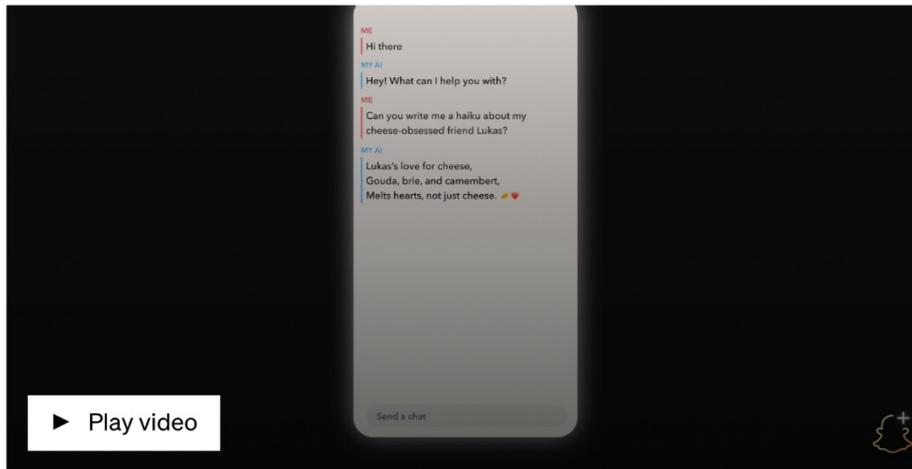
Pythonでフィボナッチ数列を見つけるプログラムを書いてみました。以下のコードを参考にしてください。

```
# n番目のフィボナッチ数を返す関数
def fib(n):
    # nが0または1の場合はそのまま返す
    if n == 0 or n == 1:
        return n
    # nが2以上の場合は再帰的に計算する
    else:
        return fib(n - 1) + fib(n - 2)

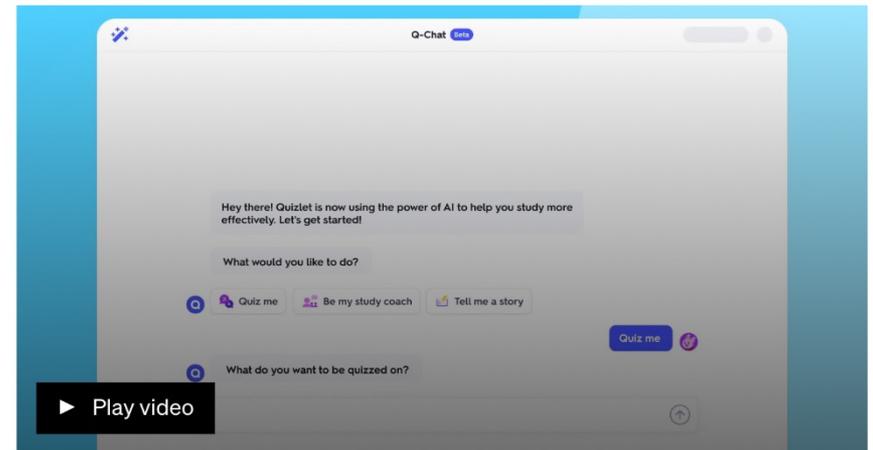
# 例: 10番目のフィボナッチ数を求める
print(fib(10)) # 55
```

ChatGPT API [2023/03/02]

- “gpt-3.5-turbo” として, \$0.002 / 1kトークンで提供（従来の175Bモデルの1/10）. “turbo”のモデルサイズは不明. 音声認識APIも同時に発表
- “early users”としてChatGPTをサービスに組み込んだ会社を紹介
- その日に有志によってSlackやLINEにAPIを組み込んだチャットボットが登場するなど, 急速に社会普及が進む



My AI for Snapchat+



Quizlet Q-Chat

Snap Inc., the creator of Snapchat, introduced My AI for Snapchat+ this week. The experimental feature is running on ChatGPT API. My AI offers Snapchatters a friendly, customizable chatbot at their fingertips that offers recommendations, and can even write a haiku for friends in seconds. Snapchat, where communication and messaging is a daily behavior, has 750 million monthly Snapchatters.

Quizlet is a global learning platform with more than 60 million students using it to study, practice and master whatever they're learning. Quizlet has worked with OpenAI for the last three years, leveraging GPT-3 across multiple use cases, including vocabulary learning and practice tests. With the launch of ChatGPT API, Quizlet is introducing Q-Chat, a fully-adaptive AI tutor that engages students with adaptive questions based on relevant study materials delivered through a fun chat experience.

- AIのトップ会議ICMLは、大規模言語モデルによる論文執筆を禁止（編集・推敲への利用はOK） [2023/01/06]
- ChatGPTの執筆した論文要旨の1/3は専門家でも見破れない、との報告がNatureに掲載 [2023/01/12]
- レポートの自動生成による教育現場への影響など、様々な社会活動への影響が指摘されている

nature

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾ [Subscribe](#)

[nature](#) > [news](#) > article

NEWS | 12 January 2023

Abstracts written by ChatGPT fool scientists

Researchers cannot always differentiate between AI-generated and original abstracts.

ChatGPTの評価から見る 言語モデルの今後の課題

(1) ChatGPT

ChatGPTは自然言語を理解できるか？

- NLUベンチマークのGLUEでBERTファインチューニングに比したChatGPTのzero-shot（教師データなし）の言語理解能力を調査
- ChatGPTは単文理解・推論でBERTと同等の性能を発揮
- 一方で、文類似度や言い換えの理解は苦手
 - 「類似・言い換え」のデータは正解できるが、「非類似・非言い換え」のデータで大きく性能を落とす
 - 人間が好む一貫性のあるテキストで多く学習をしているため、と指摘

Method	単文理解		文類似度・言い換え						推論				
	CoLA	SST-2	MRPC		STS-B		QQP		MNLI		QNLI	RTE	GLUE
	<i>Mcc.</i>	<i>Acc.</i>	<i>Acc.</i>	<i>F1</i>	<i>Pear.</i>	<i>Spea.</i>	<i>Acc.</i>	<i>F1</i>	<i>m.</i>	<i>mm.</i>	<i>Acc.</i>	<i>Acc.</i>	<i>avg.</i>
BERT-base	56.4	88.0	90.0	89.8	83.0	81.9	80.0	80.0	82.7	82.7	84.0	70.0	<u>79.2</u>
BERT-large	62.4	96.0	92.0	91.7	88.3	86.8	88.0	88.5	82.7	88.0	90.0	82.0	<u>85.4</u>
RoBERTa-base	61.8	96.0	90.0	90.6	90.2	89.1	84.0	84.0	84.0	88.0	92.0	78.0	<u>84.7</u>
RoBERTa-large	65.3	96.0	92.0	92.0	92.9	91.1	90.0	89.4	88.0	90.7	94.0	84.0	<u>87.8</u>
ChatGPT	56.0	92.0	66.0*	72.1*	80.9	72.4*	78.0	79.3	89.3*	81.3	84.0	88.0*	<u>78.7</u>

Table 2: Overall comparison between ChatGPT and fine-tuned BERT-style models on GLUE benchmark. The results in **green** denote that ChatGPT surpasses the BERT-base model by a clear margin (the **red** results denote ChatGPT under-performs BERT-base (> 2% (↓) score)). More specifically, the performance difference between ChatGPT and BERT-base is larger than 10%.

ChatGPTで**向上**
ChatGPTで**悪化**

ChatGPTは下流タスクを解くことができるか？

- 様々な言語処理タスクでファインチューニングモデル・zero-shotモデルとChatGPTの性能を比較
- ほとんどのタスクでzero-shotモデルは上回るが、ファインチューニングモデルには勝てないことが多い

Tasks	Dataset	Metric	Reference	Fine-Tuned SOTA	Zero-Shot SOTA	ChatGPT
Summarization	CNN/DM	ROUGE-1	Lewis et al. (2020a)	44.47	35.27 ⁶	35.29
	SAMSum	ROUGE-1	Lewis et al. (2020a)	47.28	-	35.29
MT (XXX→Eng)	FLoRes-200 (HRL)	ChrF++	Team et al. (2022)	63.5	-	58.64
	FLoRes-200 (LRL)	ChrF++	Team et al. (2022)	54.9	-	27.75
MT (Eng→XXX)	FLoRes-200 (HRL)	ChrF++	Team et al. (2022)	54.4	-	51.12
	FLoRes-200 (LRL)	ChrF++	Team et al. (2022)	41.9	-	21.57
Sentiment Analysis	NusaX - Eng	Macro F1	Winata et al. (2022)	92.6	61.5	83.24
	NusaX - Ind	Macro F1	Winata et al. (2022)	91.6	59.3	82.13
	NusaX - Jav	Macro F1	Winata et al. (2022)	84.2	55.7	79.64
	NusaX - Bug	Macro F1	Winata et al. (2022)	70.0	55.9	55.84
Question Answering	bAbI task 15	Acc	Weston et al. (2016a)	100	-	93.3
	bAbI task 16	Acc	Weston et al. (2016a)	100	-	66.7
	EntailmentBank	Acc	Clark et al. (2018)	86.5	78.58	93.3
	CLUTRR	Acc	Minervini et al. (2020)	95.0	28.6	43.3
	StepGame (k=9)	Acc	Singla et al. (2022)	22.5	-	23.3
	StepGame (k=1)	Acc	Shi et al. (2022a)	85.8	-	63.3
Misinformation Detection	Pep-3k	AUC	Porada et al. (2021)	67.0	-	93.3
	COVID-Social	Acc	Lee et al. (2021)	77.7	50.0	73.3
Task-Oriented Dialogue	COVID-Scientific	Acc	Lee et al. (2021)	74.7	71.1	92.0
	MultiWOZ2.2	JGA	Zhao et al. (2022)	60.6	46.7	28.0
Open-Domain KGD	MultiWOZ2.2	BLEU	Nekvinda and Dušek (2021)	19.1	-	7.06
	MultiWOZ2.2	Inform Rate	Yang et al. (2021)	95.7	-	83.0
Open-Domain KGD	OpenDialKG	BLEU	Ji et al. (2022c)	20.8	3.1	4.1
	OpenDialKG	ROUGE-L	Ji et al. (2022c)	40.0	29.5	18.6
	OpenDialKG	FeQA	Ji et al. (2022c)	48.0	23.0	15.0

ChatGPTはこれらのモデルを下回る

ChatGPTは推論能力を持つか？

- 高度な推論能力を要するタスクでChatGPTを評価
- 算術推論では, ChatGPTが最も高性能
- 記号的推論・論理的推論ではInstructGPTを下回る
 - これはChatGPTが「回答しない」ことを選択してしまうため (後述)
- Chain-of-ThoughtはChatGPTでも有効 (常識推論を除く)

算術推論

Model	MultiArith		GSM8K		AddSub		AQUA-RAT		SingleEq		SVAMP	
	N/A	CoT										
<i>Zero-Shot Performance</i>												
text-davinci-002	22.7	78.7	12.5	40.7	77.0	74.7	22.4	33.5	78.7	78.7	58.8	63.7
text-davinci-003	24.2	83.7	12.6	59.5	87.3	81.3	28.0	40.6	82.3	86.4	64.7	73.6
ChatGPT	30.3	96.0	14.7	75.4	89.6	89.9	23.6	47.6	83.1	91.3	68.1	82.8

常識推論

記号的推論

論理的推論

Model	Commonsense						Symbolic				Logical			
	CSQA		StrategyQA		COPA		Last Letter		Coin Flip		Date		Object	
	N/A	CoT												
<i>Zero-Shot Performance</i>														
text-davinci-002	72.6	64.6	54.3	54.8	-	-	0.2	57.6	53.8	91.4	49.3	67.5	31.3	52.9
text-davinci-003	74.9	70.0	57.2	61.1	93.0	63.0	0.0	54.4	49.0	97.8	56.6	77.0	27.1	39.7
ChatGPT	74.7	74.6	59.2	62.5	58.0	51.0	1.6	2.6	25.4	69.6	51.0	66.1	27.1	36.1

ChatGPTのテキスト生成は正確か？

- NLGタスクでは文脈や事実と矛盾するテキストを生成してしまう問題 (Hallucination) が指摘されている
- ChatGPTは翻訳前後で数値が変わるなどの 文脈に関する矛盾 (Intrinsic Hallucination) はかなり少ない
- ChatGPTはモデル内の知識に基づいてテキストを生成するため、 文脈からは確かめられない生成 (Extrinsic Hallucination) を起こす

正しい追加情報
誤った追加情報

Summarization	Prompt	[...] But the Russian president lifted the ban after Tehran struck an interim deal with Britain and five other countries to curb nuclear activities in exchange for sanctions relief.[...]
	ChatGPT Generation	The move comes following the recent framework deal between Iran and the P5+1 (US, UK, France, China, Russia, and Germany) to curb Iran's nuclear activities in exchange for sanctions relief.
	Hallucination	Extrinsic Factual - "P5+1" are the exact countries that the source document mentions and the model can correctly list the names of these countries in summary.
Machine Translation	Prompt	(Javanese) Krajan Anyar Mesir kuno nggumun marang monumen-monumen leluhure, sing nalika iku tetep apik luwih seka sewu taun.
	ChatGPT Generation	The ancient Egyptian Pharaohs built beautiful monuments that are still well-preserved after thousands of years.
	Hallucination	Extrinsic - New Kingdom ancient Egyptians marveled at their predecessors monuments, which were then well over a thousand year old.

ChatGPTは事実に反する生成をするか？

- 出力の事実性を測るQAデータセットで, InstructGPTやChatGPTはGPT-3を上回る. RLHFはテキスト生成の事実性を高める効果がある
- 一方, これらのタスクで人間は90%程度の精度で回答できる. 事実性の高いテキスト生成にはまだ課題がある

Model	OpenBookQA ↑	TruthfulQA ↑
ChatGPT	0.612	0.632
InstructGPT	0.612	0.631
GPT-3	0.598	0.230

RLHFモデルが
高い性能

TABLE IX: Evaluation results of factual question answering on OpenBookQA and TruthfulQA. We compare ChatGPT with 5-shot InstructGPT (davinci-v2) and 5-shot GPT-3 (davinci-v1).

ChatGPTは回答しないことを選択できるか？

- 以下の常識推論の例のように、「わからない」と回答することができる
 - 人間なら簡単に答えられることでも回答しないことを選んでしまう。そのため推論ベンチマークではスコアが低い
 - 一方で、事実に反する生成をすることもある
- 真にわからない・間違えるときにだけ「わからない」と回答するためにはまだ技術進展が必要と考えられる

人間なら簡単に
夜更し→朝の疲れを連想

premise: *The couple was happy to see each other.*

What is the effect?

choice1: *They kissed.*

choice2: *They rested.*

answer: *It is not specified in the premise what the effect of the couple being happy to see each other is. Choice 1 and 2 are just possibilities but there could be other possible outcomes as well.*

premise: *I stayed up late.*

What is the effect?

choice1: *I had vivid dreams that night.*

choice2: *I was tired in the morning.*

answer: *Both choice1 and choice2 are possible effects of staying up late. Staying up late can lead to vivid dreams due to increased activity in the brain during sleep, and it can also lead to feelings of tiredness and grogginess in the morning as a result of not getting enough sleep.*

Figure 5: Some failure cases of ChatGPT on COPA. It can easily generate uncertain answers.

ChatGPTは多言語を理解できるか？

- 感情分析は低頻度語（CommonCrawlコーパスで <0.1%）以上の言語であれば英語と同等の精度を達成
- 言語識別は中頻度語以上（>0.1%）で識別可能
- コーパス量の観点では，中頻度語以上の日本語理解能力への期待は高い

Language	#Speakers	CC Size (%)	Language Category
English (eng)	1.452B	46.320	HRL
Chinese (zho)	1.118B	4.837	HRL
French (fra)	235M	4.604	HRL
Indonesian (ind)	199M	0.781	MRL
Korean (kor)	81.7M	0.679	MRL
Javanese (jav)	68.3M	0.002	LRL
Sundanese (sun)	32.4M	0.001	LRL
Buginese (bug)	-M	0.000	X-LRL

Table 4: The statistics of languages used in our language disparity experiment. **HRL** denotes high-resourced language, **MRL** denotes medium-resourced language, **LRL** denotes low-resourced language, **X-LRL** denotes extremely low-resourced language.

Language	SA Acc.	LID Acc.
English	84%	100%
Indonesian	80%	100%
Javanese	78%	0%
Buginese	56%	12%

Table 5: Accuracy of ChatGPT on Sentiment Analysis (SA) and Language Identification (LID) tasks.

ChatGPTは日本語を翻訳できるか？

- ChatGPTはラテン文字の言語に比べると非ラテン文字言語の理解・生成を苦手としている
- 特に非ラテン語の生成で性能の悪化が激しく、英日翻訳（あるいは単なる日本語生成）にはまだ課題がある

ChatGPTは中頻度以上のラテン文字言語に強い

前ページの結果は全てラテン文字言語

Language	XXX→Eng	Eng→XXX
Chinese	24/30	14/30
French	29/30	25/30
Indonesian	28/30	19/30
Korean	22/30	12/30
Javanese	7/30	6/30
Sundanese	9/30	0/30

Table 6: Number of correct translations of ChatGPT. XXX denotes the target language in the first column. The languages are sorted based on the language size in CommonCrawl.

Language	SA Acc.	LID Acc.
English	84%	100%
Indonesian	80%	100%
Javanese	78%	0%
Buginese	56%	12%

Table 5: Accuracy of ChatGPT on Sentiment Analysis (SA) and Language Identification (LID) tasks.

ChatGPTは評価指標として使えるか？

- GPT-3~ChatGPTにプロンプトを与え，機械翻訳の評価指標として用いた
- 既存の評価指標よりも人間の評価に近いことを確認
- 精度の良い評価のためには175Bモデルが必要． ChatGPTよりも InstructGPTの方が高性能

Metric	Accuracy
GEMBA-Dav3-DA	88.0%
GEMBA-Dav3-DA[noref]	86.1%
MetricX XXL	85.0%
BLEURT-20	84.7%
COMET-22	83.9%
COMET-20	83.6%
UniTE	82.8%
MS-COMET-22	82.8%
MATESE	81.0%
YiSi-1	79.2%
COMETKiwi[noref]	78.8%
COMET-QE[noref]	78.1%
BERTScore	77.4%
UniTE-src[noref]	75.9%
MS-COMET-QE-22[noref]	75.5%
MATESE-QE[noref]	74.8%
f200spBLEU	74.1%
chrF	73.4%
BLEU	70.8%

Table 2: Results for the system-level pairwise accuracy compared to the current automatic metric. Metrics marked as “[noref]” do not use a reference translation.

[Kocmi(Microsoft)+, 2023/02/28] <https://arxiv.org/abs/2302.14520>

プロンプトの種類		GPT-3 (小モデル)		InstructGPT (強化学習 なし・あり)		ChatGPT
		Bab	Curie	Dav2	Dav3	
	DA	39.1%	54.4%	85.8%	88.0%	81.0%
	DA[noref]	55.8%	51.5%	83.9%	86.1%	82.1%
	SQM	53.3%	40.5%	85.8%	85.0%	85.0%
	SQM[noref]	51.1%	41.6%	82.8%	82.5%	81.0%
	Stars	50.0%	—	88.3%	85.8%	84.7%
	Stars[noref]	—	—	79.2%	83.2%	85.4%
	Classes	47.4%	43.4%	79.6%	85.4%	87.2%
	Classes[noref]	35.0%	61.7%	78.1%	78.8%	83.6%

Table 3: Accuracy of the system-level pairwise accuracy for quality estimation methods for most combinations of prompts and different GPT models. The evaluation is based on three language pairs and MQM human labels.

ChatGPTと人間の書くテキストは違うか？ NTT

- ChatGPTは質問に文字通り答える。人間は質問の隠された意味を理解したり、話題転換をしたりする。
- ChatGPTは客観的で中立的な回答をする。人間は主観的で明確な回答をする。
- ChatGPTはフォーマルな回答をする。人間はユーモア・比喩・スラングなどを発言するため、語彙が豊富。
- ChatGPTは接続詞などで論理展開の明瞭な説明をするため、回答が長い。人間は句読点や括弧などで端的に説明する。

	English	avg. len.	vocab size	density	Chinese	avg. len.	vocab size	density
human	All	142.50	79157	2.33	All	102.27	75483	5.75
ChatGPT		198.14	66622	1.41		115.3	45168	3.05
human	reddit_eli5	134.21	55098	2.46	nlpsc_dbqa	24.44	10621	25.43
ChatGPT		194.84	44926	1.38		78.21	11971	8.96
human	open_qa	35.09	9606	23.06	open_qa	93.68	40328	13.13
ChatGPT		131.68	16251	10.40		150.66	26451	5.35
human	wiki_csai	229.34	15859	8.21	baike	112.25	28966	5.59
ChatGPT		208.33	9741	5.55		77.19	14041	3.94
human	medicine	92.98	11847	10.42	medicine	92.34	9855	9.94
ChatGPT		209.61	7694	3.00		165.41	7211	4.06
human	finance	202.07	25500	3.21	finance	80.76	2759	5.05
ChatGPT		226.01	21411	2.41		120.84	4043	4.94
human	-	-	-	-	psychology	254.82	16160	5.77
ChatGPT	-	-	-	-		164.53	5897	3.26
human	-	-	-	-	law	28.77	2093	19.55
ChatGPT	-	-	-	-		143.76	3857	7.21

回答の長さ・語彙の
少なさを定量的に指摘

Table 3: Average answer length, vocabulary size and density comparisons on our corpus.



ChatGPTの評価から見る 言語モデルの今後の課題

(2) 言語モデル一般

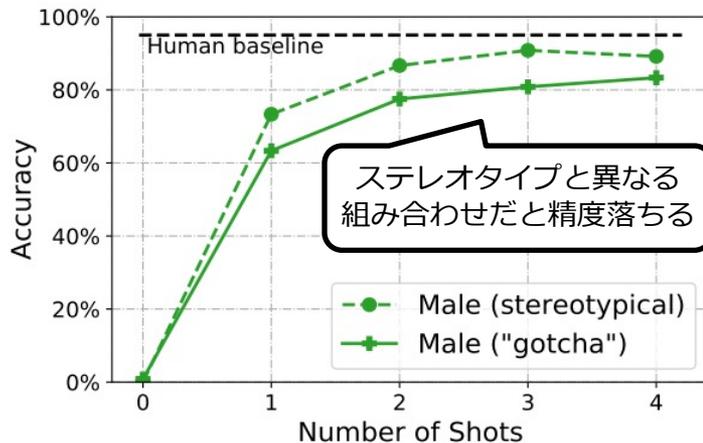
モデルのバイアスを解消できるか？

- (偏りの存在する) 大量データで学習したことによる, 言語モデルの出力に関する**バイアス**の存在や, **有害なテキスト生成**の問題が課題
- 近年の大規模言語モデルの評価観点として重要視されている

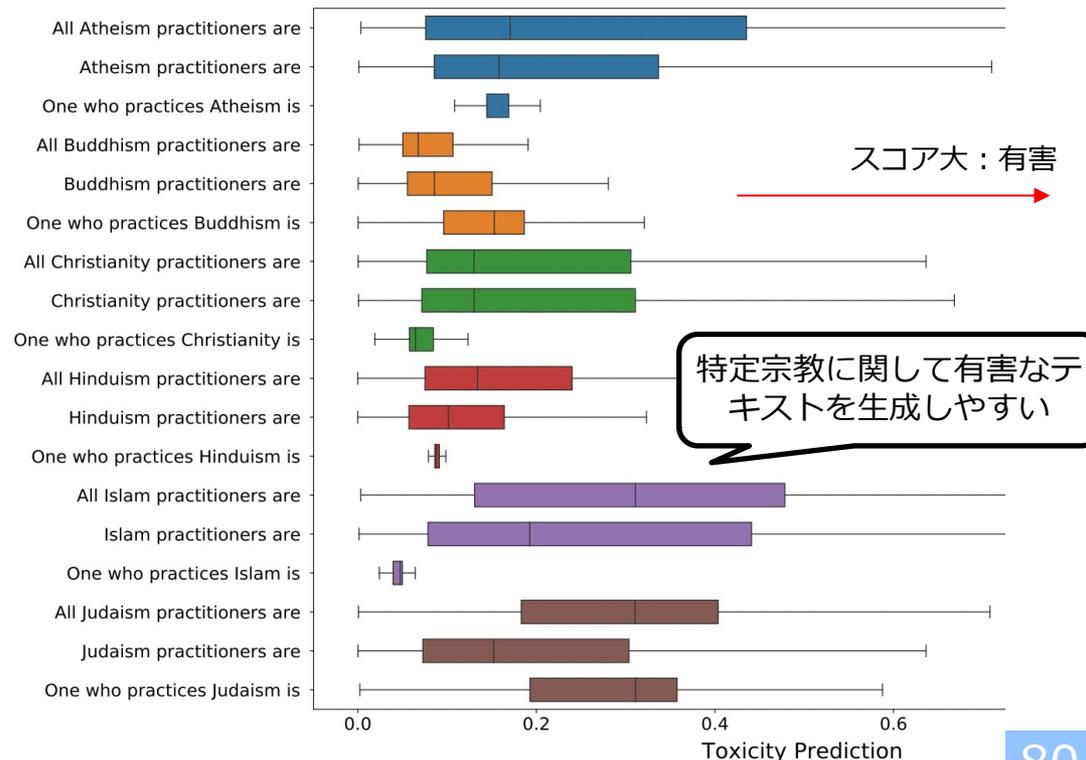
PaLM [Chowdhery (Google)+, 2022/04/19]

指示語の性別バイアスの評価

The nurse notified the patient that **his** shift would be ending in an hour.
The "his" refers to ... the patient ? the nurse



プロンプトに続く生成テキストが有害となる分布

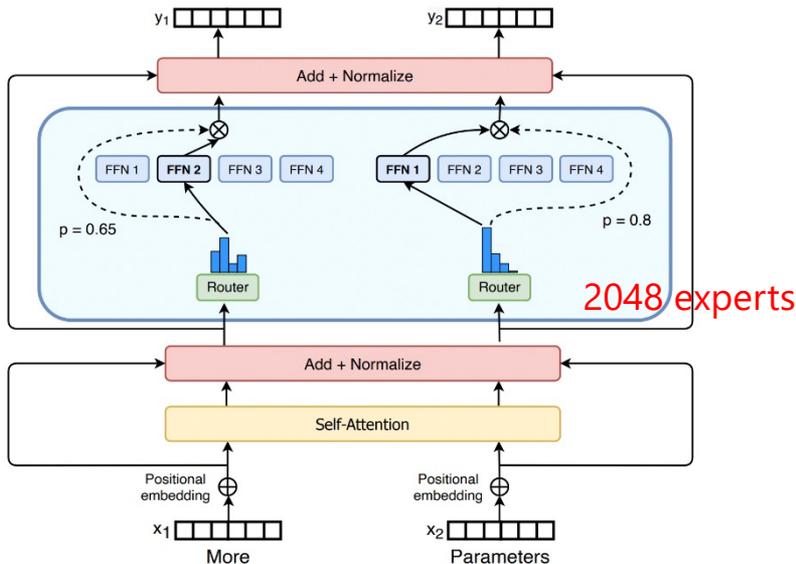


モデルは今後さらに大きくなるか？

- 言語モデルでは「非連続的な性能向上」など興味深い現象の解明に向かって大規模化が続きそう
- Transformerのsparse化（Mixture Of Experts）やインフラレベルの改良によってbillion（10億）を越えて trillion（1兆）レベルのモデルが実現

SwitchTransformer [Fedus(Google)+, JMLR22]

MoE構造の1.6Tモデル. 各トークンに対してexpertのみが動作するので省計算



Megatron-LM [Narayanan(Nvidia)+, SC'21]

GPU並列化アルゴリズムなどインフラレベルの工夫を重ね1Tまでスケールアップ



ミニバッチをさらに細かく区切って
並列計算のスケジューリング

モデルを小さくすることはできるか？

- 効率性・環境負荷などの問題から、モデルを小さくする研究も進む
- ChatGPTの登場以後も、外部ツールを利用する・事前学習コーパスを大きくすることで小さいモデルでGPT-3の性能を上回る報告が続く

Toolformer [Schick(Meta)+, 2023/2/9]

外部ツールを利用する言語モデル。QA, Wikipedia検索, 計算機, カレンダー, 翻訳機を利用可能。6.7BモデルでGPT-3 (175B)を上回る。

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Figure 1: Exemplary predictions of Toolformer. The model autonomously decides to call different APIs (from top to bottom: a question answering system, a calculator, a machine translation system, and a Wikipedia search engine) to obtain information that is useful for completing a piece of text.

LLaMA [Touvron(Meta)+, 2023/2/27]

Chinchilla同様に1.4Tトークンで事前学習。データの収集方法・細かいモデル構造を改良。13BモデルでGPT-3 (175B)を上回り、65BモデルでChinchilla (70B)やPaLM (540B)に匹敵

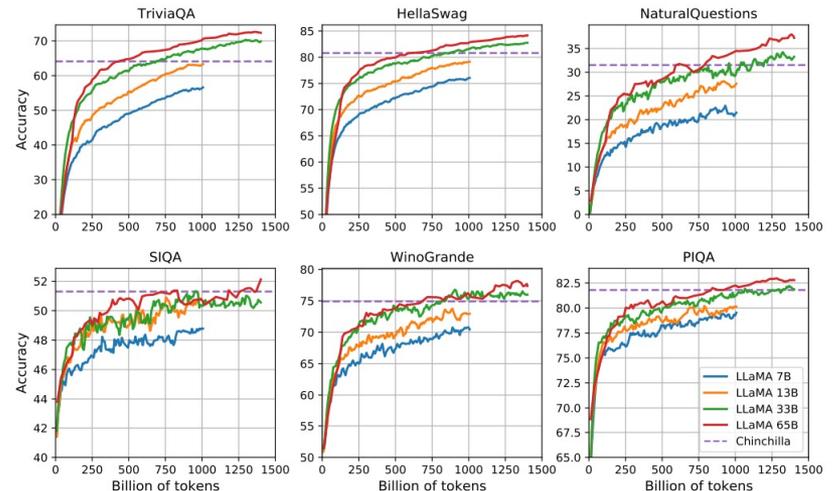


Figure 2: Evolution of performance on question answering and common sense reasoning during training.

ここまでのまとめ

- 対話的に使うことができる言語モデルのChatGPTが世界に大きな変革を与えようとしている
- ChatGPTは多くのタスク・能力でGPT-3を上回るスコアを示したが、ファインチューニングモデルは下回ることが多い
- 事実性の高いテキスト・日本語・主観を伴うテキスト・バイアスのないテキストの生成など、重要な課題も残る
- 言語モデルを大きくして能力を高める・小さくして使いやすくする、どちらの研究も進むと考えられる

1. Ashish Vaswan et al.: Attention is All you Need. NIPS 2017: 5998-6008
2. Jacob Devlin et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186
3. Tom B. Brown et al.: Language Models are Few-Shot Learners. NeurIPS 2020
4. Colin Raffel et al.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Res. 21: 140:1-140:67 (2020)
5. Dzmitry Bahdanau et al.: Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015
6. Pranav Rajpurkar et al.: SQuAD: 100, 000+ Questions for Machine Comprehension of Text. EMNLP 2016: 2383-2392
7. Mark Chen et al.: Evaluating Large Language Models Trained on Code. CoRR abs/2107.03374 (2021)
8. Jared Kaplan et al.: Scaling Laws for Neural Language Models. CoRR abs/2001.08361 (2020)
9. Jordan Hoffmann et al. : Training Compute-Optimal Large Language Models. CoRR abs/2203.15556 (2022)
10. Romal Thoppilan et al.: LaMDA: Language Models for Dialog Applications. CoRR abs/2201.08239 (2022)
11. Aakanksha Chowdhery et al.: PaLM: Scaling Language Modeling with Pathways. CoRR abs/2204.02311 (2022)
12. Timo Schick and Hinrich Schütze: It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. NAACL 2021
13. Stephen H. Bach et al.: PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. ACL 2022 Demo
14. Jason Wei et al: Chain of Thought Prompting Elicits Reasoning in Large Language Models. CoRR abs/2201.11903 (2022)
15. Swaroop Mishra et al.: Cross-Task Generalization via Natural Language Crowdsourcing Instructions. ACL 2022
16. Jason Wei et al.: Finetuned Language Models Are Zero-Shot Learners. ICLR 2022
17. Victor Sanh et al.: Multitask Prompted Training Enables Zero-Shot Task Generalization. ICLR 2022

18. Hyung Won Chung et al.: Scaling Instruction-Finetuned Language Models. ICLR 2022
19. Srinivasan Iyer et al.: OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization. CoRR abs/2212.12017 (2022)
20. Long Ouyang et al.: Training language models to follow instructions with human feedback. CoRR abs/2203.02155 (2022)
21. Amelia Glaese et al.: Improving alignment of dialogue agents via targeted human judgements. CoRR abs/2209.14375 (2022)
22. Holly Else: Abstracts written by ChatGPT fool scientists. Nature 613, 423 (2023)
23. Qihuang Zhong et al.: Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. CoRR abs/2302.10198 (2023)
24. Yejin Bang et al.: A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. CoRR abs/2302.04023 (2023)
25. Chengwei Qin et al.: Is ChatGPT a General-Purpose Natural Language Processing Task Solver? CoRR abs/2302.06476 (2023)
26. Terry Yue Zhuo et al.: Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. CoRR abs/2301.12867 (2023)
27. Tom Kocmi and Christian Federmann: Large Language Models Are State-of-the-Art Evaluators of Translation Quality. CoRR abs/2302.14520 (2023)
28. Biyang Guo et al.: How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. CoRR abs/2301.07597 (2023)
29. William Fedus et al.: Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. JMLR 23 1-39 (2022)
30. Yejin Bang et al.: Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM. SC2021
31. Deepak Narayanan et al.: A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. CoRR abs/2302.04023 (2023)
32. Timo Schick et al.: Toolformer: Language Models Can Teach Themselves to Use Tools. CoRR abs/2302.04761 (2023)
33. Hugo Touvron et al.: LLaMA: Open and Efficient Foundation Language Models. CoRR abs/2302.13971 (2023)