

人文社会ビジネス科学学術院 ビジネス科学研究群 2024年度 春C

テキストマイニングの実践

day 2

スケジュール

day 2

- 講義 – 自然言語処理の最新動向
- 講義 – テキストマイニングの手順
- 講義&演習 – データ理解

day 3

- 講義&演習 – 演習環境の準備
- 講義&演習 – テキスト解析 (1)
- 講義&演習 – テキスト解析 (2)

day 4

- 講義&演習 – テキスト分析 (1)
- 演習 – テキスト分析 (実践編)

day 5

- 演習 – テキスト分析 (実践編)

自然言語処理の最新動向

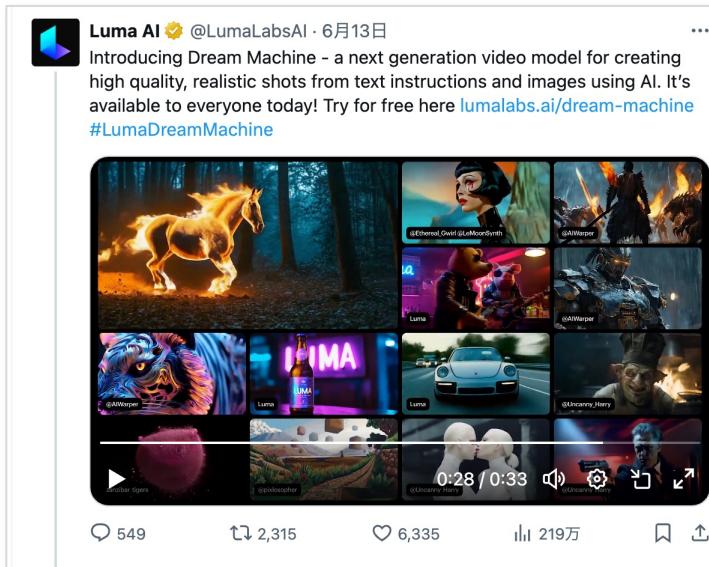
目次

- 大規模言語モデル(LLM)はどんなもの?
- 大規模言語モデル(LLM)の成り立ち
- テキストマイニングの未来図

大規模言語モデル (LLM) はどんなもの?

使ったことがありますか? ➡ 始まりは Transformer という発明!

動画生成ができる

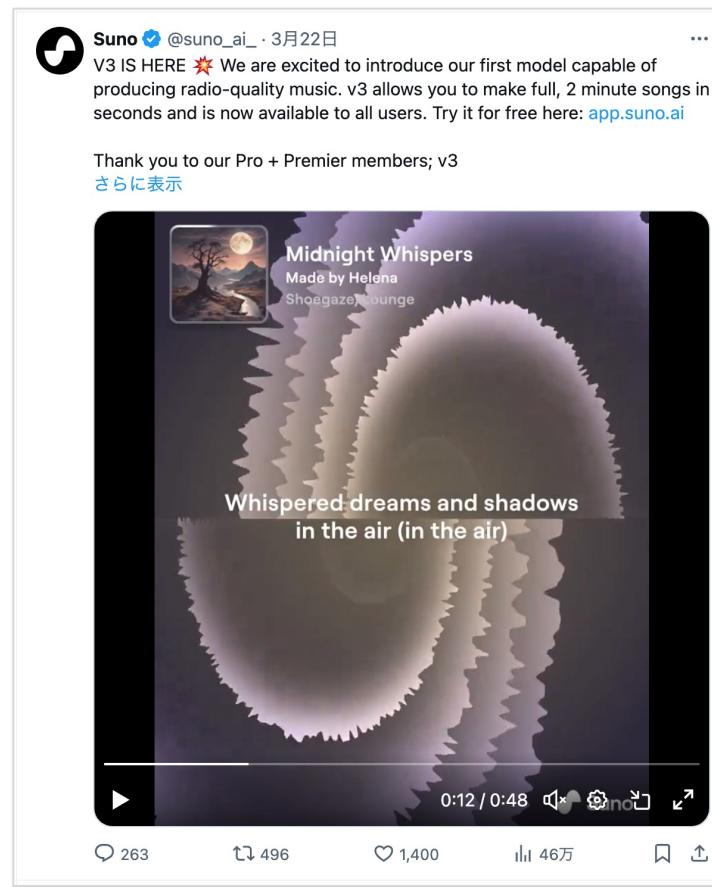


出所: [X \(@Luma AI\)](#)

台本生成ができる
歌詞生成ができる

大規模言語モデル(LLM)

音楽生成ができる



出所: [X \(@Suno\)](#)

CM が作れる

台本: 生成AI (LLM)
動画: 生成AI (動画生成)
歌詞: 生成AI (LLM)
音楽: 生成AI (音楽生成)



出所: [X \(@CultureCrave\)](#)

※ 上の例は具体例ではなくイメージ。OpenAIが開発中の Soraで作成されたもので、AIが商業目的で活用された実例として歓迎する声が上がった一方で、不快な動画などの批判も多数寄せられた。

大規模言語モデル(LLM)、使ってますか?

- OpenAI GPT-4o: 自然言語の入力で、何でもできる
プログラミングができる

コード

次に、以下のコードを使用してワードクラウドを作成します。

```
python
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# テキストを定義
text = """
OpenAI's mission is to ensure that artificial general intelligence (AGI) benefits all humanity. AGI refers to highly autonomous systems that output meaningful results across all domains of knowledge.
"""

# ワードクラウドの設定
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(text)

# ワードクラウドを表示
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off') # 軸を非表示にする
plt.show()
```

このコードでは、指定したテキストからワードクラウドを生成し、表示します。`WordCloud`オブジェクトのパラメータを調整することで、出力の見せ方をカスタマイズできます。例えば、`background_color`を変更することで背景色を設定できます。

曖昧な指示でもサンプルコードを生成

- 表作成(データ整理)ができる

以下の文章を読んで表にまとめてください

ホンダの軽自動車「N-BOX」が首位から陥落した。日本自動車販売協会連合会（自販連）と全国軽自動車協会連合会（全軽自協）が発表した月別販売台数によると、N-BOXは5月の販売台数が14,582台で、前年同月比4.4%増だった。しかし、前年同月比91.9%増だったスズキのスペーシアが首位に躍り出た。現在、生産を停止している。

以下に文章を表にまとめました：

順位	車種	メーカー	車種分類	5月の販売台数	前年同月比	備考
1	スペーシア	スズキ	軽自動車	15,160台	91.9%増	2023年11月に発売した全面改良車が好調な販売に寄与した
2	N-BOX	ホンダ	軽自動車	14,582台	4.4%増	10ヶ月連続首位を維持できず、2位に陥落
3	ヤリス	トヨタ	小型車	-	-	4月の2位から1つ順位を下げた
4	カローラ	トヨタ				
5	シエンタ	トヨタ				
32	プリウス	トヨタ				

前年同月の販売台数も追加してください

以下に前年同月の販売台数も追加した表をまとめました：

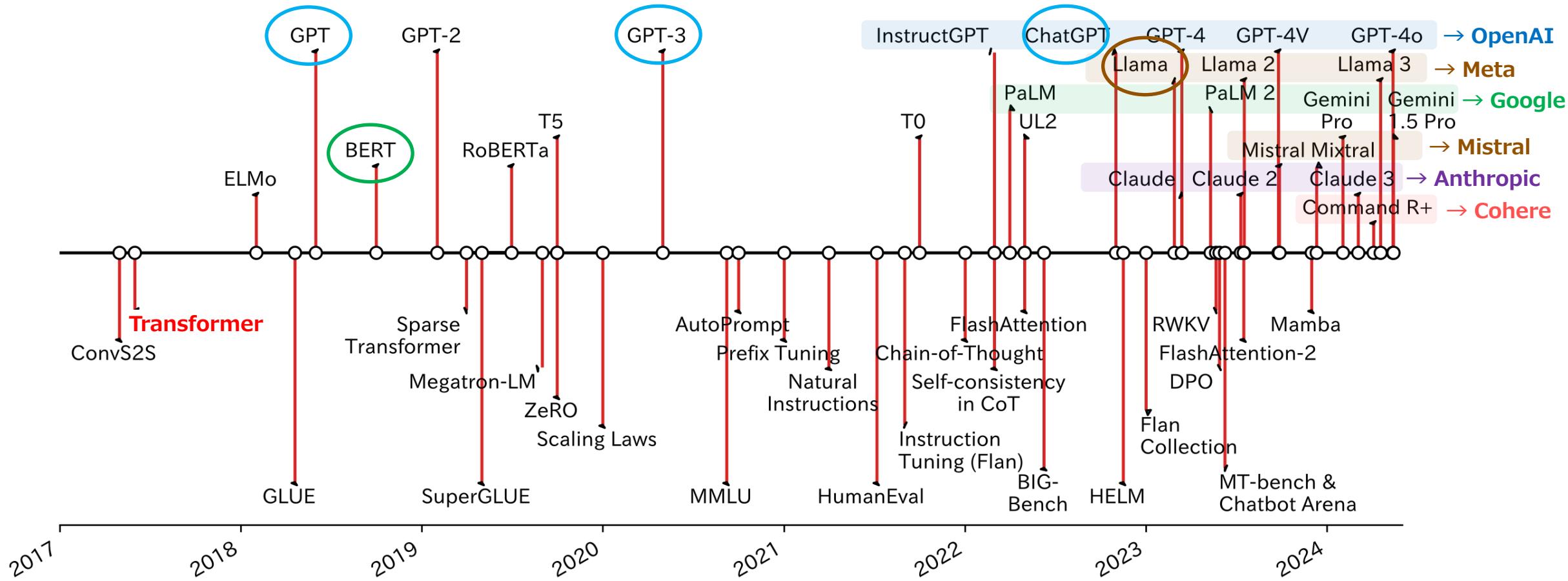
順位	車種	メーカー	車種分類	5月の販売台数	前年同月比	前年同月の販売台数	備考
1	スペーシア	スズキ	軽自動車	15,160台	91.9%増	7,895台	2023年11月に発売した全面改良車が好調な販売に寄与した
2	N-BOX	ホンダ	軽自動車	14,582台	4.4%増	13,968台	10ヶ月連続首位を維持できず、2位に陥落
3	ヤリス	トヨタ	小型車	-	-	-	4月の2位から1つ順位を下げた

大規模言語モデル(LLM)、使ってますか?

● Google NotebookLM (Gemini 1.5 Pro): 論文の読み方が変わる?!

The image shows two screenshots of the Google NotebookLM interface. The left screenshot illustrates the process of registering a PDF document. A red circle highlights the 'ソース' (Sources) section where a PDF file named '2406.13121v1.pdf' is listed. A red box labeled 'PDFを登録' (Register PDF) is overlaid on this area. Another red circle highlights the 'CICプロンプト' (CIC Prompt) section, which contains detailed instructions about the CiC (Corpus-in-Context) prompt design. A red box labeled 'チャットする' (Chat) is overlaid on the bottom right of this section. The right screenshot shows the results of a query. A red circle highlights the 'メモに残す' (Save to Memo) button at the top right of the response. Another red circle highlights the 'CICプロンプト' section, which now includes a summary of the LLMs used in the paper. A red box labeled 'メモに残す' (Save to Memo) is overlaid on the top right of this section. The bottom of both screenshots shows a message from NotebookLM stating 'NotebookLM はまだ不正確な回答をすることがあるため、ご自身で事実確認されることをお勧めします' (NotebookLM may provide inaccurate answers, so we recommend you verify the facts yourself).

大規模言語モデル(LLM)の歴史 (2017-2024年)



言語モデルの
アーキテクチャの発展

事前学習と
ファインチューニング

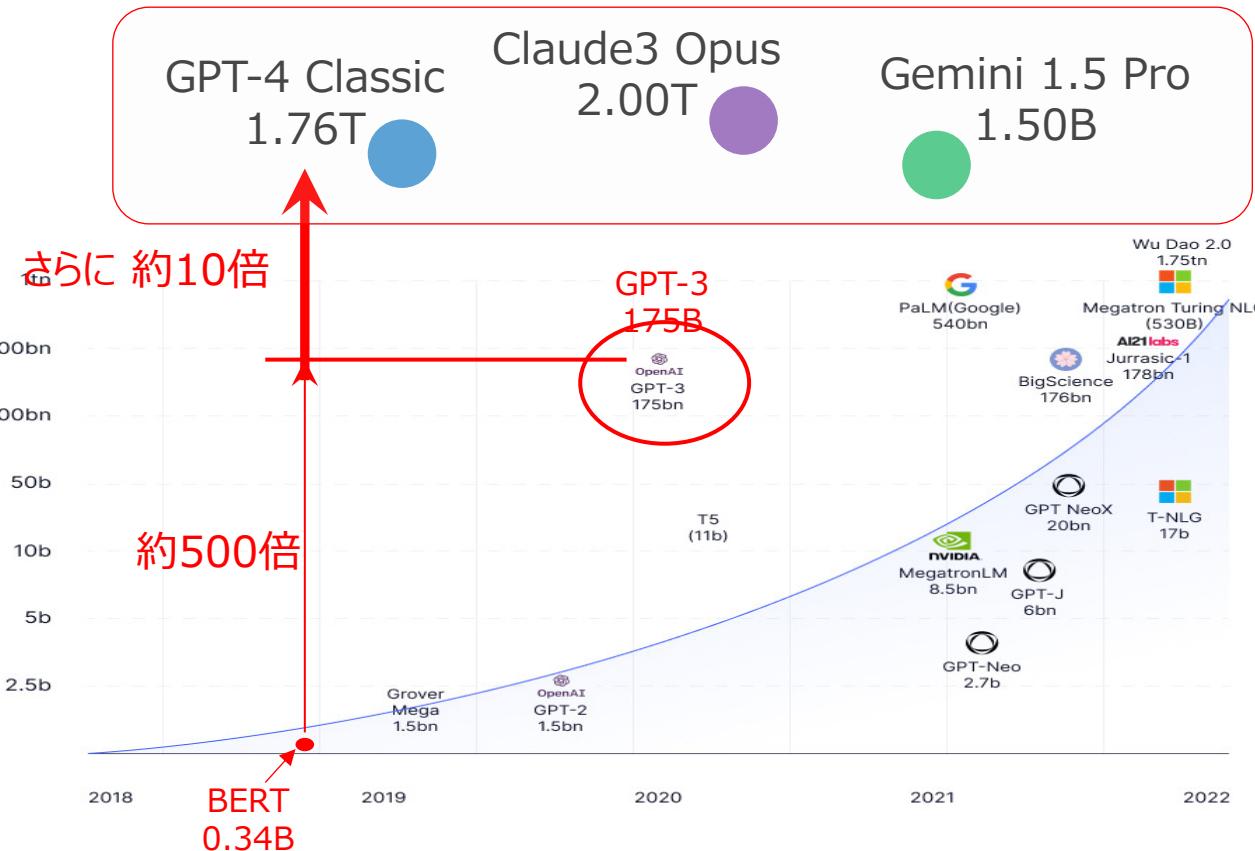
言語モデル
の大規模化

LLMの能力の引き上げ
(指示チューニング、アライメント、思考の連鎖)

汎用的な知能を実現するLLMの開発競争

大規模言語モデル(LLM)は、どのくらい大きいのか

- 大量の計算機資源で、大量データを大きなモデルで学習すれば良いモデルができる
"Scaling laws" [Kaplan (OpenAI)+, 2020/01]



Llama 3 (Meta) の学習:

パラメタ数	テキストデータ	計算時間 (GPU時間)	計算環境 (GPU数)
700億 (70B)	約15兆トークン (43TB 程度)	6.4M GPU時間 (731 GPU年)	24K 枚 (H100)



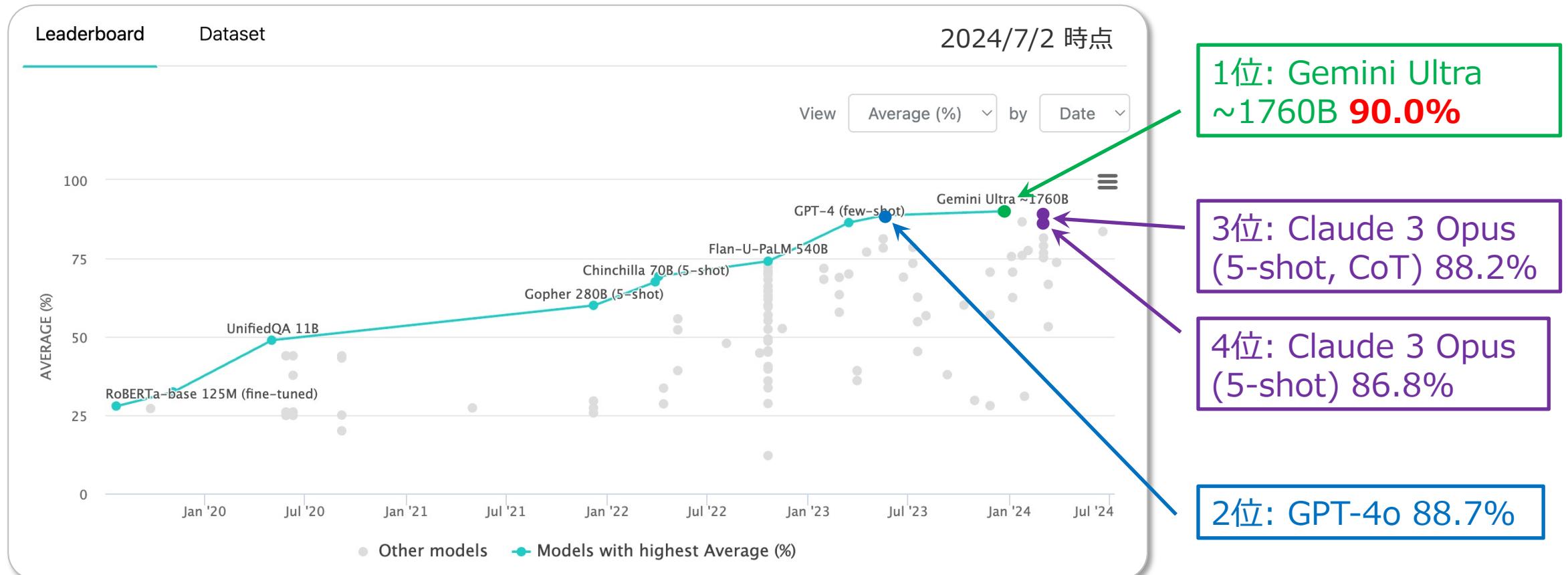
AWS H100 の時間単価で 126億円
※p5.48xlarge(8GPUs): \$98.32/h

さくら高火力 PHY の月額単価で 34億円
※高火力PHY(8GPUs): ¥3,046,120/月

出典: [Llama 3 のモデルカード](#)、Andrej Karpathy 氏のX

大規模言語モデル(LLM)は、どのくらい賢いか

- 2023年に Google Gemini Ultra が数学、歴史、情報科学、法律などの57の領域の問題で構成される MMLU ベンチマークで、**人間の専門家**によるスコア (89.8%) を超えるスコアを達成

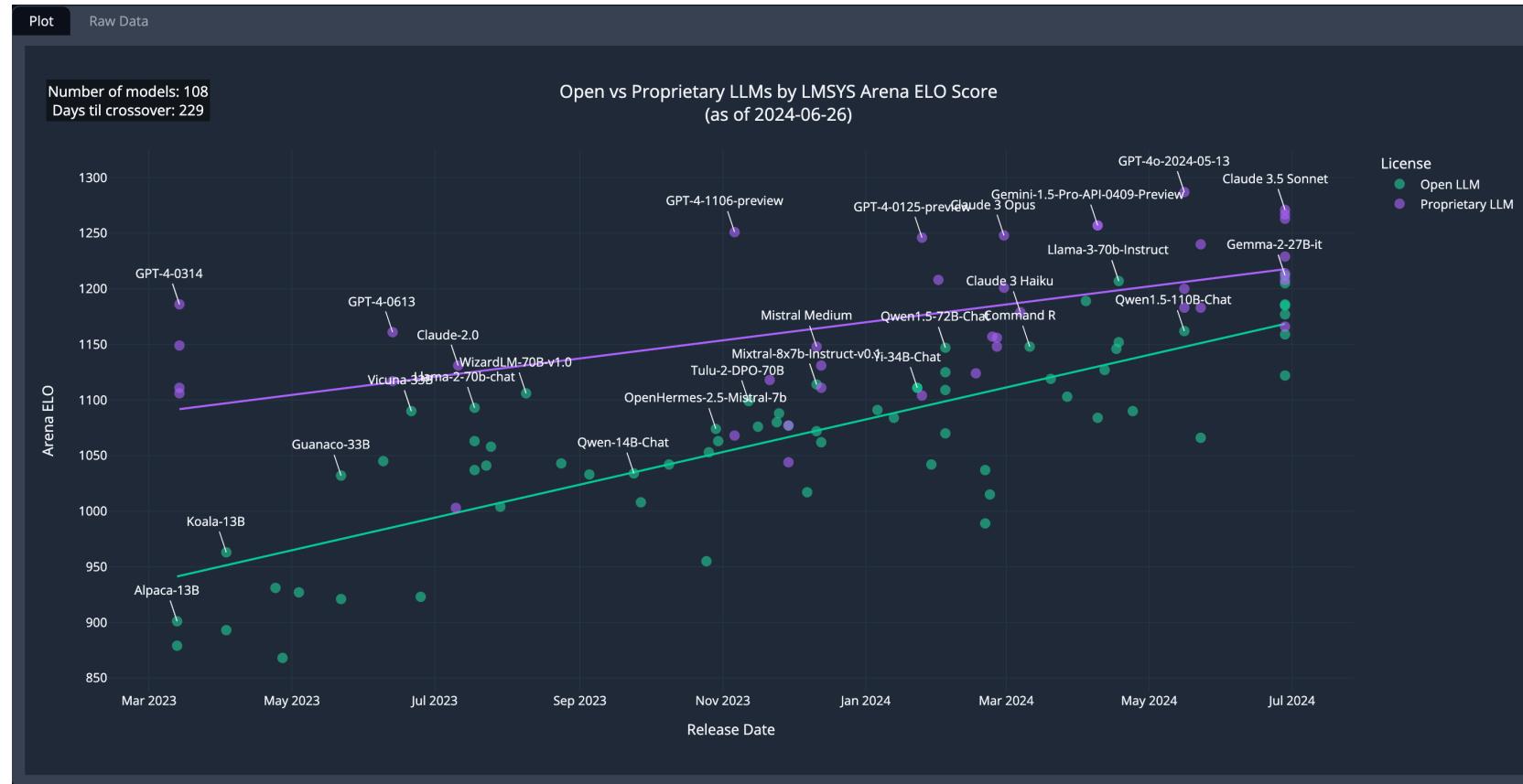


出典: [Paperwithcode の MMLU リーダーボード](#)

オープンな大規模言語モデル(LLM)

- 2023年はじめ頃から短い期間で性能が大きく改善

- オープン(緑)とクローズド(紫)との性能差が狭くなっている → LLMの“民主化”が進む?!



出典: [Chatbot Arena リーダーボード](#)

Chatbot Arena:

- ・異なる2つのLLMの出力に対し、どちらが優れているのかをオンラインでユーザーに投票してもらう
- ・人間から見た性能の評価を行う世界標準的な仕組みとなっている

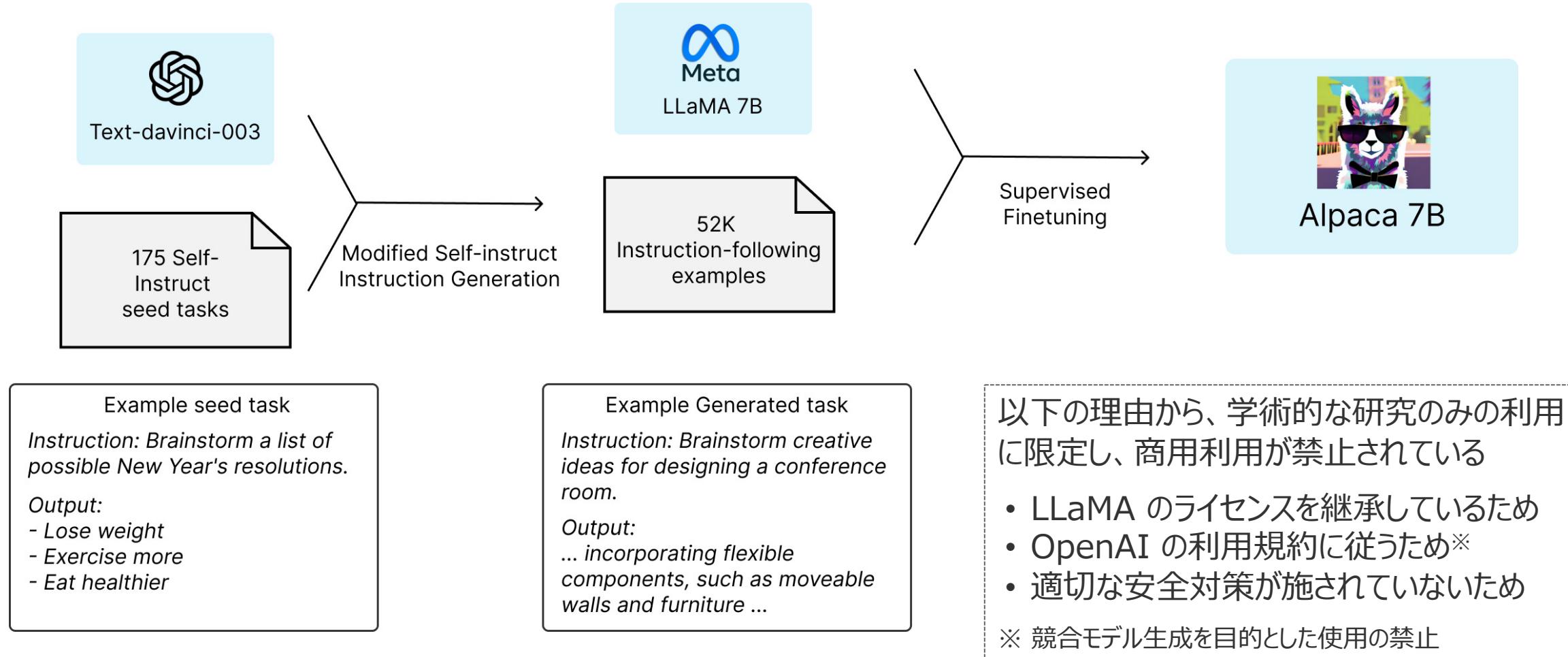
オープンな大規模言語モデル(LLM) ※ 2023/5月末時点

2023/3月の Alpaca 登場以降、ものすごい勢いで GPT-3を超える ローカルLLMモデル が登場、競争が激化

分類	モデル名	アーキ	提供元	リリース	サイズ	ライセンス	日本語	例:「ペンギンはなぜ空を飛べないのですか?」
クラウド	GPT-4	クローズド	OpenAI	2023/3/14	175B	有償API	◎	ペンギンが飛べない理由はいくつかあります。主な理由は、彼らの身体…
	Claude	クローズド	Anthropic	2023/3/14	50B	有償API	◎	ペンギンは空を飛べない主な理由は次のとおりです。1. 翼がないため…
研究利用のみ	LLaMa	LLaMa	Meta	2023/2/24	7-65B	× 制限あり	△	ペンギンは空の中の風に乗り飛ぶことができます。
	Alpaca (LLaMa)	LLaMa	Stanford	2023/3/13	7B	× 制限あり	△	ペンギンは、空の下で彼は翼を叩くことができます。しかし、彼は午前…
	Alpaca-LoRA (LLaMa)	LLaMa	Eric Wang	2023/3/17	7B	× 制限あり	△	ペンギンは空を飛べないためには、彼は彼の背骨を抜く必要があります…
	Guanaco (LLaMa)	LLaMa	Joseph Cheung	2023/3/19	7B	× 制限あり	△	ペンギンが空を飛べない理由は、空を飛ぶことができないためです。ペ…
	GPT4All (LLaMa)	LLaMa	Nomic AI	2023/3/29	13B	× 制限あり	×	…
	Vicuna (LLaMa)	LLaMa	LMSYS	2023/3/30	13B	× 制限あり	○	ペンギンは、自然選択によって空を飛ぶことができないとされています。…
	Koala (LLaMa)	LLaMa	BAIR	2023/4/4	7-13B	× 制限あり	×	I'm sorry, but I'm not quite sure what you mean by …
	Stable-Vicuna (LLaMa)	LLaMa	Stability AI	2023/4/28	13B	× 制限あり	?	…
商用利用可能	Cerebras-GPT	GPT-2	Cerebras	2023/3/28	13B	Apache 2.0	△	ペンギンが空を飛べない理由としては、空を飛ぶことができないことが挙…
	Dolly2.0 (GPT-J-Alpaca)	GPT-J	Databricks	2023/4/12	6B	Apache 2.0	△	よくわかりませんが、ペンギンは空を飛べないのではなく、空を飛ぶことを…
	StableLM-Alpha	GPTNeoX	Stability AI	2023/4/19	7B	CC BY-SA-4.0	△	空を飛べるためには、空気を回し、目的地を知りません。ペンギンは、…
	GPT4All (GPT-J)	GPT-J	Nomic AI	2023/4/24	6B	Apache 2.0	×	It seems like you are having trouble with your writing. …
	RedPajama-INCITE	GPTNeoX	Together	2023/5/5	3-7B	Apache 2.0	△	ペンギンは空を飛べないのでですか。
	MPT-7B	MPT	MosaicML	2023/5/5	7B	Apache 2.0	△	もちろん、飛んでくれます！私は飛んでくれます！でも、飛んでくれない…
日本語特化 (商用利用可能)	open-calm	GPTNeoX	CyberAgent	2023/5/16	1-7B	CC BY-SA-4.0	△	ペンギンが空を飛べない理由は、水中で息ができないから、というのを…
	japanese-gpt-neox-3.6b	GPTNeoX	rinna	2023/5/17	3.6B	MIT	△+	ペンギンは、翼のような推進力を生み出す筋肉がなく、空気力学的に…

(参考) Alpaca [Taori+ (Stanford Univ.), 2023/3/13]

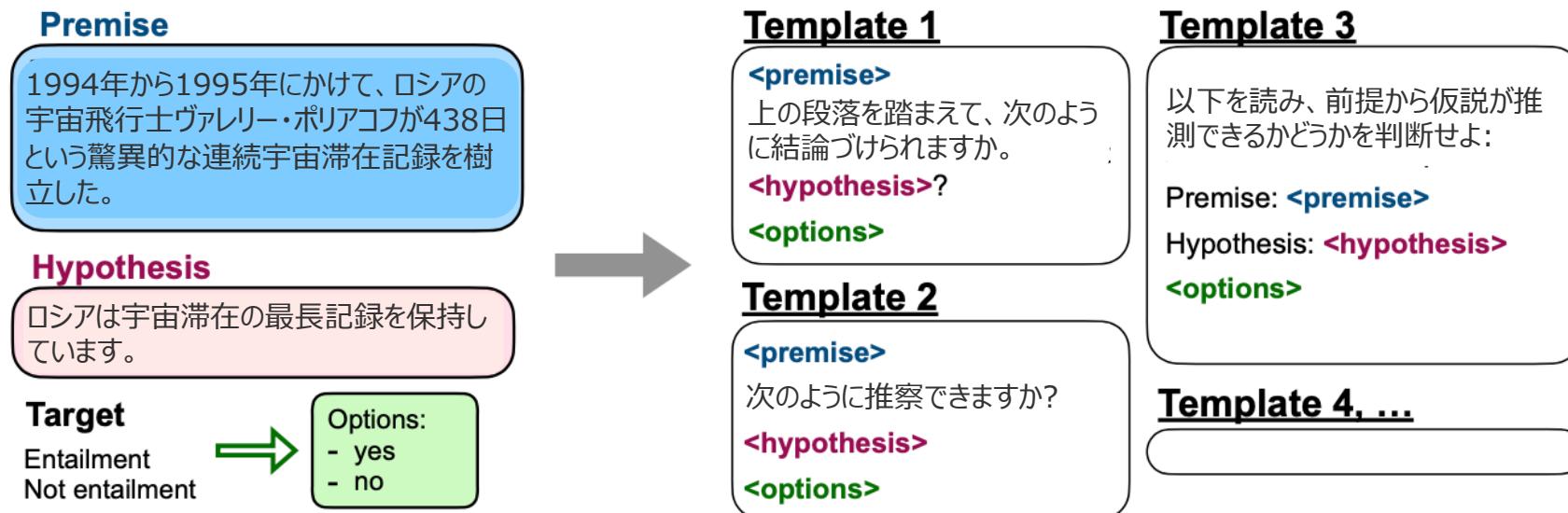
GPT-3 で生成した命令追従型のデータセットで LLaMA (Meta)をファインチューニング(Instruction Tuning)したモデルで、はるかに小さな環境で簡単・安価に GPT-3 に近い性能の再現ができる



引用: <https://crfm.stanford.edu/2023/03/13/alpaca.html>

(参考) Instruction Tuning (FLAN) [Wei+,2022]

- GPT-3 など言語モデルの構造を変えずに、**複数のタスク**でファインチューニングする方法
- タスク毎にテンプレートを用意し、「プロンプト(タスクの指示と事例)+出力」の形式に変換した学習データで言語モデルを追加学習する (=Instruction Tuning)
 - ゼロショットで解くタスクにおいて、GPT-3よりも高い精度を達成
 - 未知のタスクや指示に対しても精度よくテキストを生成



図：自然言語推論(NLI)タスクの例（前提出文が仮設を含意するか否かを自動判定するタスク）

引用: <https://arxiv.org/pdf/2109.01652.pdf>

日本におけるオープンな大規模言語モデル(LLM)

年	月	日	モデル	開発元	規模
2022	11	30	nlp-waseda/gpt2-xl-japanese	早稲田大学	1.5
2023	3	16	okazaki-lab/japanese-gpt2-medium-unidic	東京工業大学	medium
	5	12	retrieva-jp/t5-small-short	Retrieva	S, B, L, XL
		17	cyberagent/open-calm-7b	CyberAgent	1, 3, 7
		31	rinna/japanese-gpt-neox-3.6b-instruction-ppo	Rinna	XS, S, M, 1, 3.6
	7	31	rinna/bilingual-gpt-neox-4b	Rinna	4
	8	8	stockmark/gpt-neox-japanese-1.4b	Stockmark	1.4
		10	stabilityai/japanese-stablelm-base-alpha-7b	Stability AI Japan	7
		18	line-corporation/japanese-large-lm-3.6b-instruction-sft	Line	1.7, 3.6
		20	matsuo-lab/weblab-10b	東京大学	10
		29	elyza/ELYZA-japanese- Llama-2 -7b	ELYZA	7
	9	28	pfnet/plamo-13b	Preferred Networks	13
	10	20	llm-jp/llm-jp-13b-v1.0	LLM-jp	13
		25	stabilityai/japanese-stablelm-base-gamma-7b	Stability AI Japan	7
		27	stockmark/stockmark-13b	Stockmark	13
	11	2	cyberagent/calm2-7b	CyberAgent	7
		2	stabilityai/japanese-stablelm-base-beta-70b	Stability AI Japan	7, 70
	12	6	moneyforward/houou-instruction-7b-v1	Money Forward Lab	7
		19	tokyotech-llm/Swallow-70b-hf	東工大, 産総研	7, 13, 70
		21	rinna/nekomata-14b	rinna	7, 14
		27	elyza/ELYZA-japanese- Llama-2 -13b	ELYZA	13

年	月	日	モデル	開発元	規模
2024	1	15	lightblue/qarasu-14B-chat-plus-unleashed	Lightblue	7, 14
		31	karakuri-ai/karakuri-lm-70b-v0.1	KARAKURI	70
	2	9	llm-jp/llm-jp-13b-dpo-lora-hh_rhf_ja-v1.1	LLM-jp	13
		13	kotoba-tech/kotomamba-2.8B-v1.0	Kotoba Technologies	2.8
		26	sambanovasystems/SambaLingo-Japanese-Base	SambaNova Systems	7
	3	11	tokyotech-llm/Swallow-MS-7b-v0.1	東工大, 産総研	7
		11	tokyotech-llm/Swallow-MX-8x7b-NVE-v0.1	東工大, 産総研	46.7 (8x7)
		21	Rakuten/RakutenAI-7B	楽天	7
	4	20	abeja/Mixtral-8x7B-v0.1-japanese	ABEJA	46.7 (8x7)
		24	leia-llm/Leia-Swallow-13b	Studio Ousia	7, 13
		30	llm-jp/llm-jp-13b-v2.0	LLM-jp	13
	5	7	rinna/ Llama-3 -youko-8b	rinna	8
		10	Fugaku-LLM/Fugaku-LLM-13B	東工大, 東北大, 富士通, 理研, サイバーエージェント, Kotoba Technologies	13
		16	stockmark/stockmark-100b	Stockmark	100
		25	elyza/ Llama-3 -ELYZA-JP-70B	ELYZA	8, 70
	7	2	tokyotech-llm/ Llama-3 -Swallow-70B-v0.1	東工大, 産総研	8, 70

: フルスクラッチで学習したモデル

: 繙続事前学習したモデル (Llama2や3、Mistralなどの高精度な学習済みモデルに追加学習する場合が多い)

出所: [岡崎, 2024] JSAI2024 チュートリアル講演1「大規模言語モデルの開発」に加筆して修正

自然言語処理はどうなる?

- 人間が書いたり話したりする言葉をコンピュータで処理する技術・研究分野

- 応用技術: 自然言語処理を応用したアプリケーション

テキスト検索

テキスト分類

テキスト要約

情報抽出

機械翻訳

質問応答
対話

など

**ChatGPT で
いいんじゃない?**

- 基盤技術: 言語を応用タスクで利用しやすい形式に変換する

言語モデル

形態素解析

固有表現抽出

構文解析

意味解析

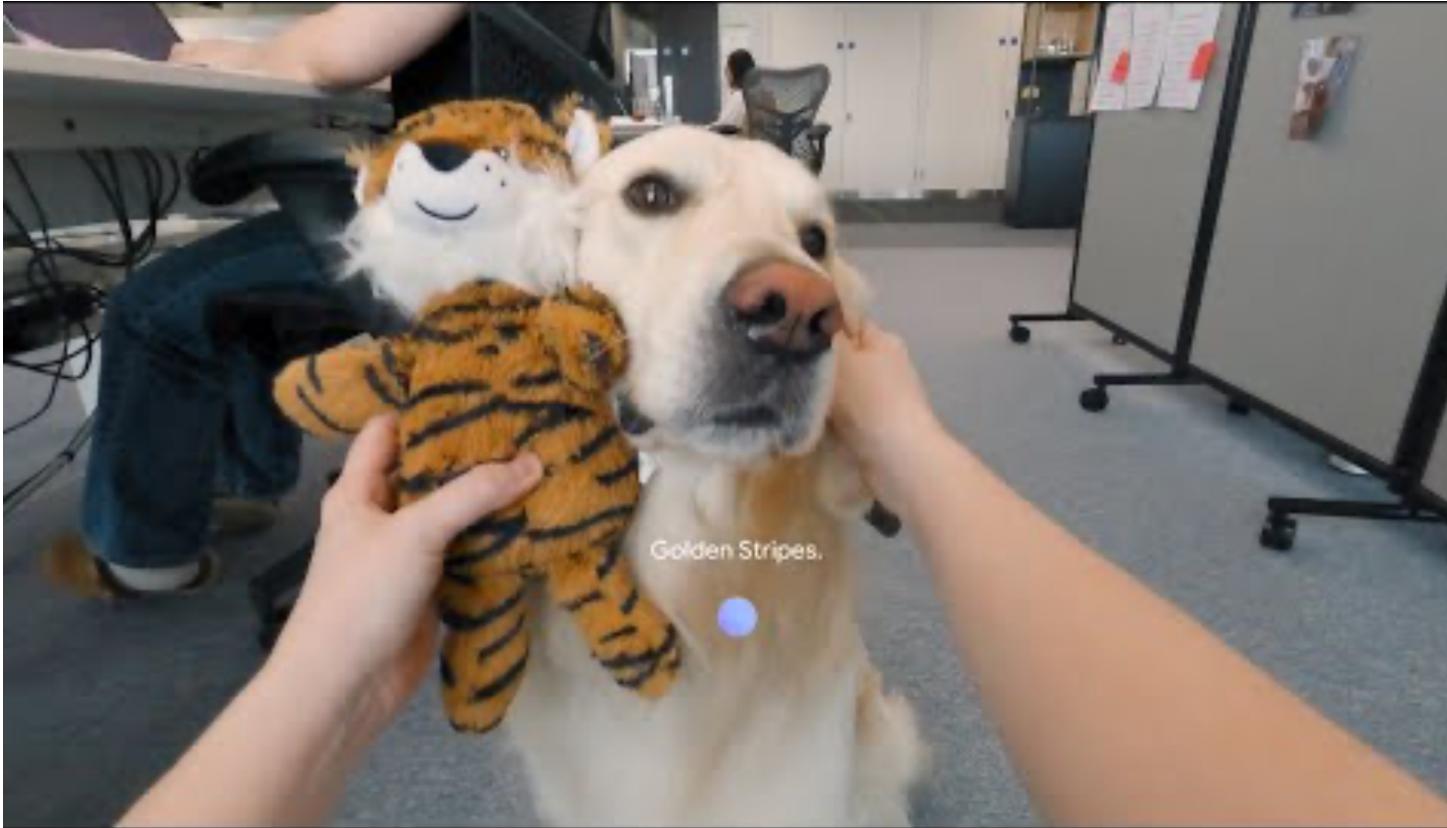
文脈解析

談話解析

など

今どこまで来ているか?

- スマホアプリから音声や映像を介したリアルタイムでのやり取りに対応する驚異的な体験
 - 5/14: OpenAI が GPT-4o 発表
 - 5/15: Google I/O で [Project Astra](#) 発表



左: Project Astra
(Google DeepMind)

大規模言語モデル(LLM)を利用するには

	フルスクラッチLLM	ローカルLLM	API 利用
説明	自らLLMをスクラッチで学習	オープンなLLMモデルを利用	クローズドLLMのサービスを利用する
メリット	<ul style="list-style-type: none">公開されている利用可能な基盤モデルではパラメータサイズが不十分・過剰特定の言語やドメインの知識のみのデータからなるモデルを作れる自社で保有する大量のデータを活用できる(例:コールセンターログ)出力結果の品質をコントロールできる	<ul style="list-style-type: none">オンプレミスで独自LLMを運用することで、情報漏洩リスクを低減可能ファインチューニングにより、プラットフォーム固有の機能、用語、コンテンツなど特定のニーズや要件に合わせてモデルを調整可能(ファインチューニング)出力結果の品質を多少はコントロールできる	<ul style="list-style-type: none">GPT-4o や Gemini などの高精度が得られるデータ管理・学習/運用環境の管理が不要
デメリット	<ul style="list-style-type: none">GPT-4o や Gemini のような高精度なモデルを構築することは難しいモデルサイズの大きさから、独自に大規模な学習/運用環境が必要(コスト高)	<ul style="list-style-type: none">GPT-4o や Gemini のような高精度なモデルを構築することは難しい独自に学習/運用環境が必要(フルスクラッチに比ると低成本)	<ul style="list-style-type: none">カスタマイズの余地が少ない一握りのLLMプロバイダーへの依存により、サービス停止の可能性がある
コスト例 (\$1=160円換算)	<ul style="list-style-type: none">LLaMA 65B のスクラッチ学習: A100 80GB × 2048枚 × 21日間=約6.5億円	<ul style="list-style-type: none">Alpaca 13B のファインチューニング: A100 80GB × 8枚 × 3時間=約20万円以下	<ul style="list-style-type: none">OpenAI のAPI料金(1Kトークン): GPT-4o 0.8円

代表的なサービス

- サービスによって、プライバシーやセキュリティ、準拠法など扱いが異なるため確認が必要

サービス名	提供者	プラットフォーム	モデル	備考
OpenAI API	OpenAI	Azure	LLM: GPT-4o , GPT-3.5 コード生成: GPT-4o , GPT-3.5 画像生成: DALL-E 音声認識: Whisper 音声合成: TTS	OpenAI社が発表した最新モデルが最も早く提供される
Azure OpenAI Service	Microsoft	Azure	• OpenAI API と同じモデル • OpenAI 以外のモデルも今後提供される予定	OpenAI がサービス提供した後、遅れてサービス提供開始
Vertex AI	Google	GCP	LLM: Gemini , PaLM コード生成: Codely 画像生成: Imagen	
Amazon Bedrock	Amazon	AWS	LLM: Claude , Amazon Titan 画像生成: Stable Diffusion	

(参考) AOAII と OpenAI の詳細比較 ※常に最新情報を確認するようにしてください

	Azure OpenAI Service	OpenAI
利用可能モデル	OpenAI GPT-4o, GPT-3.5, GPT-3, Codex, Embeddings, DALL-E	OpenAI GPT-4o, GPT-3.5, GPT-3, Codex, Embeddings, DALL-E, Whisper(音声認識)
サービス提供速度	OpenAI がサービス提供した後、遅れてサービス提供開始	最新モデルが先に提供される場合が多い
セキュリティ	<ul style="list-style-type: none"> • Azure のセキュリティ基準に準拠 • APIキーによる認証とAzure AD認証に対応 • Azure 仮想ネットワークによる保護 	<ul style="list-style-type: none"> • OpenAI のセキュリティポリシーに準拠 (脆弱性開示ポリシー) • APIキーによる認証
データ	<ul style="list-style-type: none"> • お客様から提供されたトレーニングデータは、お客様のモデルの fine-tuning (微調整) にのみ使用される • マイクロソフトのモデルをトレーニング/改善するために使用することはない (参考)。 • データは悪用/誤用の監視目的で30日間保持され、承認されたマイクロソフト社員がレビューする可能性がある (参考)。保持されないよう要求可能。 	<ul style="list-style-type: none"> • API経由のデータは OpenAI のモデルをトレーニング/改善するために使用することはない (参考) • データは悪用/誤用の監視目的で30日間保持され、OpenAI社員/サードパーティ業者がレビューする可能性がある。保持されないよう要求可能。
プライバシー	<ul style="list-style-type: none"> • マイクロソフトの声明およびAzure OpenAI Serviceの製品ポリシーに準拠 • 日本の法律を準拠法とし、東京地裁裁判所を管轄裁判所として契約可能 	<ul style="list-style-type: none"> • OpenAI のプライバシーポリシーに準拠。カリフォルニア州法に準拠
価格	モデル利用価格は同一 (Azure の価格体系に基づく)	モデル利用価格は同一 (OpenAI の価格体系に基づく)
SLA	<ul style="list-style-type: none"> • 99.9%以上の稼働率を保証 (詳細) 	SLAは提供されていない

出典: <https://zenn.dev/microsoft/articles/e0419765f7079a>

知っておくべき代表的なリスク

項目	概要	考えられる事例
Hallucinations	幻覚、でっち上げ。特定のソースとの関係で無意味な内容や真実でない内容を作り出す傾向がある	でっち上げの情報を元に、顧客への提案資料を作成
Harmful content	ポリシーに反するコンテンツや、個人・集団・社会に害を及ぼす可能性のあるコンテンツを生成する可能性がある	LLMを組み込んだサービスを提供している場合、顧客に対して、有害なコンテンツを出力
Disinformation and influence operations	ニュース記事、ツイート、対話、メールなど、もっともらしく現実的で的を射たコンテンツを生成することが可能である	社員が業務中にフェイクニュースを作成し、SNS投稿
Privacy	一般に利用可能なデータソースから学習しており、これは、個人情報を含む場合がある。その結果、個人の特定を試みるために使用される可能性がある	LLMを組み込んだサービスを提供している場合、顧客がLLMから訓練データ等の情報を抽出
Cybersecurity	ソーシャルエンジニアリングのいくつかのサブタスク(フィッシングメールの作成など)や、ソースコード内等の脆弱性の説明が可能	プログラム作成の際に、セキュアなコードへの変換を依頼し、プログラム内に埋め込まれた機微な情報(クレデンシャル等)が流出

引用: [中島,2023] AI-SCHOLAR主催「What is GPT」講演資料 (中島氏が [GPT-4 Technical Report](#) から抜粋して加筆修正)

參考資料 (1/2)

- [Vaswani+, NIPS2017] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
- [西田,2022] 西田京介. "自然言語処理とVision-and-Language." JSAI2022 チュートリアル講演資料 (2022).
- [Devlin+ (Google), NAACL19] Devlin, Jacob, Ming-Wei Chang, and Kenton Lee. "Google, KT, Language, AI: BERT: pre-training of deep bidirectional transformers for language understanding." Proceedings of NAACL-HLT. 2019.
- [Brown+ (OpenAI), NeurIPS2020] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
- [Chowdhery+, 2022] Chowdhery, Aakanksha, et al. "Palm: Scaling language modeling with pathways." arXiv preprint arXiv:2204.02311 (2022).
- [Wei(Google)+, 2022] Wei, Jason, et al. "Chain of thought prompting elicits reasoning in large language models." arXiv preprint arXiv:2201.11903 (2022).
- [Kojima, 2023] Kojima, Takeshi, et al. "Large language models are zero-shot reasoners."
- [Touvron+ (Meta), 2023/2/27] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).
- [Wei+, 2022] Wei, Jason, et al. "Finetuned language models are zero-shot learners." arXiv preprint arXiv:2109.01652 (2021).
- [Taori+ (Stanford Univ.), 2023/3/13] Taori, Rohan, et al. "Alpaca: A strong, replicable instruction-following model." Stanford Center for Research on Foundation Models. <https://crfm.stanford.edu/2023/03/13/alpaca.html> 3.6 (2023): 7.
- [The Vicuna Team, 2023/3/31] Chiang, Wei-Lin, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality." See <https://vicuna.lmsys.org> (accessed 14 April 2023) (2023).
- [Yao+'23, ICLR] Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." arXiv preprint arXiv:2210.03629 (2022).

参考資料 (2/2)

- [Honda+,2022] Shion Honda and Hidehisa Arai, "AI開発の新たなパラダイム「基盤モデル」とは." Recruit Data Blog (accessed 25 Jan 2023).
- [OpenAI,2023] OpenAI, R. "Introducing ChatGPT." <https://openai.com/blog/chatgpt> (accessed 5 April 2023).
- [OpenAI,2023] OpenAI, R. "GPT-4 technical report." arXiv (2023): 2303-08774.
- [Shohei N.,2023] Shohei N. "[比較表] Azure OpenAIと本家OpenAI APIの比較表." <https://zenn.dev/microsoft/articles/e0419765f7079a> (accessed 5 April 2023).
- [中島,2023] 中島佑允. "大規模言語モデルを活用するために知っておきたいこと." AI-SCHOLAR主催「What is GPT」講演資料 (2023).
- [Lambersy,2022] "How GPT-3 Writing Tools Work & 4 Things To Be Careful With When Using Them." <https://www.textcortex.com/post/how-gpt-3-writing-tools-work> (accessed 5 April 2023).
- [Harris+,1954] Harris, Zellig S. "Distributional structure." Word 10.2-3 (1954): 146-162.
- [Lin,2002] Lin, Jeng-Jong. "Applying a co-occurrence matrix to automatic inspection of weaving density for woven fabrics." Textile research journal 72.6 (2002): 486-490.
- [Mikolov+,2013] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
- [Mikolov+,2010] Mikolov, Tomas, et al. "Recurrent neural network based language model." Interspeech. Vol. 2. No. 3. 2010.
- [Sutskever+,NIPS2014] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems 27 (2014).
- [Bahdanau+,2015] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

大規模言語モデル (LLM) の成り立ち

(再掲) 自然言語処理とは

- 人間が書いたり話したりする言葉をコンピュータで処理する技術・研究分野
- 応用技術：自然言語処理を応用したアプリケーション

テキスト検索 テキスト分類 テキスト要約 情報抽出
機械翻訳 質問応答 対話 など

- 基盤技術：言語を応用タスクで利用しやすい形式に変換する

言語モデル 形態素解析 固有表現抽出 構文解析
意味解析 文脈解析 談話解析 など

言語モデル (Language Model: LM) とは

- 単語の並びの生成確率をモデル化したもの (確率的言語モデル)

単語の並び $\Rightarrow y_1, y_2, \dots, y_T$ 生成確率 $\Rightarrow P(y_1, y_2, \dots, y_T)$

- 特定の単語の次に来る単語を予測できる

$$y^* = \operatorname{argmax}_{y \in V} P(\text{日本}, \text{の}, \text{首都}, \text{は}, y)$$

$P(\text{日本}, \text{の}, \text{首都}, \text{は}, \text{ロンドン})$	= 0.00000043
$P(\text{日本}, \text{の}, \text{首都}, \text{は}, \text{パリ})$	= 0.00000082
$P(\text{日本}, \text{の}, \text{首都}, \text{は}, \text{東京})$	= 0.00000103
$P(\text{日本}, \text{の}, \text{首都}, \text{は}, \dots)$	= ...

計算された確率が最大値を
取る単語を選択する
↓
東京

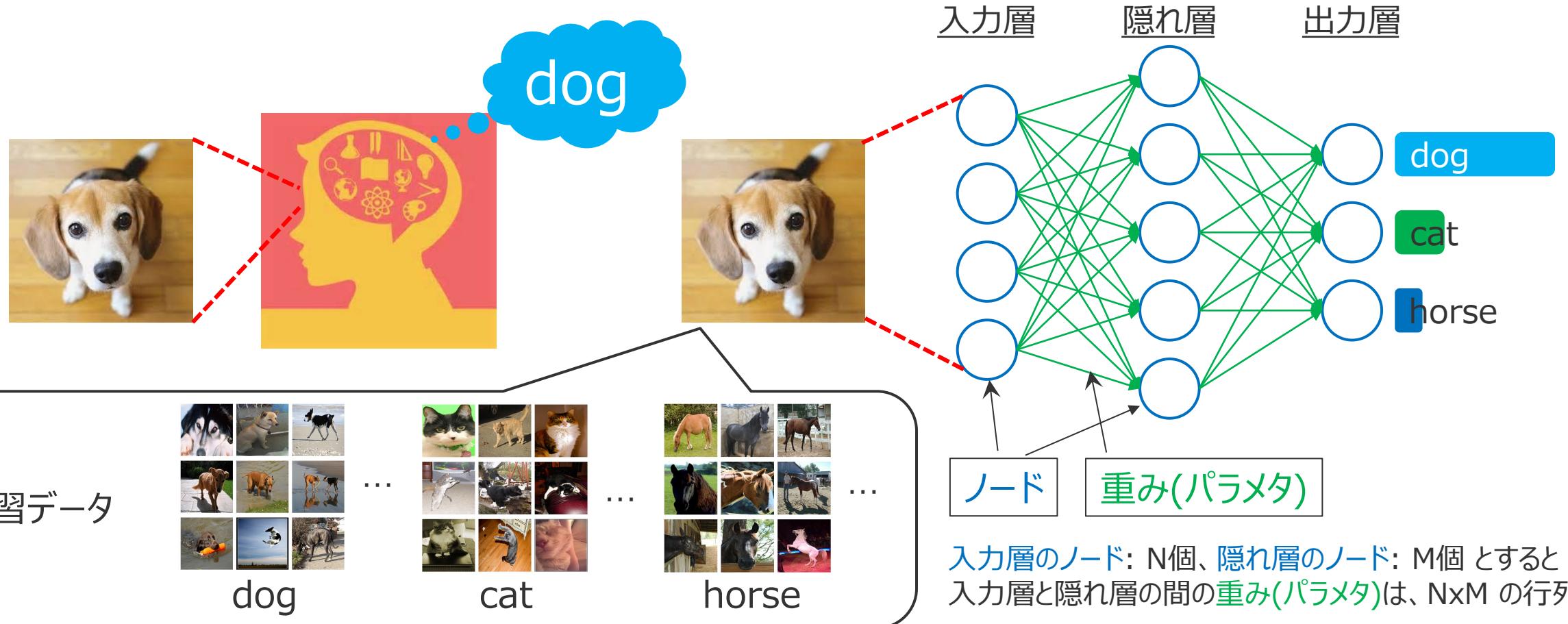
- 単語の並びに従ってモデル化することで文脈も考慮 \rightarrow N-gram言語モデル

$$y^* = \operatorname{argmax}_{y \in V} P(\text{日本}, \text{の}, \text{首都}, \text{は}, y) = \operatorname{argmax}_{y \in V} P(y | \text{日本}, \text{の}, \text{首都}, \text{は})$$

yの前方N-1個の単語(トークン)

ニューラルネットワークとは

- 脳の神経細胞(ニューロン)の働きを模した機械学習の手法
- ちなみに、機械学習とはデータを学習しパラメータを獲得すること



- 分布仮説 [Harris+, 1954]

- 単語の意味はその周囲の単語から形成されるという仮説
→ 似た文脈で出現する単語は意味が似ている

文1: 昨日、りんごを食べた。りんごジュースを飲んだ。りんごの皮をむいた。

文2: 昨日、りんごを食べた。ぶどうジュースを買った。ぶどうの皮をむいた。

文3: 昨日、自転車に乗った。自転車を修理した。自転車を買った。

共起による
ベクトル表現
[Lin, 2002]

	…	食べる	…	飲む	…	修理	…
りんご	0	1	0	1	0	0	0
ぶどう	0	1	0	1	0	0	0
自転車	0	0	0	0	0	1	0
:							

← 語彙数(数万～数十万)分の疎なベクトルになる →

似てる
似てない

- 分布仮説 [Harris+, 1954]

- 単語の意味はその周囲の単語から形成されるという仮説

→ 似た文脈で出現する単語は意味が似ている

- 各意味を複数の次元で分散して表現する (=分散表現)

→ 次元は低次元(例えば100次元)で、値は実数値

单語埋め込み
(word embedding)
とも呼ばれる

これらの実数値をニューラルネットワークで求める

共起による
ベクトル表現
[Lin, 2002]

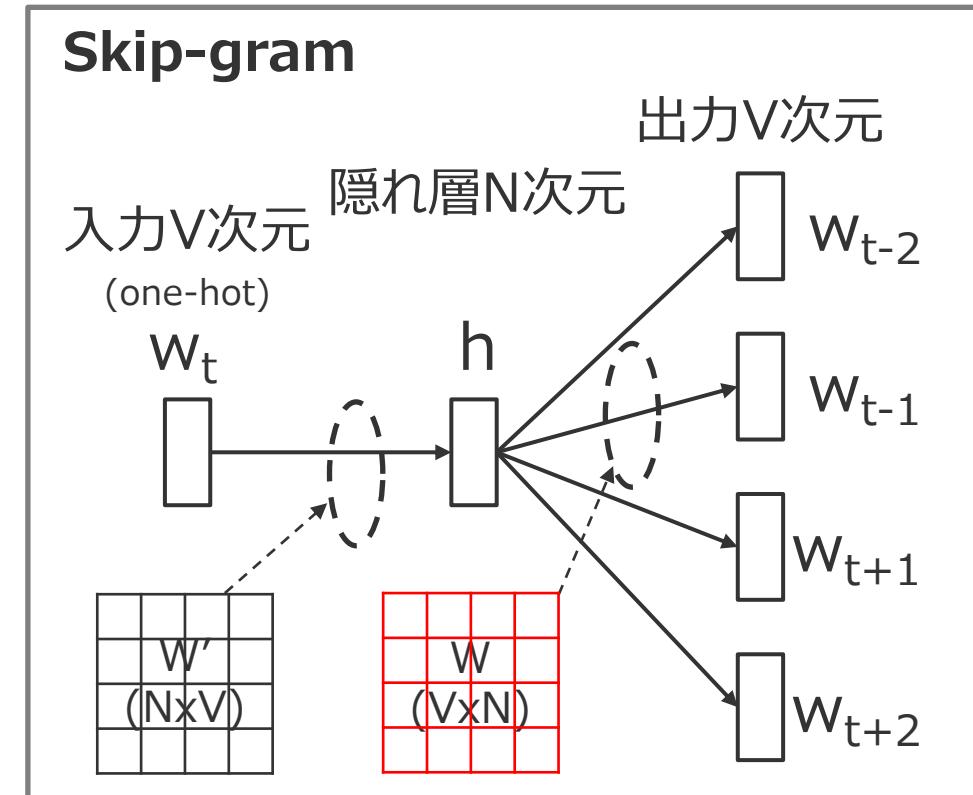
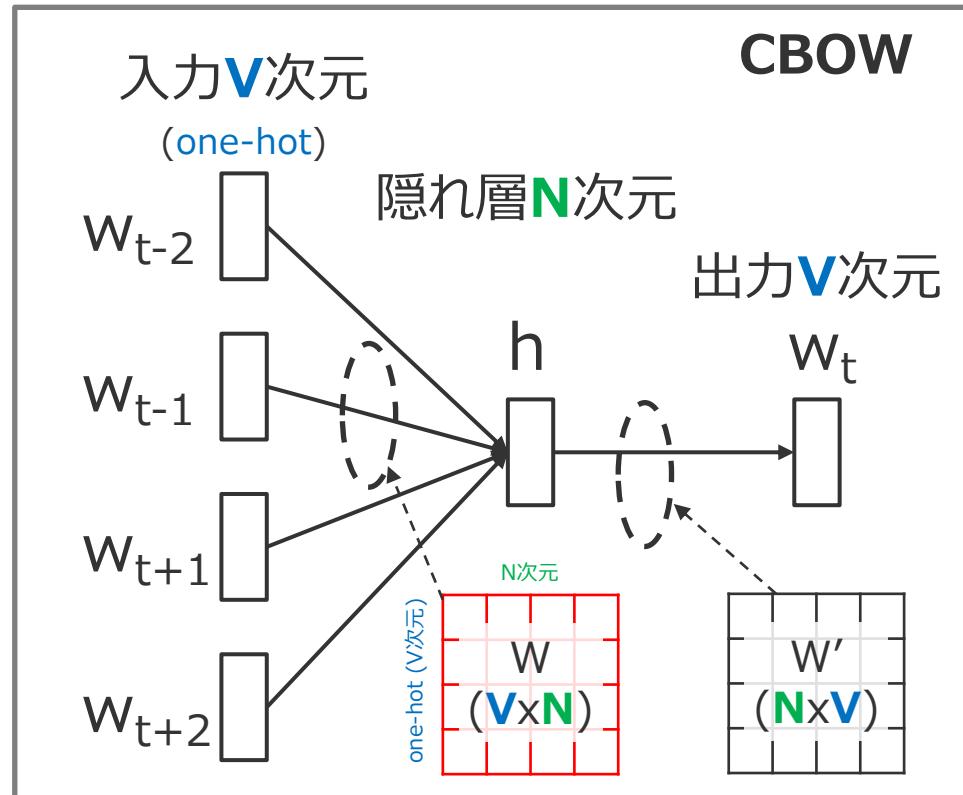
次元→	0	1	…	50	…	98	99
りんご	1.07	-1.08		1.48		0.46	0.48
ぶどう	1.95	-1.53		0.36		-0.61	-0.44
自転車	0.67	1.44		-1.50		0.10	0.67
:							

← 高々数百次元の密なベクトル →

似てる
似てない

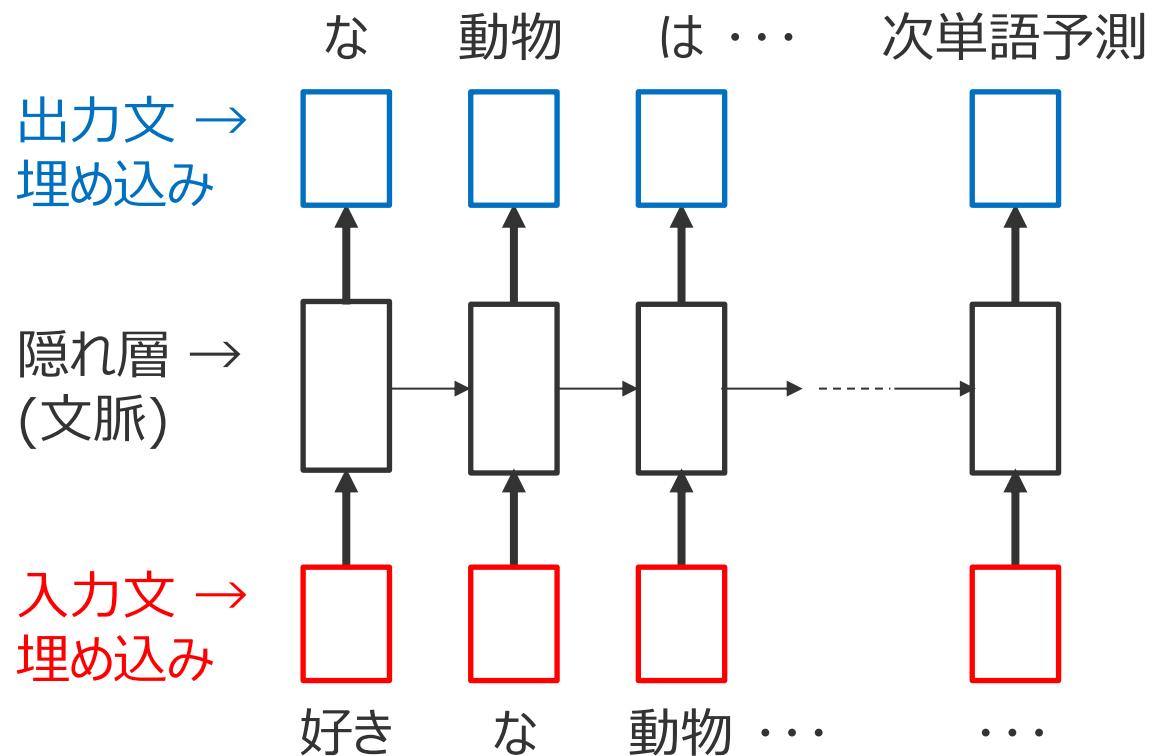
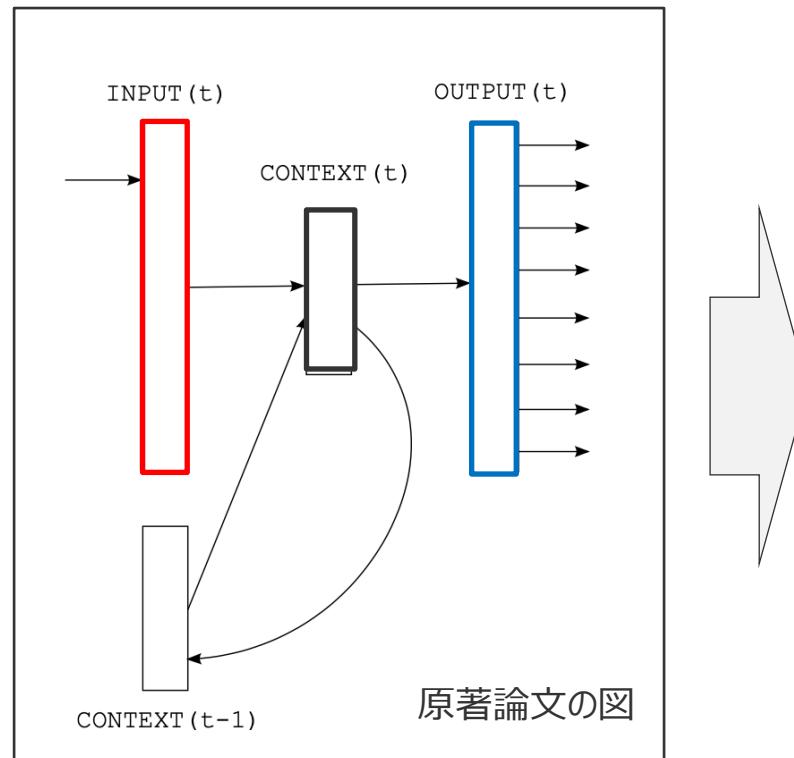
- 代表格: word2vec [[Mikolov+, 2013](#)]

- 従来の単語ベクトルも類似度の比較はできていたが、足したり引いたりできなかった
 - $\text{king} - \text{man} + \text{woman} = \text{queen}$ が有名



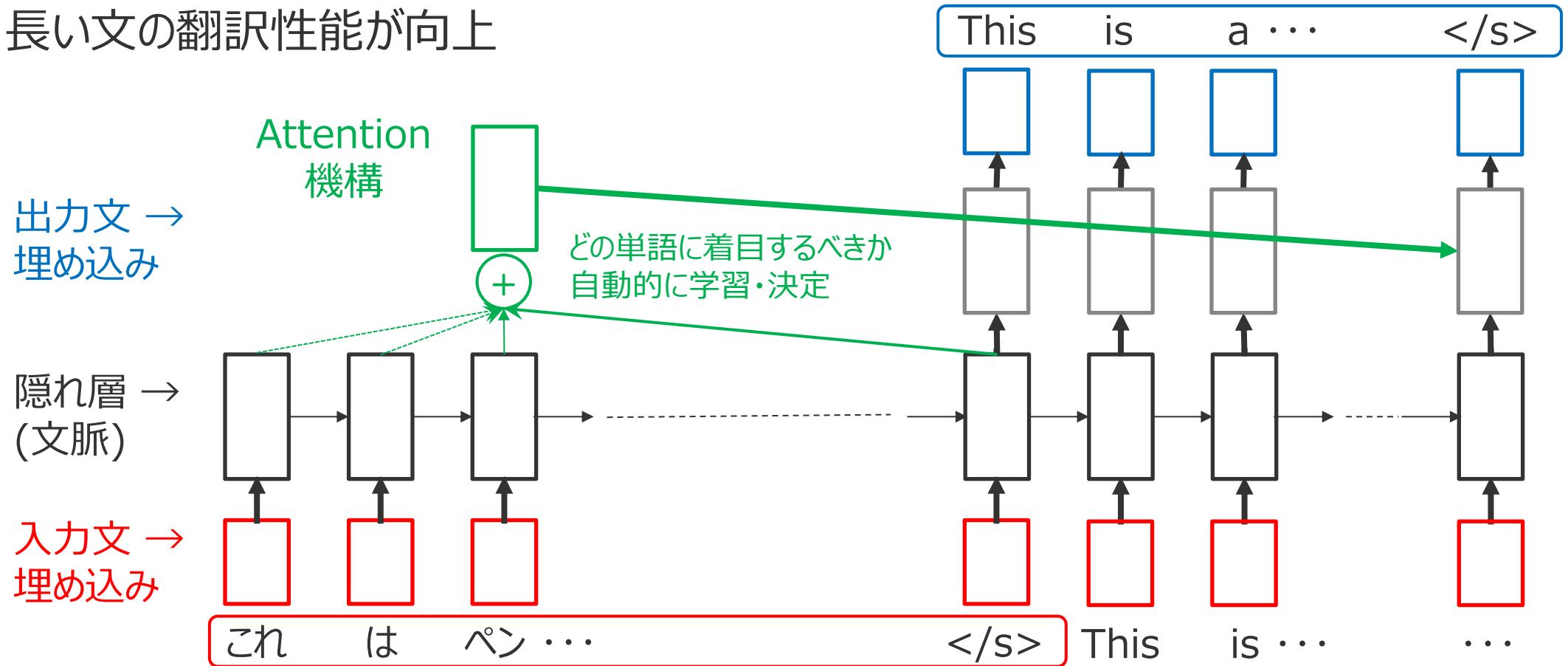
- RNN 言語モデル [Mikolov+, 2010]

- RNN(系列データを対象とするNN)を使った言語モデルで、次の単語を予測する
- 隠れ層に過去の履歴(文脈を考慮した情報)が埋め込んでいく



エンコーダー-デコーダー型の機械翻訳モデル

- Seq2Seq [[Sutskever+, NIPS2014](#)]: ニューラル機械翻訳の基本となったモデル
- Attention 機構 [[Bahdanau+, 2015](#)] により、
長い文の翻訳性能が向上

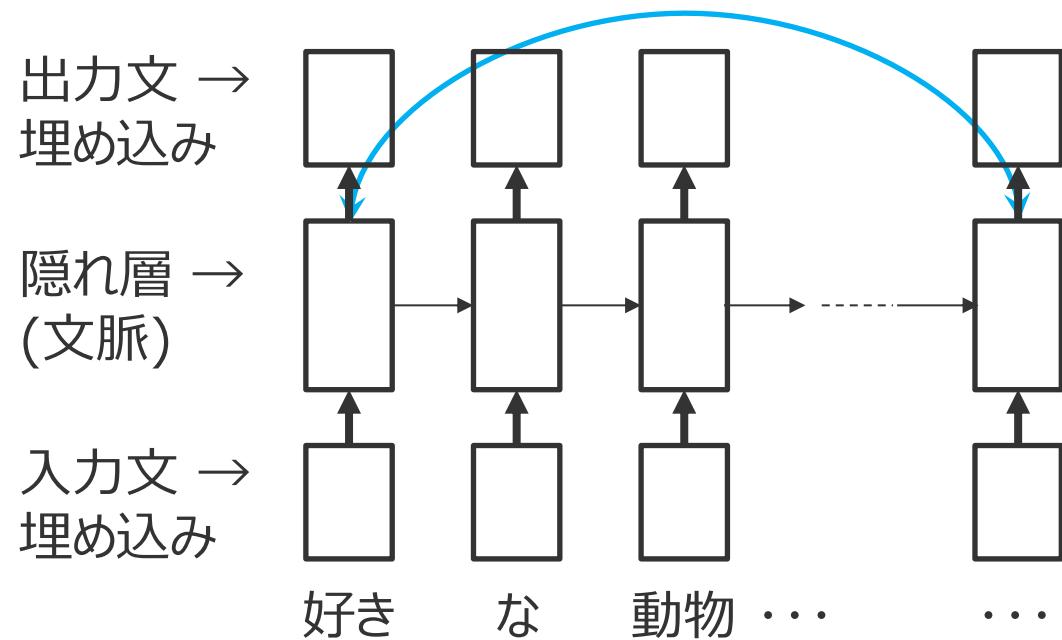


Self-Attention の登場

- 従来の文脈理解は、長期依存性の理解に限界があった
- 離れた単語の関係性も直接考慮できる Self-Attention が性能向上に大きく寄与

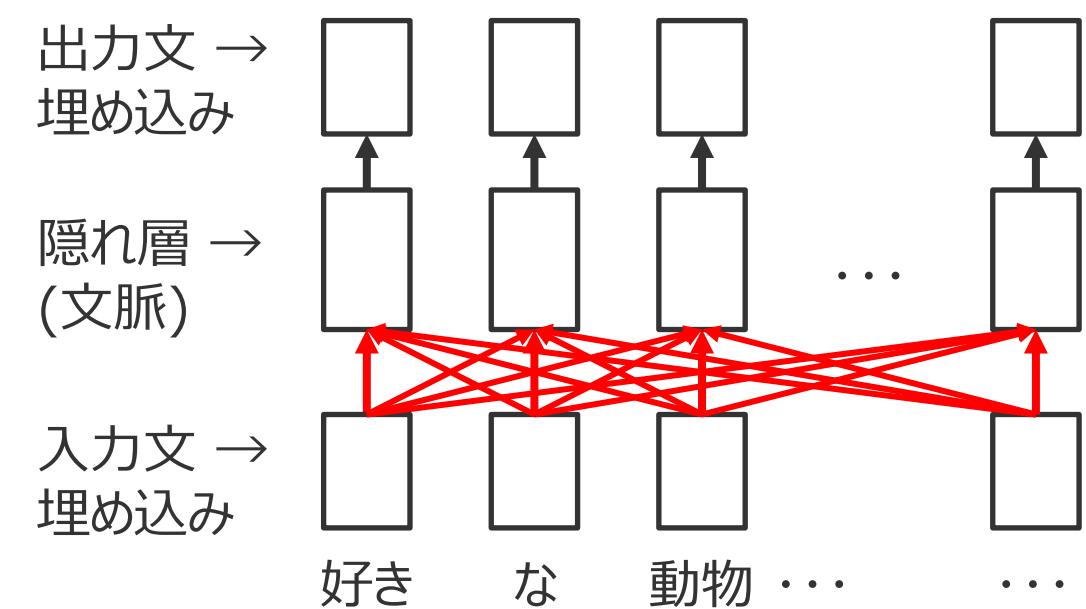
従来(LSTM)

遠く離れた単語の関係性
を捕まえにくい



Self-Attention

遠く離れた単語も直接
関係性を考慮できる



Transformer [Vaswani+, NIPS2017]

- 単語間の関係を RNN や CNN を用いずアテンションのみを用いて表現したエンコーダデコーダ型モデルにより、機械翻訳で圧倒的な SOTA を達成

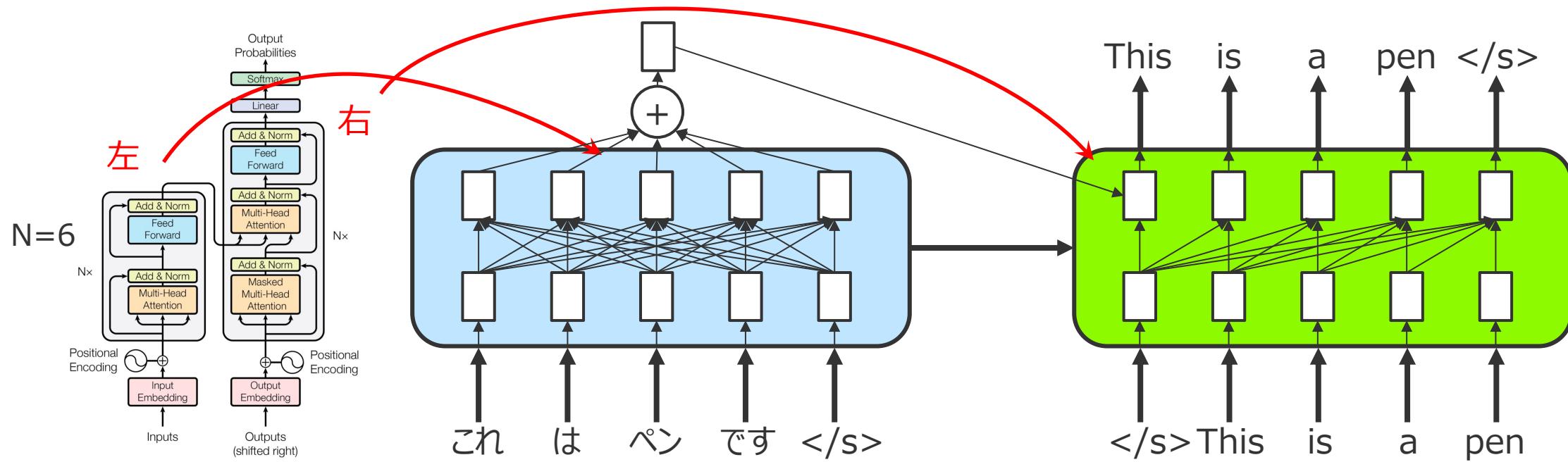


Figure 1: The Transformer - model architecture.

よく見る図

Transformer エンコーダ

引用: [西田,2022] JSAI2022 チュートリアル講演資料の一部を修正して作成

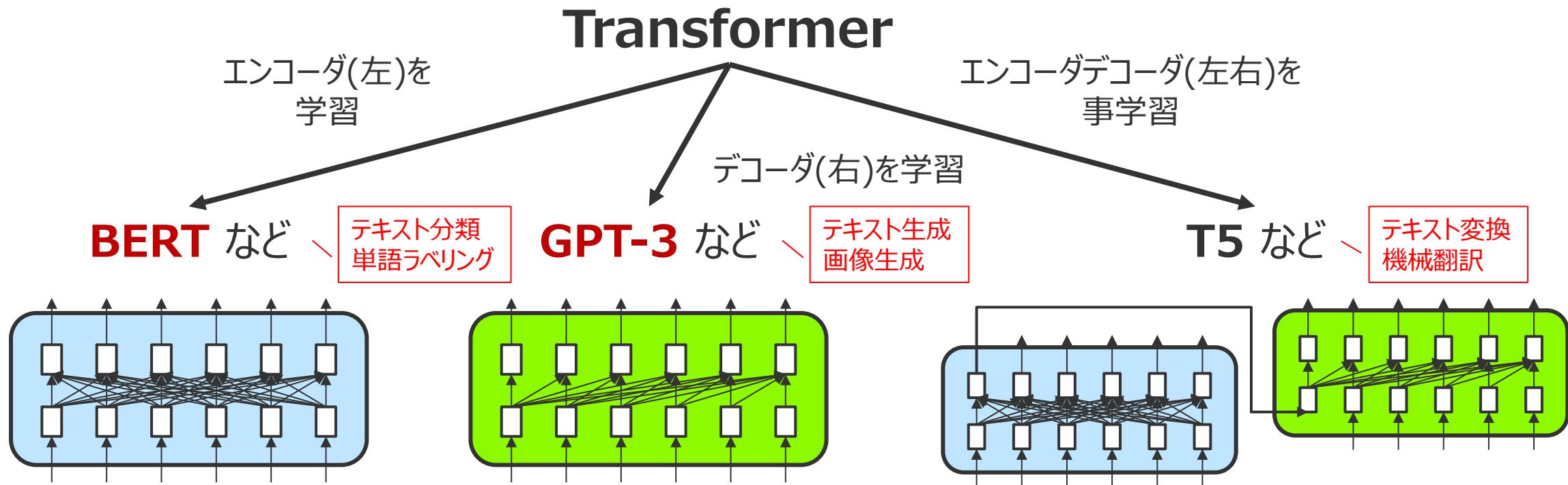
双向型のNNモデル
(右側の単語も使う)

Transformer デコーダ

自己回帰型のNNモデル
(出力を入力に戻す)

Transformer [Vaswani+, NIPS2017]

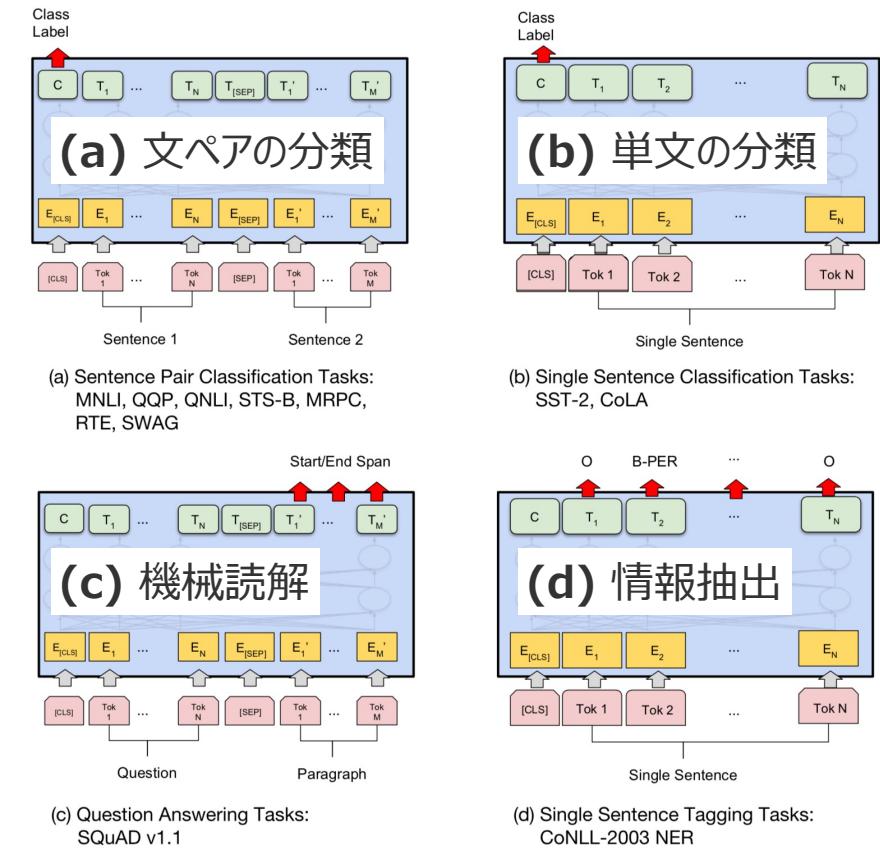
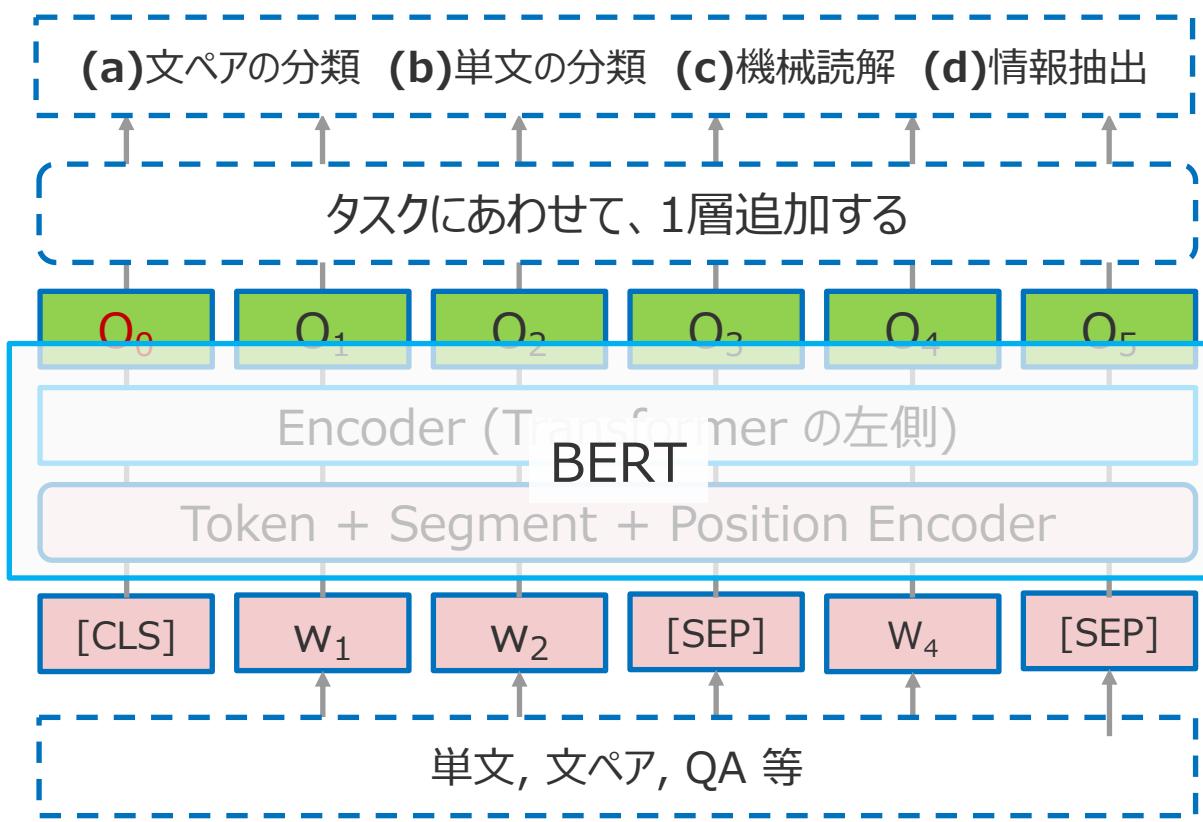
- 近年の基盤モデルの殆どがモデルの一部に Transformer を採用
- コンピュータビジョンの分野にも Transformer が高い性能を發揮



引用: [西田,2022] JSAI2022 チュートリアル講演資料の一部を修正して作成

BERT [Devlin+ (Google), NAACL19]

- 機械読解タスクで人間のスコアを超える、多数のNLPタスクで高性能を出し注目を浴びた
- 双方向 Transformer ブロックを24層重ねた言語モデル
- 出力層をタスク毎に1層のみ追加して、様々なタスクに適応できる



GPT-3 [Brown+ (OpenAI), NeurIPS2020]

- GPT-1(1億パラメタ), GPT-2(15億パラメタ)と同じ自己回帰モデルだが超大規模
- 超大量(3000億トークン)のテキスト, 超巨大(96層)の Transformer デコーダで 1750億※のパラメタを学習 (例: BERTは, 3.3億トークン, 24層, 3.4億パラメタ)
- タスクの説明もテキストとして入力し, 様々なタスク(マルチタスク)を実現
 - Zero-shot: タスク説明のみを与え全くサンプルを与えない
 - One-shot: タスク説明と1つのサンプルのみを与える
 - Few-shot: タスク説明と少數(10から100)のサンプルを与える



※ 2022年4月に Google が公開した PaLM は 5400億 [Chowdhery+, 2022]

GPT-3 [Brown+ (OpenAI), NeurIPS2020]

The three settings we explore for in-context learning

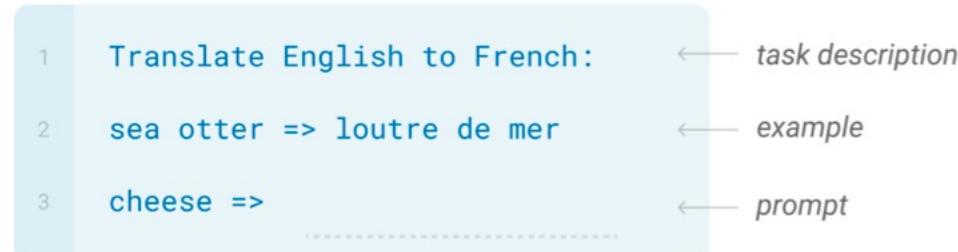
Zero-shot

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



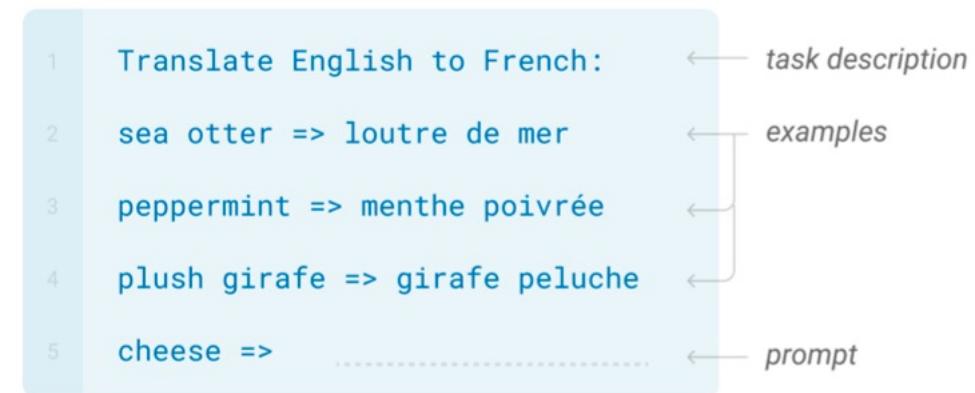
One-shot

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



Few-shot

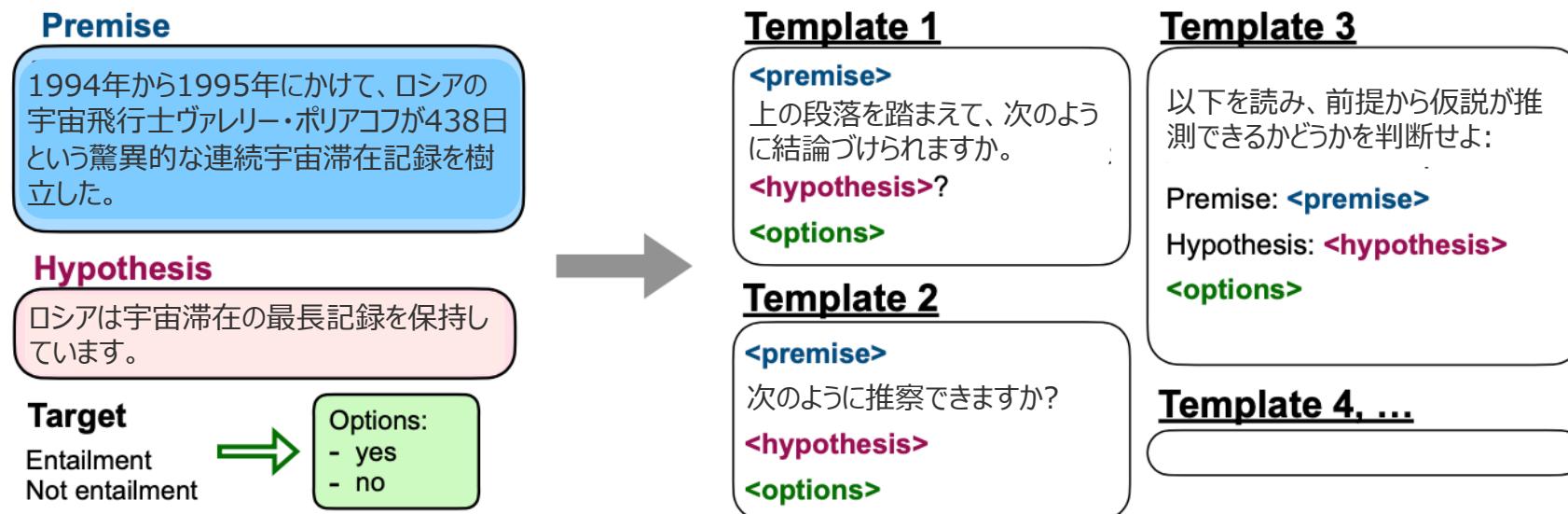
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



In-context learning:
タスクをその場で学習する能力

(再掲) Instruction Tuning (FLAN) [Wei+,2022]

- GPT-3 など言語モデルの構造を変えずに、**複数のタスク**でファインチューニングする方法
- タスク毎にテンプレートを用意し、「プロンプト(タスクの指示と事例)+出力」の形式に変換した学習データで言語モデルを追加学習する (=Instruction Tuning)
 - ゼロショットで解くタスクにおいて、GPT-3よりも高い精度を達成
 - 未知のタスクや指示に対しても精度よくテキストを生成



図：自然言語推論(NLI)タスクの例（前提出文が仮設を含意するか否かを自動判定するタスク）

引用: <https://arxiv.org/pdf/2109.01652.pdf>

ChatGPT — 大規模基盤モデル GPT-3 をファインチューニングしたモデル

事前学習した大規模言語モデル GPT-3 に、人間の質問に答えるように「**Instruction Tuning**」と人間の好みに合った答えを出すように(=アライメントするように)「**RLHF**」を加えた

Step1

人間の用意した望ましい回答で事前学習モデル(GPT-3)を fine-tuning

A prompt is sampled from our prompt dataset.

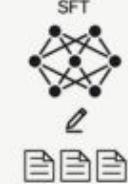
Explain reinforcement learning to a 6 year old.



We give treats and punishments to teach...

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



引用: <https://openai.com/blog/chatgpt/>

Step2

Step1の出力に人間がランク付けし、報酬モデルを学習

A prompt and several model outputs are sampled.

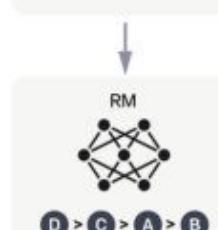
Explain reinforcement learning to a 6 year old.

- A In reinforcement learning, the agent is...
- B Explain rewards...
- C In machine learning...
- D We give treats and punishments to teach...



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step3

Step2で学習した報酬モデルを使ってPPOで強化学習

A new prompt is sampled from the dataset.



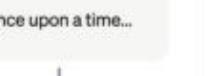
Write a story about otters.



The PPO model is initialized from the supervised policy.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Instruction Tuning:

人間の指示に対して、望ましい出力をするように学習する方法 (→前頁)

RLHF: 人間のフィードバックに基づいて強化学習する方法

- モデルの出力に対して人間がスコアを付与
- スコアを最大化するように報酬モデルを学習
- 学習した報酬モデルを用いて、人間が好む出力をするように強化学習する

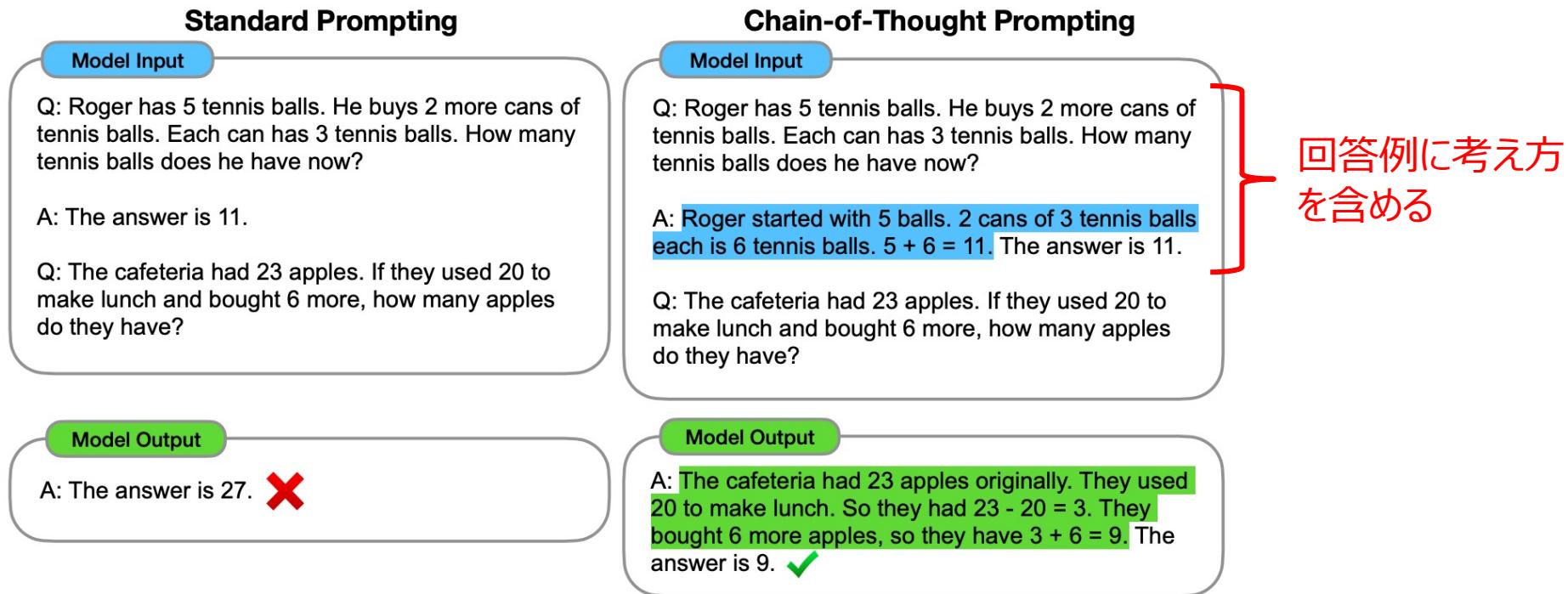
プロンプトエンジニアリングの手法

- 様々な手法やTipsが溢れている → 論文などで効果が検証された手法を使うのが良い

Few-shot Learning	少数の入力データ例を与えることで、モデルにタスクを学習させる手法。
Chain of Thought (CoT)	モデルに推論プロセスを段階的に説明することで、複雑な推論問題に対する精度を向上させる手法。
Self-consistency	同じ問題に対して複数の異なる回答を生成し、それらの中で最も妥当な回答を選択することで、信頼性を向上させる手法。
Recursively Criticizes and Improves (RCI)	モデルが自身の生成した回答を批判し、それを基に改善を繰り返すことで、より洗練された回答を生成する手法。
ReAct (Reason + Act)	推論タスクと行動タスクを組み合わせ、モデルが環境と相互作用しながらタスクを遂行する手法。

Chain-of-Thought

- 「考え方」を与えることで、推論能力が大きく向上する [[Wei\(Google\)+,2022](#)]
- ReAct や Self Consistency も CoT の考え方を継承している

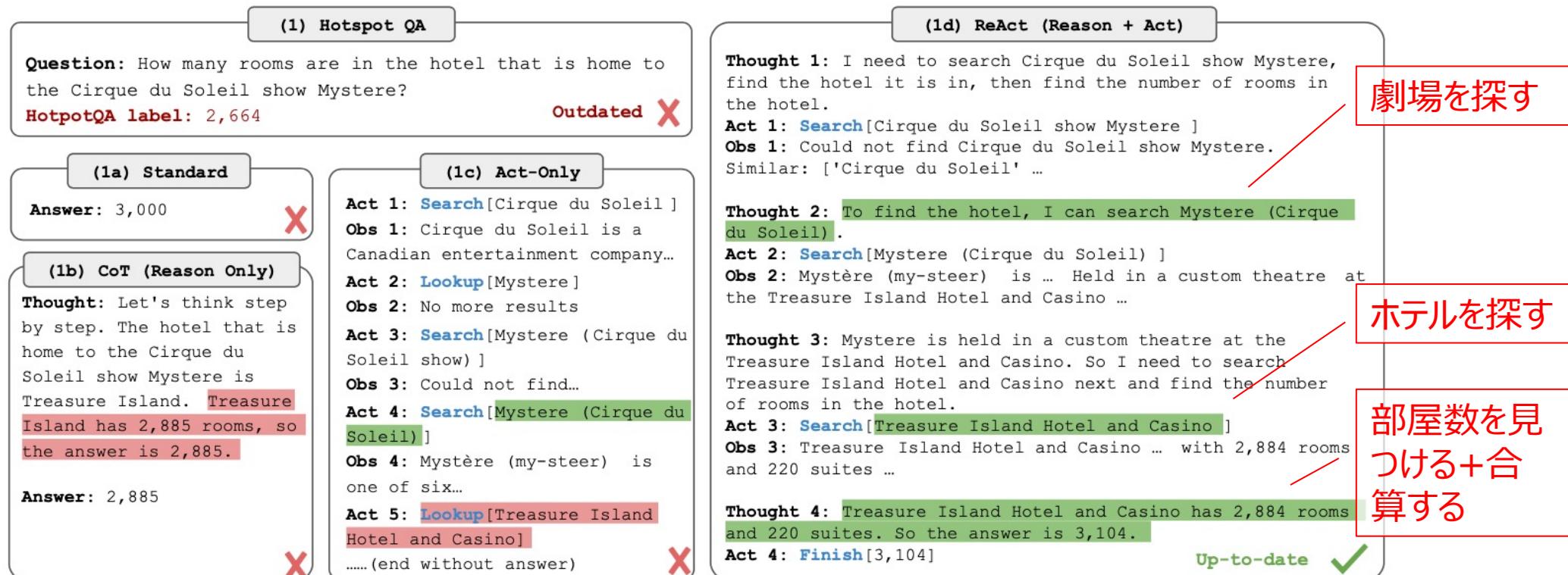


- 「ステップバイステップで」をプロンプトに追加すると計算などが正確になりやすい [[Kojima+,2022](#)]

引用 <https://arxiv.org/pdf/2201.11903.pdf>

ReAct (Reason + Act)

- LLM による推論の過程に思考と行動のフェーズを導入することで、回答精度を向上
- 必要な行動(Act)とその理由(Reason)を思考する
- 思考をもとに行動し、得られた結果から再度次に必要な行動とその理由を思考する



引用: <https://arxiv.org/abs/2210.03629>

ReAct の動作イメージ

Input: レオ・ディカプリオの恋人は誰でしょう？彼女の現在の年齢は0.43乗すると何歳でしょう？

prompt:

次の質問に、できる限り答えてください。あなたは以下のツールにアクセスすることができます：

Search: 検索エンジン。時事問題に関する質問に答える必要があるときに便利です。入力は検索クエリである必要があります。

Calculator: 数学に関する質問に答える必要があるときに便利です。

次のような書式で入力します：

質問: 答えなければならない入力された問題

思考: 常に何をすべきかを考えるべき

行動: 取るべき行動 [Search, Calculator] のいずれかでなければならない

行動の入力: アクションへの入力

観察: 行動の結果

... (この思考・行動・行動の入力・観察は、N回繰り返すことができます)

思考: 最終的な答えがわかった

最終回答: 入力された元の質問に対する最終的な回答

始める！

質問: レオ・ディカプリオの恋人は誰でしょう？彼女の現在の年齢は0.43乗すると何歳でしょう？

LLMへの入力

レオ・ディカプリオのガールフレンドが誰なのか、彼女の現在の年齢を調べたい。
そして、彼女の年齢を0.43乗で計算してみる。

行動: Search

行動の入力: レオ・ディカプリオの恋人とその年齢

Searchの
実行

観察: レオナルド・ディカプリオは、恋人のカミラ・モローネと数ヶ月で別れた後、彼の恋愛に関する長年の説を証明したようだ ...

思考:

LLMへの入力

レオ・ディカプリオが最近、恋人のカミラ・モローネと別れたことがわかった。0.43乗の年齢を計算する前に、彼女の年齢を知る必要があります。

行動: Search

行動の入力: カミラ・モローネ年齢

Searchの
実行

観察: 25 years

思考:

LLMへの入力

Camila Morroneは25歳です。これで、彼女の年齢を0.43乗で計算できる。

行動: Calculator

行動の入力: 25 ** 0.43

Calculatorの
実行

観察: 回答: 3.991298452658078

思考:

LLMへの入力

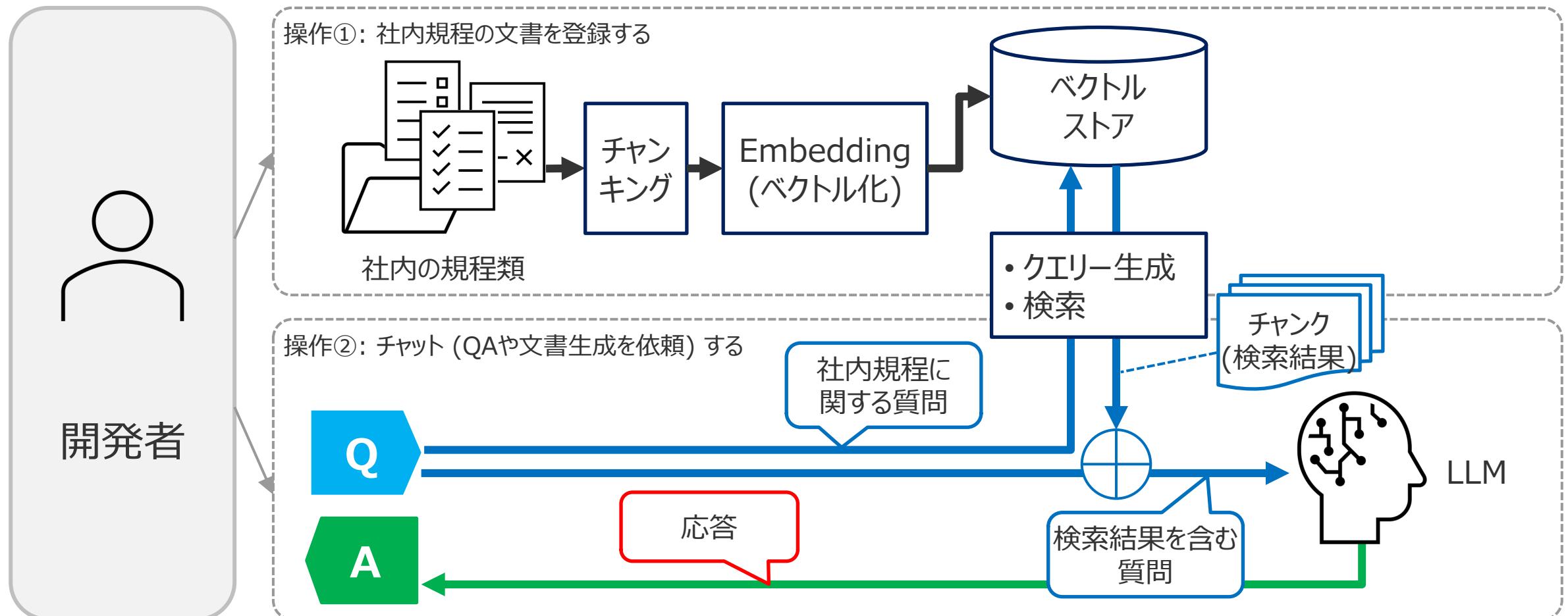
カミラ・モローネの年齢を0.43乗で計算すると、約3.99になります。

最終回答: カミラ・モローネの年齢を0.43乗すると約3.99になります。

テキストマイニングの未来図

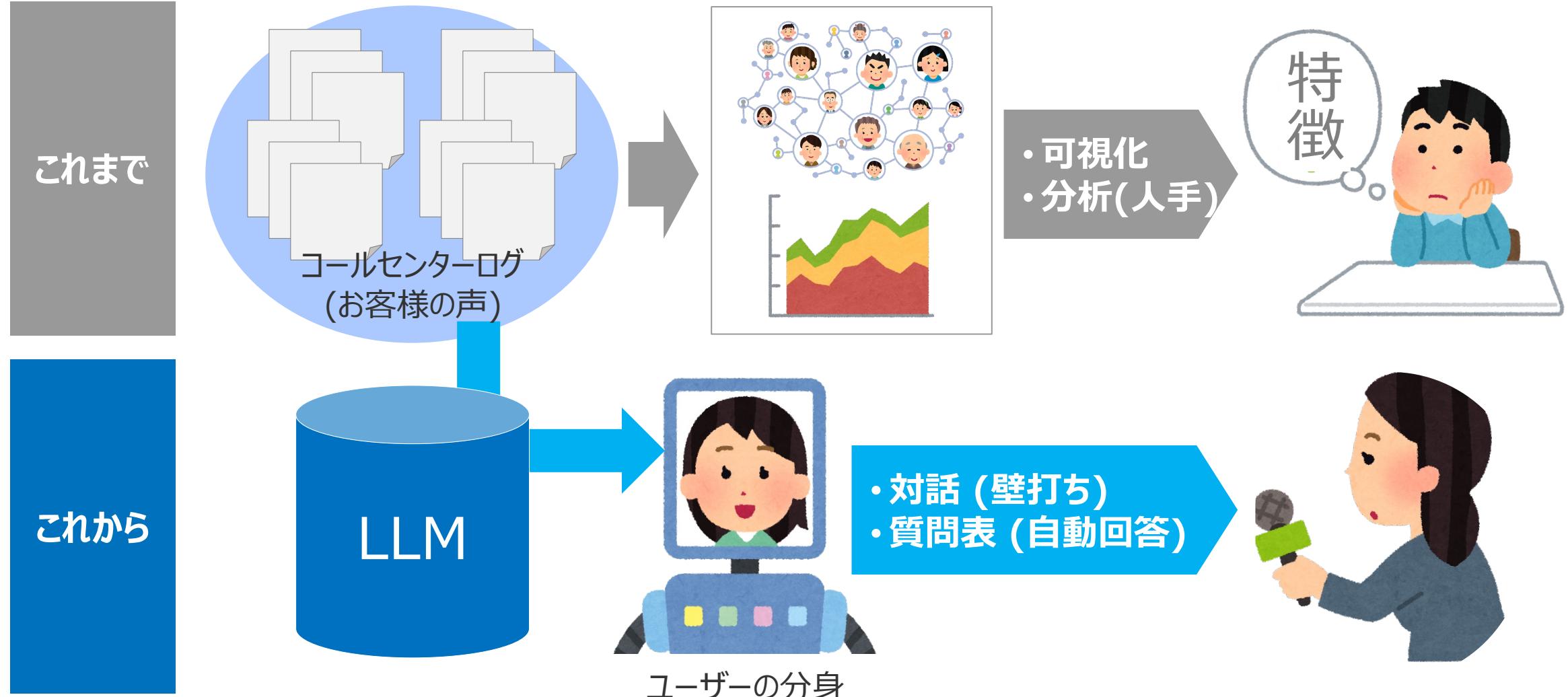
RAG (Retrieval-Augmented Generation)

- 外部の情報や事実を検索して、検索で得られた正確な情報に基づいて、大規模言語モデル(LLM)に回答を生成させる仕組み（社内規程やマニュアルQA）



テキストマイニングの未来図

お客様の声の分析の例 … LLM がユーザー(あるいはユーザーの集合)になりきり(=分身)、ユーザーの分身である LLM との対話やヒアリングを通して、ユーザー(あるいはユーザーの集合)の潜在的なニーズや隠された課題を探る



どこまで来たか?

● ChatGPT による EXCEL データの分析例 ※日本語の解析ができないなど現時点では課題も多い

① このデータは宿泊者によるクチコミです。エリアはいくつありますか?

ChatGPT の回答は必ずしも正しいとは限りません。重要な情報は確認するようにしてください。

	カテゴリー	エリア	施設番号	施設名
1	A_レジャー	01_登別	80732	登別カルルン 湯元オロフレ
2	A_レジャー	01_登別	149334	天然温泉 モ ドーミーイン 蘭 (ドーミー 御宿野乃ア グループ)

このデータは宿泊者によるクチコミです。エリアはいくつありますか?

このデータには「エリア」列があります。エリアの数を数えます。 [→]

このデータには、10の異なるエリアがあります。 [→]

□ □ □ □ □

エリアごとのクチコミ数を教えてください。 最高評価の施設を教えてください。

② ChatGPT にメッセージを送信する

ChatGPT の回答は必ずしも正しいとは限りません。重要な情報は確認するようにしてください。

① 湯布院と登別のコメントを読んでエリアごとの評判を比較してください。

ChatGPT の回答は必ずしも正しいとは限りません。重要な情報は確認するようにしてください。

湯布院のクチコミ

- 食事とお風呂の評価が高い: 多くのコメントで食事の質や量が高評価されており、部屋に設置された風呂（特に半露天風呂）も好評です。ただし、設備の一部（布団の厚さなど）については改善の余地があるようです。
- 観光地へのアクセスが良好: 位置が観光地に近いため便利という意見が多いです。そのため、訪問者は観光後に簡単に宿に戻れる利便性を楽しんでいます。
- 個別のサービスに対する感謝: 貸し切り風呂の利用が可能で、清掃が行き届いている点も評価されています。また、スタッフの親切な対応が印象的な点として挙げられています。

登別のクチコミ

- お風呂の設備が評価の中心: 登別のコメントでは、サウナや露天風呂などのお風呂の設備が特に良かったとの意見が多いです。
- 家族向けのサービス: 子供向けのサービスが良いと評価されており、家族連れにとって魅力的な宿泊地であることが伺えます。
- 清掃面での改善点: いくつかのコメントで清掃が十分でない点が指摘されており、特に大浴場入口の清掃状態や部屋の掃除が不十分だったとの声があります。また、寝具の質に関する不満も見られます。

② ChatGPT にメッセージを送信する

ChatGPT の回答は必ずしも正しいとは限りません。重要な情報は確認するようにしてください。

まとめ

- 大規模言語モデル(LLM)はどんなもの?
 - 生成AIは、我々の生活やビジネスに大きな影響を与え、利用範囲の拡大が進む
 - クローズドなLLM の開発と同様、ローカルLLM の開発競争も激しく、その差が縮まる
 - LLM によって、今後多くの新しいサービスや応用が期待される
- 大規模言語モデルの成り立ち
 - 従来の言語モデルは文章を生成(次の単語を予測)していた
 - LLMは Transformerで文章生成する仕組みを超大規模化 → 知識と読解力を得た
- テキストマイニングの未来予想図
 - RAG によって、より手軽に誰でもテキストマイニングを実現できるようになる(予想)

テキストマイニングの手順

テキストマイニング

- 驚異的な大量の文書データに記述されている多種多様な内容を対象として、その相関関係や出現傾向などから新たな知識を発見する [那須川,1999]
- 市場調査や販売戦略の立案、製品やサービス改善、顧客対応の改善に役立てたい
 - アンケート、レビューサイトのクチコミ、ツイートなど
- 最近では、報道番組などで Twitter 分析を取り上げることも多い
 - 震災、選挙、新型コロナウィルスなど

クチコミ分析の例 — コックroach

- パッケージ描かれたイラストが嫌 → 変更後、前年比2倍の出荷



出典: http://www.kincho.co.jp/seihin/insecticide/go_aerosol/gokiburi_u_spray/index.html

クチコミサイトの例 — 楽天トラベル

● ホテルのクチコミ数: 1,393万件 ※年間約60~80万

The screenshot shows the Rakuten Travel website at <https://travel.rakuten.co.jp/review/>. The main heading is 'お客様の声' (Customer Reviews) with the number '13,929,531'. Below it is a search bar for reviews and filters for domestic and overseas stays. A green banner at the bottom left lists new reviews from CANDEO HOTELS and BIO-RESORT HOTEL & SPA OSE. On the right, a box highlights customer feedback and service icons for domestic stays, high-speed buses, pet hotels, and overseas stays.

経年変化:

- 780万件 (2015)
 - 836万件 (2016)
 - 900万件 (2017)
 - 973万件 (2018)
 - 1,042万件 (2019)
 - 1,098万件 (2020)
 - 1,165万件 (2021)
 - 1,237万件 (2022)
 - 1,325万件 (2023)
 - **1,393万件 (今回)**
- ※ 2024/5/19現在

Rakuten Hotel Review Page

楽天宿・航空券・ツアー予約

国内旅行 国内ツアー レンタカー 高速バス 海外旅行 海外ツアー 海外航空券 海外ホテル 割引クーポン 懸賞広場 観光案内

楽天トラベルトップ > 全国 > 千葉県 > 外房(鴨川・勝浦・御宿・茂原) > 鴨川温泉 > 鴨川シーウールドホテル クチコミ・感想・情報

鴨川シーウールドホテル

★★★★★ 4.12 クチコミ・お客さまの声(886件) この宿泊施設をお気に入りに追加 メルマガ 幹事さん機能 友達登録 シェアする 3 フォト 写真 リビング

日程からプランを探す

- 国内宿泊
- ANA 航空券+宿泊
- JAL 航空券+宿泊
- 日帰り・ディユース
- 日付未定
- チェックイン 2015/06/21
- チェックアウト 2015/06/22
- ご利用部屋数 1 部屋
- ご利用人数 1部屋目：大人(1人)子供(0人)
- 金額(1部屋1泊あたり消費税込) 下限(制限なし) 上限(制限なし)
- 検索

地図

最近見た宿泊施設

11軒の閲覧履歴があります [ページ1/6]

鴨川シーウールドホテル 着めぐらしの宿・夢みさき もっと見る

施設間違情報

- 鴨川シーウールドホテル
- ★トップページ★
- 鴨川シーウールドホテル
- ★鴨シニュース★

楽天カード入会で2,000ポイントプレゼント カード GORA 楽天市場

楽天トラベルの使い方 サイトマップ ヘルプ Languages 予約の確認・キャンセル 30%以上ポイントバック!

ようこそ、楽天トラベルへ 会員登録 ログイン 予約の確認・キャンセル

鴨川シーウールドホテルのクチコミ・感想・情報

総合評価 ★★★★★ 4.12 アンケート件数: 886件

評価内訳
5点 236件
4点 302件
3点 47件
2点 15件
1点 9件

項目別評価
サービス ★★★★★ 4.11
立地 ★★★★★ 4.61
部屋 ★★★★★ 3.53
設備・アメニティ ★★★★★ 3.62
風呂 ★★★★★ 3.53
食事 ★★★★★ 4.10

サービス 立地 食事 部屋 設備・アメニティ 風呂 食事

鴨川シーウールドホテルのクチコミ・お客さまの声

最新の投稿順 評価が高い順 (総合 | サービス | 立地 | 部屋 | 設備・アメニティ | 風呂 | 食事)

772件中 1~20件表示 [1 | 2 | 3 | 4 | 5 | ... 全 39 ページ] ●次の20件

RENDZE_VOUSさんの 鴨川シーウールドホテル のクチコミ (感想・情報)

RENDZE_VOUSさん (3件) [30代/女性] 2015年06月17日 19:20:22

入園パスポート付き、バイキングの夕食と朝食付きでとてもお得な価格で泊まれました。

2日間ともシーウールドに立園出来るのでとても便利です。バイキングも種類が豊富でおいしく楽しくいただきました。温泉は湯船のひととしかないので少しうつりませんが、メインはシーウールドなので仕方ないかなと。

お部屋は古くて他の部屋の物音(子供が走り回る足音など)が気になりました。

量の上に座ったら立ちあちこちかゆくなりました。オーシャンビューで部屋の前の海が海なりうれしいのですが、波の音が大き聞こえて、なかなか眠れませんでした。設備が古くから仕方ないのかな。

総評すると部屋はちょっと微妙ですがお安いですし満足しています。シーウールドのシャチのショーや金曜日と土曜日に見ましたが、全然レベルが違った(笑)。

平日は練習程度のレベルでした。土曜日にもう一度見てよかったです。シャチ以外のショーは平日でもしっかりしていました。

レビューを評価してください このレビューは参考になりましたか?

旅行の目的 … レジャー 同伴者 … 家族 宿泊年月 … 2015年06月

ご利用の宿泊プラン いい値! パリュープラン

ご利用のお部屋 wa5シーウールドが見える特別室禁煙 [洋室]

不適切なレビューを報告する はいいいえ

鴨川シーウールドホテル クチコミ・感想・情報

総合 ★★★★★ 2 投稿者さんの 鴨川シーウールドホテル のクチコミ (感想・情報)

投稿者さん 2015年06月11日 17:03:57

良かったところ
・部屋からの景色(朝日最高でした)
・食事(品数が多く、朝夕とも良かったです)
・フロントの方の対応(お姉さんがとても頑張っていました)以上。

掃除が行き届いているとの口コミを多く見ましたが、そろは思いませんでした。

気にかかる事は多々ありましたが、フロントのお姉さんが一生懸命で、その笑顔に救われた思いです。

フロントスタッフへのお言葉、誠にありがとうございます。

モチベーションアップに繋がりますので、お客様からの声として、スタッフと共有させて頂きます。

機会がございましたら、またご利用をお待ちしております。

旅行の目的 … レジャー 同伴者 … 家族 宿泊年月 … 2015年06月

ご利用の宿泊プラン [洋室 禁煙・特別室] お部屋からシャチやイルカも見える シーウールドと海一望宿泊プラン

ご利用のお部屋 [wa5シーウールドが見える特別室禁煙 [洋室]]

総合 ★★★★★ 4 投稿者さんの 鴨川シーウールドホテル のクチコミ (感想・情報)

投稿者さん 2015年06月11日 07:33:49

夫、2歳半と5ヶ月の子どもの4人で宿泊しました。

【立地】当たり前ですが鴨川シーウールドにとても近く、ゆっくり館内を見学できました。

【部屋】至って普通です。(古いからか、勝の声は少し聞こえます。) トイレ掃除などはしっかりされていました。清浄機などもTEL一本すぐに届けて下さいました。

【食事】夜朝共にバイキング。イスですが子ども用イス、エプロン、ベビーベッドを用意して下さっています。キッズスペースも食事時間中に専門のスタッフの方がおりゆっくり食事ができます。

【風呂】小さな子ども(赤ちゃん)用のグッズ(ベビーベッド、コーナー、バス、おもちゃ、泡ソーブ、支えのあるイス)が揃っていました。お子さん連れも多く気兼ねなく楽しめました。しかしお風呂がひとつしかないのに、温泉を楽しむという雰囲気ではなく、銭湯のお湯が温泉という感じです。

また、23時頃にお風呂に行くと、アメニティやシャンプーが空だったのは少し残念でした。

【サービス】受付スタッフの皆さんとても親切、丁寧です。チェックアウト後に子どもの薬を冷蔵庫にいれておいて欲しいとダメ元で頼むと快く入

いい値! パリュープラン
【最安料金 (目安)】 10,186円~ (消費税込11,000円~)

【当日15:50からアシカと一緒に写真】笑ラジカと一緒にパチリ付プラン [最安料金 (目安)] 10,278円~ (消費税込11,100円~)

【13:40~エコ・アーチームコミュニケーションタイム 1日3組限定】 [最安料金 (目安)] 10,278円~ (消費税込11,100円~)

【夜の水族館探検】 3月~10月の火・木曜日限定プラン [最安料金 (目安)] 10,278円~ (消費税込11,100円~)

【14:50からイルカと一緒にパチリ 2室限 定】 鴨川シーウールド体験プラン [最安料金 (目安)] 10,463円~ (消費税込11,300円~)

今しかない!★アビ料理付シーウールド入園バスポート付で大満足♪5月・6月の月~木曜日限定プラン [最安料金 (目安)] 10,926円~ (消費税込11,800円~)

【便利な赤ちゃんグッズ付】初!お泊りはお母さんも嬉しい★赤ちゃんなつ得プラン [最安料金 (目安)] 10,926円~ (消費税込11,800円~)

お子様にも大好評!オーシャンビュープラン [最安料金 (目安)] 11,112円~ (消費税込12,000円~)

【80cmのジャンボサイズ】海の王者シャチぬいぐるみ付プラン [最安料金 (目安)] 11,204円~ (消費税込12,100円~)

房総2大テーマパーク満喫「マザーフ 农場チケット」付プラン [最安料金 (目安)] 11,389円~ (消費税込12,300円~)

【当日14:50~イルカ

HARADA Tomohiko

travel.rakuten.co.jp/HOTEL/2910/review.html

鴨川シーウールドホテル クチコミ・感想・情報

総合 ★★★★★ 2 投稿者さんの 鴨川シーウールドホテル のクチコミ (感想・情報)

投稿者さん 2015年06月11日 19:32:50

この度は、ご利用頂きました誠にありがとうございました。

客室内清掃の件、大変申し訳ございませんでした。

重要改善として、早急に対応いたします。

今後は、この様な事の無いように、清掃・点検を強化いたします。

フロントスタッフへのお言葉、誠にありがとうございます。

モチベーションアップに繋がりますので、お客様からの声として、スタッフと共有させて頂きます。

機会がございましたら、またご利用をお待ちしております。

旅行の目的 … レジャー 同伴者 … 家族 宿泊年月 … 2015年06月

ご利用の宿泊プラン [洋室 禁煙・特別室] お部屋からシャチやイルカも見える シーウールドと海一望宿泊プラン

ご利用のお部屋 [wa5シーウールドが見える特別室禁煙 [洋室]]

総合 ★★★★★ 4 投稿者さんの 鴨川シーウールドホテル のクチコミ (感想・情報)

投稿者さん 2015年06月11日 19:25:49

夫、2歳半と5ヶ月の子どもの4人で宿泊しました。

この度は、ご利用頂きました誠にありがとうございました。

詳細にご感想頂きまして、ありがとうございます。

今後の参考にさせて顶きます。

また、スタッフ対応に際しまして、お褒めのお言葉を頂戴しまして、とても嬉しく思います。

モチベーションアップに繋がりますので、お客様からの声として、スタッフと共有させて頂きます。

最後に、「アメニティ・シャンプー」の件、大変申し訳ございませんでした。

早急に対応をして、改善を行います。

貴重なご意見を、ありがとうございます。

機会がございましたら、またご利用をお待ちしております。

旅行の目的 … レジャー 同伴者 … 家族 宿泊年月 … 2015年06月

ご利用の宿泊プラン いい値! パリュープラン

ご利用のお部屋 [wa5海側和室 (10畳バス・トイレ付) タイプ]

いい値! パリュープラン
【最安料金 (目安)】 10,186円~ (消費税込11,000円~)

【当日15:50からアシカと一緒に写真】笑ラジカと一緒にパチリ付プラン [最安料金 (目安)] 10,278円~ (消費税込11,100円~)

【13:40~エコ・アーチームコミュニケーションタイム 1日3組限定】 [最安料金 (目安)] 10,278円~ (消費税込11,100円~)

【夜の水族館探検】 3月~10月の火・木曜日限定プラン [最安料金 (目安)] 10,278円~ (消費税込11,100円~)

【14:50からイルカと一緒にパチリ 2室限 定】 鴨川シーウールド体験プラン [最安料金 (目安)] 10,463円~ (消費税込11,300円~)

今しかない!★アビ料理付シーウールド入園バスポート付で大満足♪5月・6月の月~木曜日限定プラン [最安料金 (目安)] 10,926円~ (消費税込11,800円~)

お子様にも大好評!オーシャンビュープラン [最安料金 (目安)] 11,112円~ (消費税込12,000円~)

【便利な赤ちゃんグッズ付】初!お泊りはお母さんも嬉しい★赤ちゃんなつ得プラン [最安料金 (目安)] 10,926円~ (消費税込11,800円~)

お子様にも大好評!オーシャンビュープラン [最安料金 (目安)] 11,112円~ (消費税込12,000円~)

【80cmのジャンボサイズ】海の王者シャチぬいぐるみ付プラン [最安料金 (目安)] 11,204円~ (消費税込12,100円~)

房総2大テーマパーク満喫「マザーフ 农場チケット」付プラン [最安料金 (目安)] 11,389円~ (消費税込12,300円~)

【当日14:50~イルカ

鴨川シーワールドホテルのクチコミ・お客様の声

[●ホテル・旅行のクチコミTOPへ](#)

総合評価

4.12

アンケート件数：886件

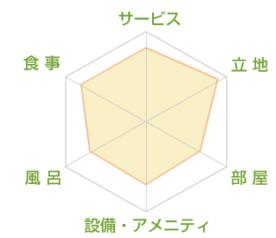
評価内訳

- 5点
- 4点
- 3点
- 2点
- 1点

236件
302件
47件
15件
9件

項目別の評価

サービス	4.11
立地	4.61
部屋	3.53
設備・アメニティ	3.62
風呂	3.53
食事	4.10



総合 2

投稿者さんの 鴨川シーワールドホテル のクチコミ (感想・情報)



投稿者さん

2015年06月11日 17:03:57

良かったところ

- ・部屋からの景色（朝日最高でした）
- ・食事（品数が多く、朝夕とも良かったです）
- ・フロントの方の対応（お姉さんがとても頑張っていました）以上。

掃除が行き届いているとの口コミを多く見ましたが、それは思いませんでした。

気にかかることは多々ありましたが、フロントのお姉さんが一生懸命で、その笑顔に救われた思います。

評価

... 総合 2

- | | |
|----------|---|
| サービス | 2 |
| 立地 | 4 |
| 部屋 | 4 |
| 設備・アメニティ | 2 |
| 風呂 | 2 |
| 食事 | 4 |

旅行の目的

... レジャー

同伴者

... 家族

宿泊年月

... 2015年06月



鴨川シーワールドホテル

2015年06月11日 19:32:50

この度は、ご利用頂きまして誠にありがとうございます。

客室内清掃の件、大変申し訳

重要改善として、早急に対応いたします。

今後は、この様な事の無いように、清掃・点検を強化いたします。

テキストデータ

フロントスタッフへのお言葉
誠にありがとうございます。

セラピーアップに繋がる
お客様からの声として、
スタッフと共有させて頂きます。

数値評価

機会がございましたら、またご利用をお待ちしております。

テキストマイニングの手順

- データをよく知る
 - データ件数や構成比を集計 → データを理解する
 - 旅行目的別の人気エリアは?
 - 同伴者別の人気エリアは?
 - 数値評価による人気エリアの差異は?
- テーマを設定する
 - 解決すべき課題を決める → 分析目的を明確にする
 - 数値評価が低い原因是?
 - 高評価の施設に学ぶ改善点は?
- テキスト分析に取り組む
 - これら課題を解決するために、テキスト分析を実施

データ理解

実習用のデータ (Webサイトクローリング)

● 楽天トラベル のクチコミデータ

- 収集期間は 2021-2022 および 2023-2024(～GW明け) の 2セット
- 以下の 10 エリアごと同数に 1,000件ずつ ランダムサンプリング
- データ件数は 1万件 × 2セット

レジャー	5エリア	登別、草津、箱根、道後、湯布院	1,000件 × 10エリア
ビジネス	5エリア	札幌、名古屋、東京、大阪、福岡	= 計10,000件

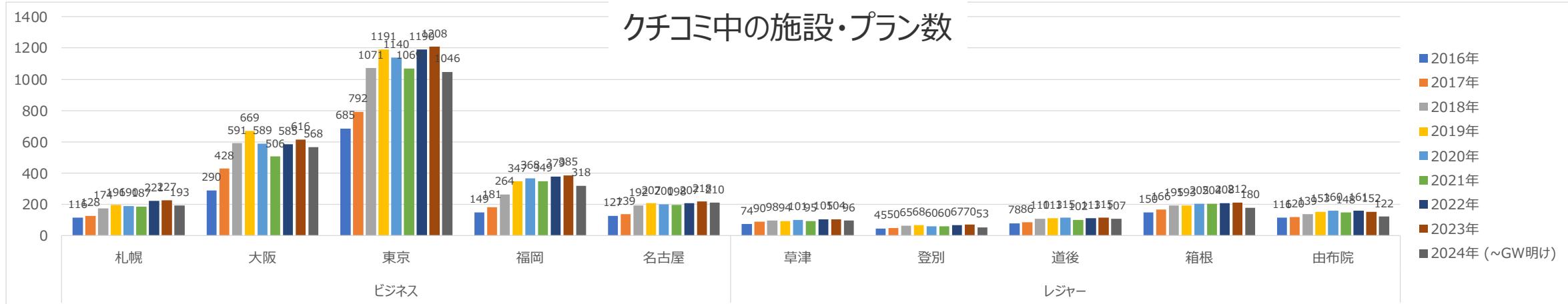
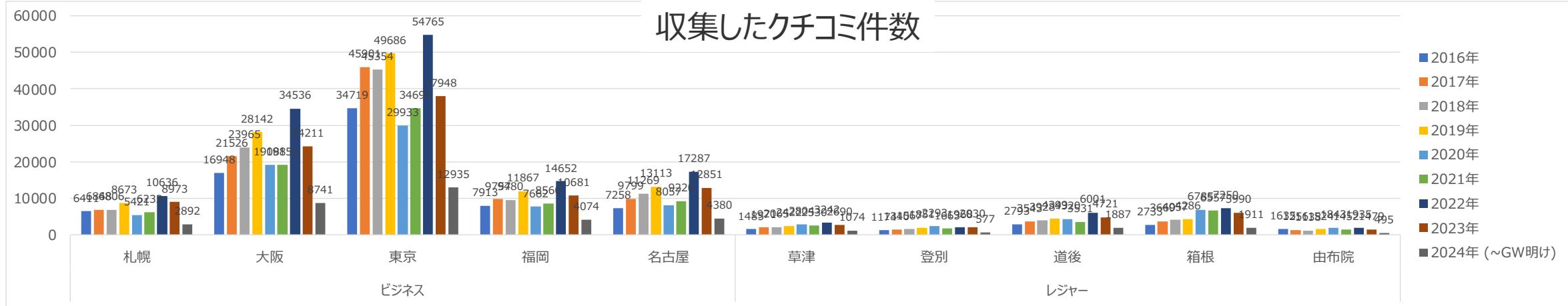
● 楽天トラベル のクチコミデータ

- 収集期間は 2021-2022 および 2023-2024(～GW明け) の 2セット
- 以下の 10 エリアごと同数に 1,000件ずつ ランダムサンプリング
- データ件数は 1万件 × 2セット
- データ項目は 18項目 (テキスト1項目+その他の属性17項目)

施設情報	4項目 カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目 コメント (テキスト)
ユーザー評価	7項目 総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事
その他の分類	2項目 旅行の目的, 同伴者
宿泊日	1項目 宿泊年月
ユーザー情報	3項目 ユーザー, 年代, 性別

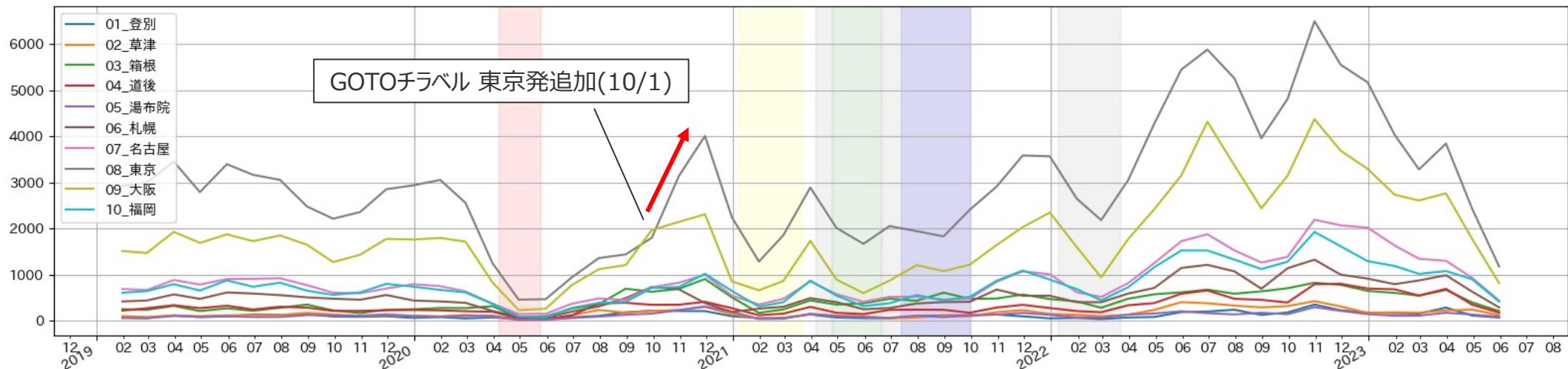
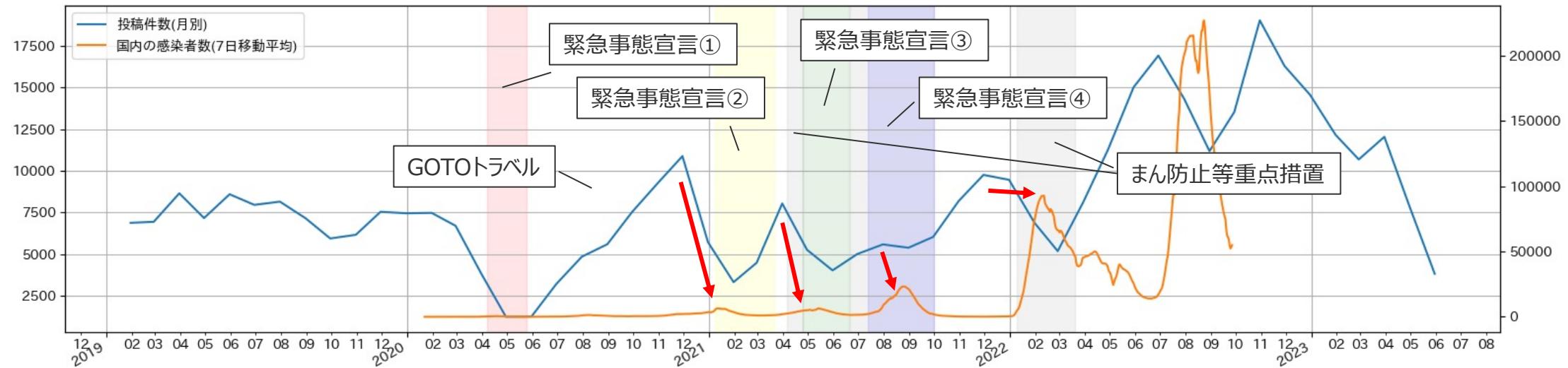
参考 — Webサイトクローリング

● 全量では 166.4万件、2021-2022は19.4万件、2023-2024は15.1万件



(参考) COVID-19 の影響

- クチコミの件数と感染者数の増減が連動 → クチコミ件数が一定の人流を反映している

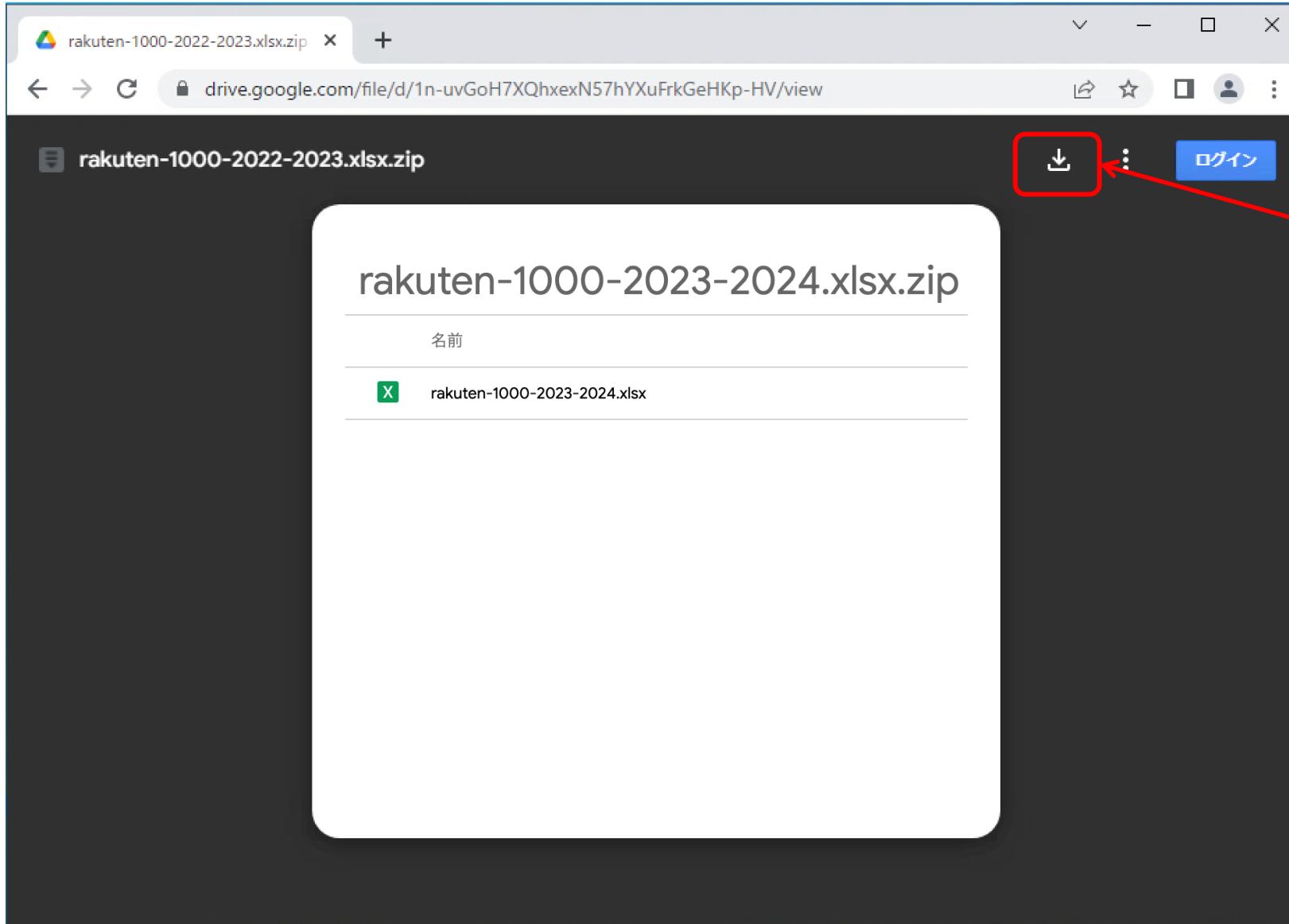


実習用データ — ファイル一覧

● 実習用データは以下の通り → 主に「**rakuten-1000-2023-2024.xlsx**」を使用する

ファイル名	件数	データセット	備考
<u>rakuten-1000-2023-2024.xlsx.zip</u>	10,000	<ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・エリアごと 1,000件 (ランダムサンプリング)・期間: 2023/1~2024 GW明け	本講義の全体を通して使用する
<u>rakuten-1000-2021-2022.xlsx.zip</u>	10,000	<ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・エリアごと 1,000件 (ランダムサンプリング)・期間: 2021/1~2022/12	演習用 (年度で比較する場合など)
<u>rakuten-all-2023-2024-tsv.zip</u>	150,537	<ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・サンプリング前の全データ・期間: 2023/1~2024 GW明け	参考用
<u>rakuten-all-2021-2022-tsv.zip</u>	193,721	<ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・サンプリング前の全データ・期間: 2021/1~2022/12	参考用
<u>rakuten-all-tsv.zip</u>	1,664,301	<ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・サンプリング前の全データ・期間: 2009/3~2024 GW明け	参考用

(参考) Google Drive ダウンロード画面



ここをクリックすると
ダウンロードが始ま
ります

テキストマイニングの手順

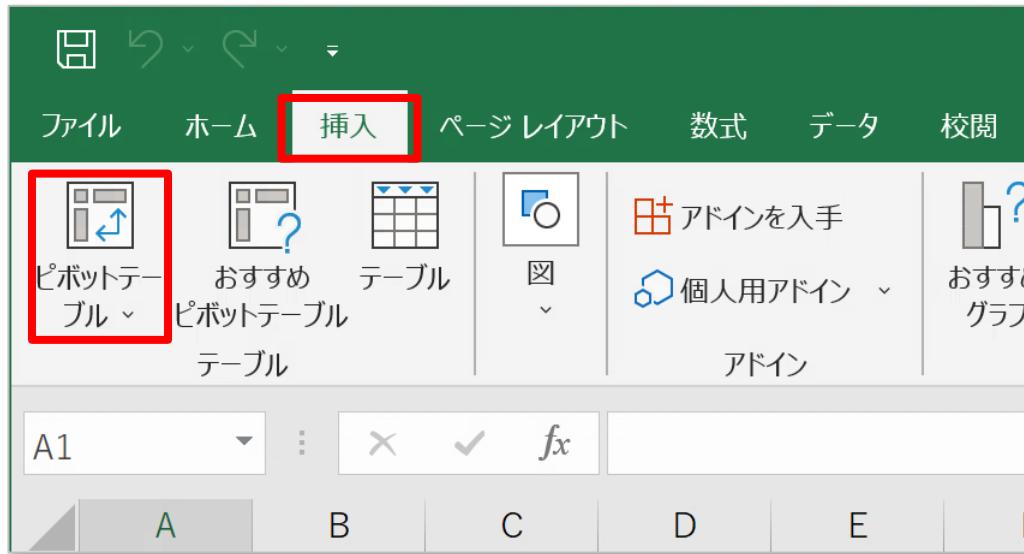
- データをよく知る
 - データ件数や構成比を集計 → データを理解する
 - 旅行目的別の人気エリアは?
 - 同伴者別の人気エリアは?
 - 数値評価による人気エリアの差異は?
- テーマを設定する
 - 解決すべき課題を決める → 分析目的を明確にする
 - 数値評価が低い原因是?
 - 高評価の施設に学ぶ改善点は?
- テキスト分析に取り組む
 - これら課題を解決するために、テキスト分析を実施

参考 — EXCEL を使った集計

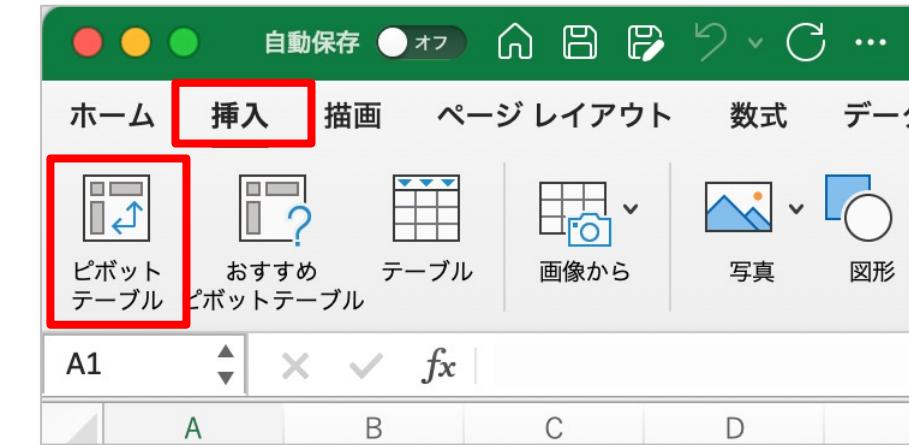
● EXCEL のピボットテーブルを使ってデータを集計する

- ① ファイル **rakuten-1000-2023-2024.xlsx** を開く
- ② A～R 列を選択し、ピボットテーブルを作成する
- ③ [挿入] タブ [テーブル] グループの [ピボットテーブル] ボタンをクリックする

Windows



Mac



データ理解 — 集計例

①件数 (エリア別)

行ラベル	個数 / コメン
■ A_レジャー	5000
01_登別	1000
02_草津	1000
03_箱根	1000
04_道後	1000
05_湯布院	1000
■ B_ビジネス	5000
06_札幌	1000
07_名古屋	1000
08_東京	1000
09_大阪	1000
10_福岡	1000
総計	10000

②投稿者の傾向 (年代別x性別)

行ラベル	個数 / コメン	列ラベル			
		男性	女性	na	総計
10代		0.00%	0.05%	0.00%	0.05%
20代		0.70%	1.16%	0.00%	1.86%
30代		1.94%	2.60%	0.00%	4.54%
40代		4.54%	3.76%	0.00%	8.30%
50代		7.95%	4.17%	0.00%	12.12%
60代		6.34%	1.96%	0.00%	8.30%
70代		1.38%	0.36%	0.00%	1.74%
80代		0.07%	0.06%	0.00%	0.13%
na		0.00%	0.00%	62.95%	62.95%
120代		0.00%	0.01%	0.00%	0.01%
総計		22.92%	14.13%	62.95%	100.00%

③投稿者の傾向 (性別xカテゴリ別)

行ラベル	個数 / コメン	列ラベル		
		A_レジャー	B_ビジネス	総計
男性		22.52%	23.32%	22.92%
女性		15.98%	12.28%	14.13%
na		61.50%	64.40%	62.95%
総計		100.00%	100.00%	100.00%

データ理解 — 集計例

④投稿者の傾向 (性別xカテゴリーエリア別)

個数 / コメント	列ラベル	A_レジャー 集計										B_ビジネス 集計			総計
		01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡				
男性	A_レジャー	24.10%	22.80%	16.20%	27.10%	22.40%	22.52%	26.40%	25.30%	21.30%	20.50%	23.10%	23.32%	22.92%	
女性	A_レジャー	16.00%	16.00%	17.20%	11.40%	19.30%	15.98%	12.70%	12.90%	11.70%	11.80%	12.30%	12.28%	14.13%	
na	A_レジャー	59.90%	61.20%	66.60%	61.50%	58.30%	61.50%	60.90%	61.80%	67.00%	67.70%	64.60%	64.40%	62.95%	
総計		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

⑤投稿者の傾向 (年代別xカテゴリーエリア別)

個数 / コメント	列ラベル	A_レジャー 集計										B_ビジネス 集計			総計
		01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡				
10代	A_レジャー	0.10%	0.00%	0.10%	0.00%	0.00%	0.04%	0.00%	0.00%	0.10%	0.20%	0.00%	0.06%	0.05%	
20代	A_レジャー	0.80%	3.40%	2.60%	1.60%	1.90%	2.06%	2.00%	2.50%	0.70%	1.60%	1.50%	1.66%	1.86%	
30代	A_レジャー	4.20%	5.00%	4.40%	4.00%	6.30%	4.78%	3.40%	4.90%	4.80%	4.30%	4.10%	4.30%	4.54%	
40代	A_レジャー	8.30%	9.00%	7.30%	6.70%	10.20%	8.30%	8.20%	8.80%	7.10%	9.20%	8.20%	8.30%	8.30%	
50代	A_レジャー	13.40%	11.60%	8.70%	13.30%	11.60%	11.72%	15.10%	12.00%	11.80%	9.80%	13.90%	12.52%	12.12%	
60代	A_レジャー	10.70%	8.10%	7.60%	10.20%	9.90%	9.30%	9.00%	8.50%	7.10%	6.10%	5.80%	7.30%	8.30%	
70代	A_レジャー	2.30%	1.60%	2.30%	2.70%	1.50%	2.08%	1.20%	1.50%	1.40%	1.00%	1.90%	1.40%	1.74%	
80代	A_レジャー	0.30%	0.10%	0.40%	0.00%	0.30%	0.22%	0.10%	0.00%	0.00%	0.10%	0.00%	0.04%	0.13%	
120代	A_レジャー	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.10%	0.00%	0.00%	0.00%	0.00%	0.02%	0.01%	
na	A_レジャー	59.90%	61.20%	66.60%	61.50%	58.30%	61.50%	60.90%	61.80%	67.00%	67.70%	64.60%	64.40%	62.95%	
総計		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

データ理解 — 集計例

⑥投稿者の傾向 (同行者別xカテゴリ-エリア別)

個数 / コメント 行ラベル	A_レジャー 集計											B_ビジネス 集計			総計
	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
一人	24.80%	13.70%	14.00%	44.40%	13.10%	22.00%	57.90%	64.30%	65.60%	57.70%	53.50%	59.80%		40.90%	
家族	63.10%	64.30%	66.10%	42.70%	69.60%	61.16%	30.90%	24.10%	23.40%	29.40%	32.10%	27.98%		44.57%	
恋人	4.80%	14.60%	11.20%	4.90%	8.30%	8.76%	4.40%	4.20%	4.10%	4.40%	3.90%	4.20%		6.48%	
友達	5.30%	5.70%	7.10%	5.50%	7.80%	6.28%	4.40%	4.40%	4.80%	6.80%	7.60%	5.60%		5.94%	
仕事仲間	1.50%	0.70%	0.50%	1.80%	0.60%	1.02%	1.90%	2.50%	1.30%	1.40%	2.40%	1.90%		1.46%	
その他	0.50%	1.00%	1.10%	0.70%	0.60%	0.78%	0.50%	0.50%	0.80%	0.30%	0.50%	0.52%		0.65%	
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%		100.00%	

⑦数値評価の構成 (総合別xカテゴリ-エリア別)

個数 / コメント 行ラベル	A_レジャー 集計											B_ビジネス 集計			総計
	01_登別	02_草津	03_箱根	04_道後	05_湯布院	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
5	41.90%	48.40%	48.90%	49.60%	67.90%	51.34%	41.70%	36.70%	36.90%	41.90%	38.50%	39.14%		45.24%	
4	41.30%	36.10%	36.90%	36.50%	22.50%	34.66%	41.90%	47.30%	41.70%	41.10%	40.90%	42.58%		38.62%	
3	9.90%	9.90%	7.80%	9.20%	4.90%	8.34%	11.60%	12.00%	14.30%	11.70%	12.60%	12.44%		10.39%	
2	4.30%	3.40%	4.30%	2.90%	3.20%	3.62%	3.20%	2.70%	4.60%	3.60%	5.00%	3.82%		3.72%	
1	2.60%	2.20%	2.10%	1.80%	1.50%	2.04%	1.60%	1.30%	2.50%	1.70%	3.00%	2.02%		2.03%	
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%		100.00%	

データ理解 — 集計例

⑧-a 数値評価の平均 (エリア別×数値評価別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
■ A_レジャー	4.25	4.25	4.13	4.05	4.29	4.29	4.30
01_登別	4.07	4.21	3.95	3.90	4.34	4.08	4.16
02_草津	4.23	4.22	4.07	3.97	4.32	4.20	4.25
03_箱根	4.24	4.12	4.18	4.05	4.29	4.33	4.26
04_道後	4.19	4.41	4.07	4.00	4.03	4.19	4.29
05_湯布院	4.51	4.28	4.37	4.35	4.46	4.61	4.52
■ B_ビジネス	3.98	4.30	4.01	3.88	3.74	4.05	4.13
06_札幌	4.05	4.30	4.09	3.93	3.79	4.12	4.19
07_名古屋	4.00	4.25	4.04	3.89	3.75	4.03	4.15
08_東京	3.93	4.38	3.94	3.82	3.70	3.99	4.06
09_大阪	4.01	4.35	4.05	3.93	3.82	4.06	4.18
10_福岡	3.93	4.24	3.96	3.84	3.64	4.01	4.07

⑧-b 数値評価の平均 (カテゴリ別×数値評価別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.25	4.25	4.13	4.05	4.29	4.29	4.30
B_ビジネス	3.98	4.30	4.01	3.88	3.74	4.05	4.13

データ理解 — 集計例

⑨-a 数値評価の平均 (20~30代, 性別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
■ A_レジャー	4.48	4.32	4.35	4.33	4.42	4.46	4.46
男性	4.38	4.23	4.30	4.18	4.38	4.48	4.44
女性	4.55	4.37	4.38	4.42	4.45	4.44	4.47
■ B_ビジネス	4.18	4.36	4.19	4.10	3.93	4.41	4.28
男性	4.16	4.33	4.13	4.10	3.93	4.35	4.23
女性	4.19	4.39	4.24	4.10	3.93	4.45	4.32

⑨-b 数値評価の平均 (40~50代, 性別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
■ A_レジャー	4.29	4.33	4.15	4.10	4.32	4.35	4.34
男性	4.23	4.29	4.10	4.05	4.29	4.33	4.32
女性	4.38	4.38	4.21	4.17	4.36	4.37	4.36
■ B_ビジネス	4.00	4.30	4.07	3.91	3.79	4.09	4.17
男性	3.93	4.27	4.00	3.85	3.73	4.07	4.12
女性	4.13	4.36	4.19	4.02	3.89	4.13	4.26

⑨-c 数値評価の平均 (60~80代, 性別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
■ A_レジャー	4.23	4.24	4.08	3.98	4.29	4.26	4.31
男性	4.21	4.20	4.05	3.92	4.26	4.27	4.27
女性	4.27	4.36	4.16	4.12	4.38	4.26	4.39
■ B_ビジネス	3.95	4.30	4.01	3.80	3.78	3.87	4.14
男性	3.91	4.24	3.98	3.77	3.75	3.79	4.13
女性	4.19	4.59	4.15	3.96	3.94	4.23	4.21

データ理解 — 集計結果の整理

観点	データの特徴	テキスト分析時に注意すべき点
年代別・性別	<ul style="list-style-type: none">約60%が年代や性別を表明していない・・	<ul style="list-style-type: none">レビュー観点がある年代や性別に偏っている可能性・・
目的別	<ul style="list-style-type: none">レジャーは家族が多い、ビジネスは一人が多い・・	<ul style="list-style-type: none">レビューの観点が性別によって偏っている可能性・・
数値評価 (総合)	<ul style="list-style-type: none">旅行目的によらず評価は高め・・	<ul style="list-style-type: none">コメントが好評価に偏っている可能性・・
数値評価 (項目ごと)	<ul style="list-style-type: none">レジャーの評価は、風呂や食事 > 設備や部屋・・	<ul style="list-style-type: none">旅行目的によって評価の観点や重みが異なっている可能性・・
全体	<ul style="list-style-type: none">・・・	

- グループワーク (~20:40)
 - データ集計によって発見した、データセットに関する特徴や傾向、テキスト分析時に注意すべき点について、グループ内で討論する
 - 前ページの表を参考に、集計結果から得られた知見を整理する

(再掲) 数値評価で違いを見るのは難しい

【再掲】⑧-a 数値評価の平均 (エリア別×数値評価別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂			
■ A_レジャー	4.25	4.25	4.13	4.05	4.29	4.29	4.29	4.50
01_登別	4.07	4.21	3.95	3.90	4.34	4.08	4.16	
02_草津	4.23	4.22	4.07	3.97	4.32	4.20	4.25	
03_箱根	4.24	4.12	4.18	4.05	4.29	4.33	4.26	
04_道後	4.19	4.41	4.07	4.00	4.03	4.19	4.29	
05_湯布院	4.51	4.28	4.37					4.52
■ B_ビジネス	3.98	4.30	4.01					4.13
06_札幌	4.05	4.30	4.09					4.19
07_名古屋	4.00	4.25	4.04	3.89	3.75			4.15
08_東京	3.93	4.38	3.94	3.82	3.70	3.99		4.06
09_大阪	4.01	4.35	4.05	3.93	3.82	4.06		4.18
10_福岡	3.93	4.24	3.96	3.91	3.64	4.01		4.07

- ユーザーの8割が4~5の評価、1~2をつけない→本音が見えない

- 同じ点数でもテキストを見れば差異があるかも

- すべての項目に回答する→どこに注目しているかよくわからない

【再掲】⑧-b 数値評価の平均 (カテゴリ別×数値評価別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.25	4.25	4.13	4.05	4.29	4.29	4.30
B_ビジネス	3.98	4.30	4.01	3.88	3.74	4.05	4.13

辻井康一 and 津田和彦「テキストマイニングを用いた宿泊レビューからの注目情報抽出方法」, デジタルプラクティス 3.4 (2012): 289-296.

【再掲】⑧-b 数値評価の平均 (カテゴリ別×数値評価別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.25	4.25	4.13	4.05	4.29	4.29	4.30
B_ビジネス	3.98	4.30	4.01	3.88	3.74	4.05	4.13

● 数値評価のみから違いを見つけるのは難しい！！

- ・ ユーザーの 8割が 4~5 の評価, 1~2をつけない
- ・ ユーザーは 注目の有無に関係なくすべての項目に回答

→ レジャーとビジネスでは、評価すべき項目も異なることを確認した

→ テキストと対応付ければ、同じ点数でも差異があることを確認した

レポート課題

- 以下を PDF ファイルで提出してください
 - データ集計により作成した「集計表」のキャプチャ (P.73~77) ※ページ番号は各スライド右下に記載
 - 作成した「集計結果の整理」の表 (P.78) ※ページ番号は各スライド右下に記載
- ※ 何らかの事情で上記2つを提出できない場合、本日の講義の感想を文章で記述してください

レポート形式	提出先	期限
PDF	manaba	次回～18:20

無償で利用できる機械学習環境

- 近年、機械学習の教育・研究を目的とした研究用ツールが相次いで登場

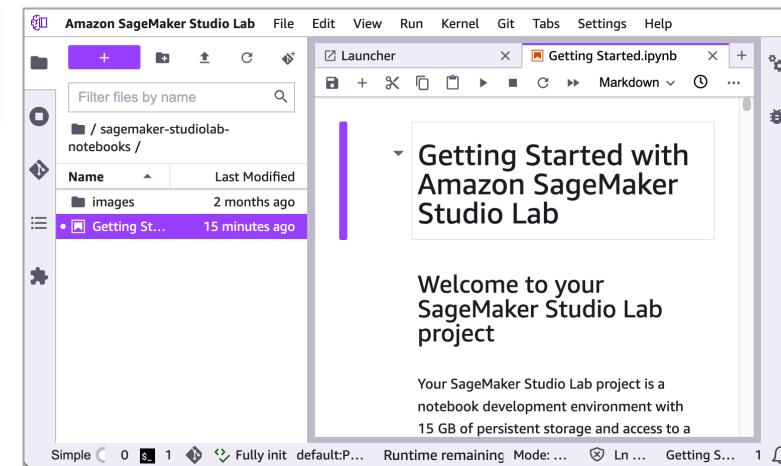
 Colaboratory

<https://colab.research.google.com>



 Amazon SageMaker Studio Lab

<https://studiolab.sagemaker.aws/>



演習で使用
↓

	Colab(無償版)	Studio Lab
GPU	T4(16GB)	T4(16GB)
最長実行時間	12時間	CPU:12時間 GPU:4時間
メモリ	12GB	15GB
ディスク	CPU:100GB GPU:78GB	15GB (永続化)
ターミナル	×	○
ランタイムの保存と再開	×	○
費用	無償	無償
その他	Googleアカウントが必要	AWSアカウントは不要 (クレカ不要)

SageMaker Studio Lab のアカウント作成

- <https://studiolab.sagemaker.aws/> にログインして、アカウントを作成してください

amazon
SageMaker Studio Lab

Sign in Request account

Request account

Request a free Amazon SageMaker Studio Lab account.

Enter your email*

Enter your first name

Enter your last name

Select your country

Enter your company or organization name

Select your occupation

Why are you interested in Amazon SageMaker Studio Lab?

Enter referral code **XXX-XXXX**

Submit request

アカウント作成手順

1. [アカウント作成フォーム](#)からアカウントの申し込みを行う
注意: リファラルコードをアカウント作成フォームに忘れずに入力ください (受講者限定です)
2. 「Account request confirmed ...」のメールを受信し、メール内のリンクからアカウントを作成する
→ リクエストの受付はすぐにメールが届きます
3. 「Verify your email ...」のメールを受信し、メール内のリンクからメールアドレスを認証する
→ リファラルコードを利用している場合は2~3分以内に結果が届きます
4. 「Your account is ready ...」のメールを受信する
→ これで「Sign in」できます

※ リファラルコードの有効期間: 2024/6/27 ~ 2024/7/12