

# 生成系 AI との付き合い方 — IoT デバイスとクラウドとの融合へ

---

Sep. 1<sup>st</sup>, 2023

SWEST25

株式会社ソラコム

テクノロジー・エバンジェリスト

松下 享平 (Max / @ma2shita)

# 自己紹介



株式会社ソラコム / テクノロジー・エバンジェリスト

松下 享平 (まつした こうへい) "Max"

静岡県民 🏞️ 新幹線通勤族 🚄

講演や執筆を中心に活動、登壇回数500以上/累計

経歴: 東証二部ハードウェアメーカーで情シス部門、  
EC 事業責任者、IoT 事業開発を経て2017年より現職

好きな言葉「論よりコード」

AWS ヒーロー (2020年受賞)



# イベントの振り返りにご活用ください！



ハッシュタグ  
**#SWEST25**  
**#SORACOM**



@SORACOM\_PR



fb.com/soracom.jp



youtube.com/c/SORACOM\_Japan



instagram.com/soracom.official



# 70分間の過ごし方 (アジェンダ)

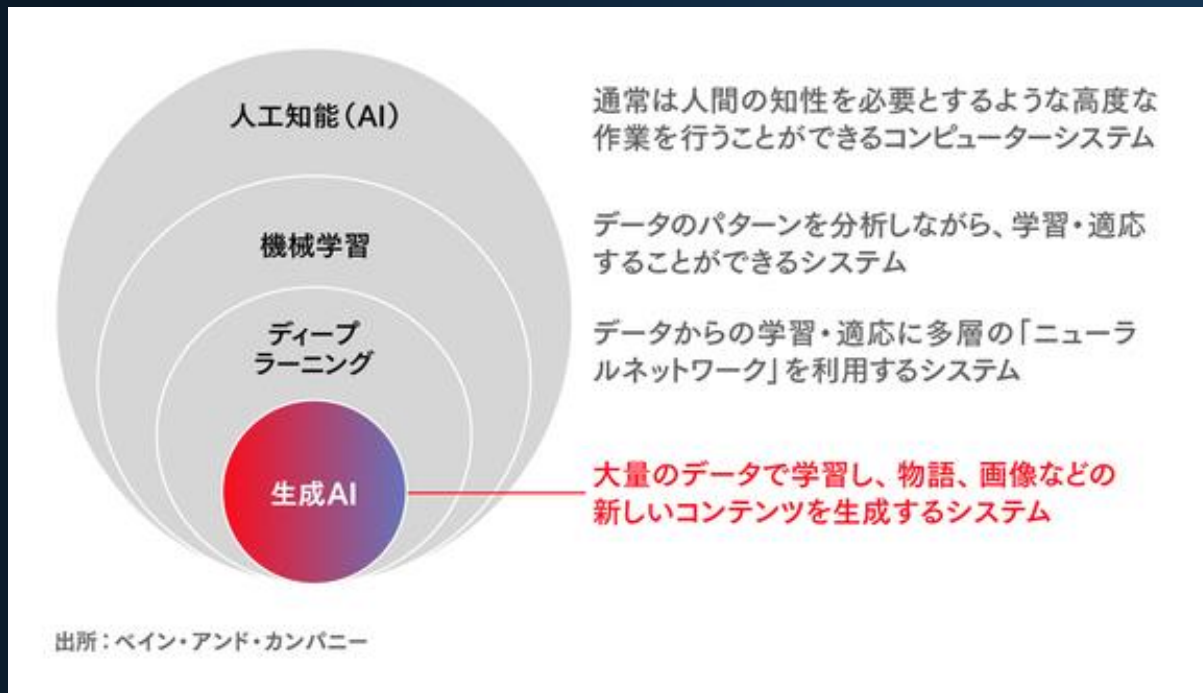
- GenAI の振り返りと、プロンプトエンジニアリング  
《 QA 》
- IoT と GenAI の関係性  
《 QA 》
- ハードウェアからの GenAI 利用の実態
- GenAI 利用と通信の活用 — 適した通信技術
- このあと、どうしたらよいのか？  
— GenAI との付き合い方  
《 QA 》



# “AI”における生成系 AI (GenAI) の位置づけ

## 生成系 AI (Generative AI; GenAI)

- 機械学習(ML)における分類「識別モデル」と「生成モデル」のうち、生成に着目した呼称。
- 具体的には画像を生成するモデルや、自然言語を扱う大規模言語モデル(LLM)を指す。

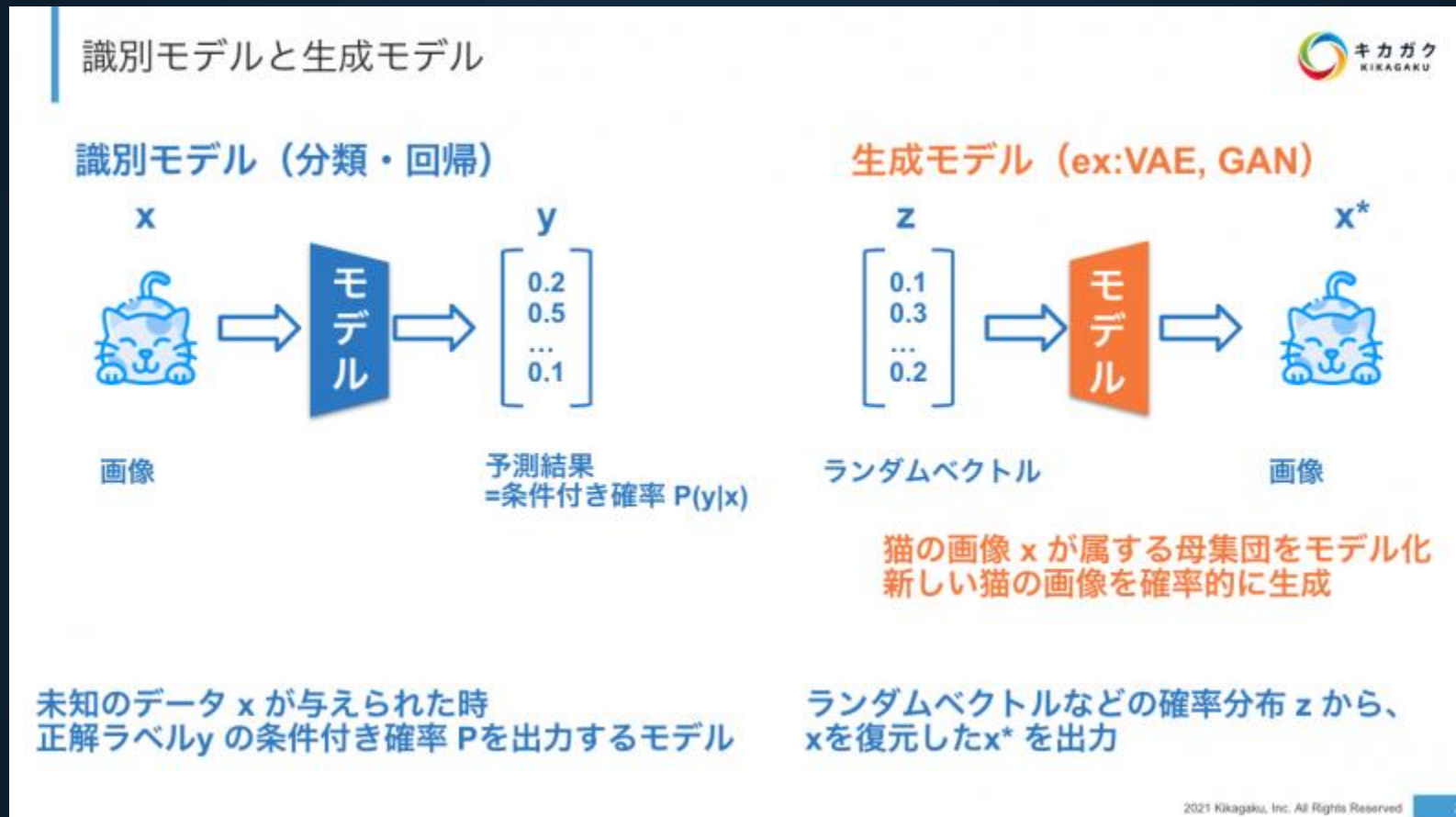


## 従来の ML や深層学習(DL)と異なる点

これまでは「学習」と「推論(生成)」の2つの作業が一体化していることが多かった。

GenAI は学習済みモデルが提供され、利用者による学習作業が不要で、成果が得られる状態。

# 識別モデルと生成モデル



出典: キカガク 谷口. “【初心者向け】 Stable Diffusion や Midjourney を支える技術 画像生成入門 1”. キカガクブログ.  
2022-10-06. <https://blog.kikagaku.co.jp/image-generation-course1>, (参照 2023-08-10).

これまでの機械学習は、識別モデル(図左側)による問題解決が主だった  
例) 画像を入力 → これは何? = 分類

# 大規模言語モデル(Large Language Model; LLM) とパラメーター数

【LLM とは】 入力単語群を基に「次の単語を予測して生成」

- 多くの LLM が公開されているが基本的には同じ

初期入力

IoT とは、遠くのモノや現場で起こっているコトをデジタル化する技術です。その用途は、

さらに入力



この入力なら  
次の単語はこれだ

出力 = 生成

...その用途は、製造業における

...その用途は、製造業における機械の監視や、

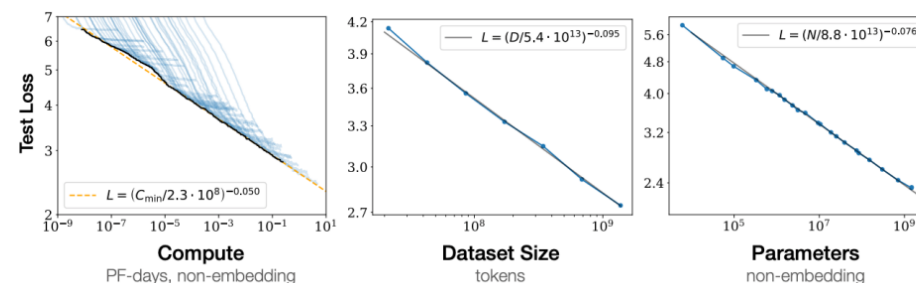
...その用途は、製造業における機械の監視や、地域防災を

## LLM の評価指標 「パラメーター数」

- パラメーターの数 = モデルの容量
- 言語や用途に特化している場合もあるため単純比較はできないが、予測精度の評価目安にはなる

### スケール則 (scaling law) : 極めて重要な法則

- データを増やし、計算能力を増やし、パラメータの数 (モデルの容量) を増やせば、精度がどんどん上がる。
- つまり、大きければ大きいほど性能があがる。
- GPT-3は、ウェブの大きなデータ (4兆単語) で事前学習。1750億パラメータ。学習に数億~数十億円のコスト。



出所: 松尾豊. "AIの進化と日本の戦略". 自民党AIの進化と実装に関するプロジェクトチーム(第2回). 2020-02-17. [https://note.com/akihisa\\_shiozaki/n/n4c126c27fd3d](https://note.com/akihisa_shiozaki/n/n4c126c27fd3d), (参照 2023-08-10).

# 例: line-corporation/japanese-large-lm

```
text = generator(  
    "おはようございます、今日の天気は",  
    max_length=30,  
    do_sample=True,  
    pad_token_id=tokenizer.pad_token_id,  
    num_return_sequences=5,  
)
```

コードの抜粋

出力の抜粋

# 下記は生成される出力の例

```
# [ {'generated_text': 'おはようございます、今日の天気は雨模様ですね。梅雨のこの時期の 朝は洗濯物が  
# {'generated_text': 'おはようございます、今日の天気は晴れ。 気温は8°C位です。 朝晩は結構冷え込む  
# {'generated_text': 'おはようございます、今日の天気は曇りです。 朝起きたら雪が軽く積もっていた。  
# {'generated_text': 'おはようございます、今日の天気は☁のち☀です。 朝の気温5°C、日中も21°Cと  
# {'generated_text': 'おはようございます、今日の天気は晴天ですが涼しい1日です、気温は午後になり低
```



# 日本国内の主だった LLM

LLM 名	パラメーター数	特化言語	提供元	公開時期	
GPT-3	1,750億	汎用	OpenAI	2020年11月	※1
GPT-4	(非公開)	汎用	OpenAI	2023年3月	※2
OpenCALM	68億	日本語	サイバーエージェント	2023年5月	※3
rinna	36億	日本語	rinna	2023年5月	※4
japanese-large-lm	36億	日本語	LINE	2023年8月	※5
Weblab-10B	100億	日・英	松尾研究室	2023年8月	※6

他にも「[NEC の LLM \(NEC\)](#)」がある。海外では「[Llama 2](#) (Meta)」 「[Claude 2](#) (Anthropic)」 「[Falcon LLM](#) (Technology Innovation Institute)」 が公開されている  
最新リストを追うならば [Open LLM Leaderboard](#)(Hugging Face) をチェック

※1 “Language Models are Few-Shot Learners”. arXiv:2005.14165.  
2020-07-22. <https://arxiv.org/abs/2005.14165>, (参照 2023-08-10).

※2 James Vincent. “OpenAI CEO Sam Altman on GPT-4: ‘people are begging to be disappointed and they will be’”. The Verge.  
2023-01-18. <https://www.theverge.com/23560328/openai-gpt-4-rumor-release-date-sam-altman-interview>, (参照 2023-08-10).

※3 “サイバーエージェント、最大68億パラメータの日本語LLM（大規模言語モデル）を一般公開 —オープンなデータで学習した商用利用可能なモデルを提供—”. サイバーエージェント プレスリリース.  
2020-05-17. <https://www.cyberagent.co.jp/news/detail/id=28817>, (参照 2023-08-10).

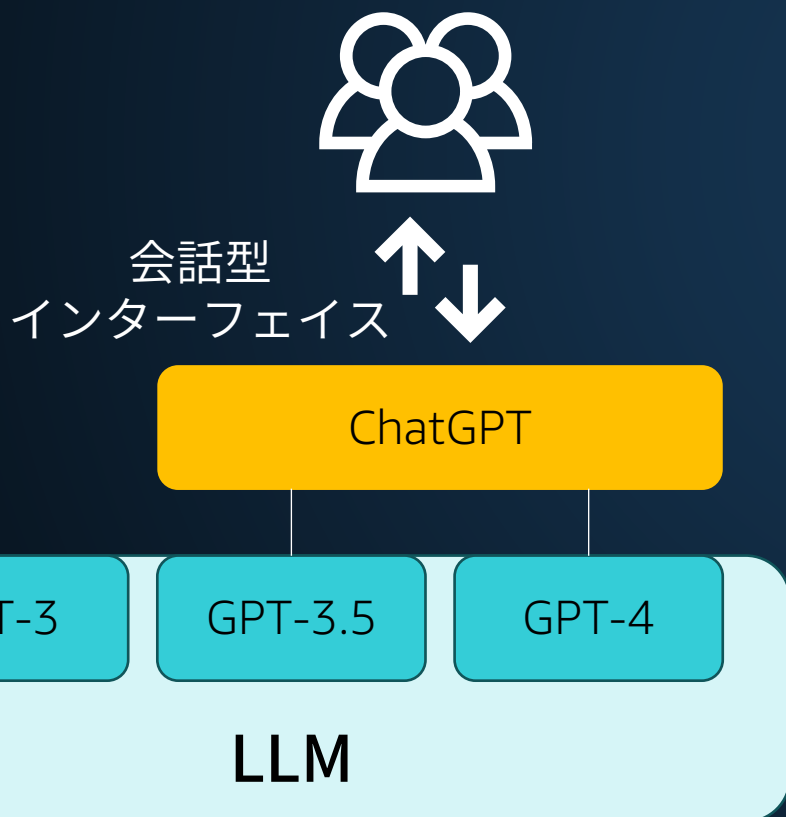
※4 “rinna、日本語に特化した36億パラメータのGPT言語モデルを公開”. rinna ニュース.  
2020-05-17. <https://rinna.co.jp/news/2023/05/20230507.html>, (参照 2023-08-10).

※5 “36億パラメータの日本語言語モデルを公開しました”. LINE Engineering.  
2020-08-14. <https://engineering.linecorp.com/ja/blog/3.6-billion-parameter-japanese-language-model>, (参照 2023-08-10).

※6 “100億パラメータサイズ・日英2ヶ国語対応の大規模言語モデル“Weblab-10B”をオープンソースで公開しました.”. 松尾研究所 プレスリリース.  
2020-08-18. <https://weblab.t.u-tokyo.ac.jp/100億パラメータサイズ・日英2ヶ国語対応の大規模/>, (参照 2023-08-10).

# LLM の一実装 「ChatGPT」

- LLM の直接実装では「文章の続きを書く」ことが可能だが、ニーズが少なかった
- ChatGPT が注目された背景は「会話型インターフェイス」にしたことで、誰でも使える = LLM 利用の民主化を実現



## ChatGPTが高度な会話に対応できるようになった理由

ChatGPTはデータセットの品質改善や強化学習を取り入れることで従来のモデルの課題を克服し、まるで人間と会話しているかのような自然な文章を生成できるようになった

### 従来のモデルの課題

- アラインメント問題
  - モデルの生成文が人間の好みと一致 (=アライン) していない
  - ネット上から収集された大規模なテキストデータを学習に用いているので、**信頼性の低いテキストや有害なテキストを生成する恐れがある**
- プロンプトの表現に敏感すぎる
  - **表現を少し砕けたものになると、モデルは意味理解が来ず、筋違いな回答を生成することがある**

### ChatGPT

- InstructGPTという先行モデルで用いられた「人間のフィードバックに基づいた強化学習」の考え方をChatGPTでも採用
  - 文の良さを評価してくれる報酬モデルを新たに学習させて、報酬が最大化するように強化学習を実施する
  - **有害なテキストを生成しないようにモデルを微調整**
- 会話データを用いてモデルのファインチューニングを行う
  - アノテーターが擬似的な「ユーザーと AIの会話」のデータを作成し、学習に活用
  - **会話ならではの砕けた表現などでも正しく意味理解し、適切な回答の生成が可能**

# Update since 8/20:

情報通信研究機構（NICT）が1790億パラメーターの大規模言語モデル（LLM）の開発を進めている。高品質な日本語データを学習させており、日本社会の常識を備えたモデルになると期待される。

小寺貴之. "1790億パラメーターに挑戦...情通機構「大規模言語モデル」日本語を猛特訓!". ニューススイッチ by 日刊工業新聞. 2023-08-21. <https://newswitch.jp/p/38168>, (参照 2023-08-21).

日本語のダジャレ「おでんの予約は？ お電話で！」がなぜ面白いかを英語で説明できることも紹介し、高い日本語能力をアピールしました。

Junichi Niino. "Google、大規模言語モデル「PaLM 2」が日本語対応したと発表。ダジャレを理解する能力も". Publickey. 2023-08-23. [https://www.publickey1.jp/blog/23/googlepalm\\_2.html](https://www.publickey1.jp/blog/23/googlepalm_2.html), (参照 2023-08-23).

Early tests have shown a fine-tuned version of GPT-3.5 Turbo can match, or even outperform, base GPT-4-level capabilities on certain narrow tasks.

(GPT-3.5 Turbo を対象に、ファインチューニングをサポート)

Andrew Peng, Michael Wu, John Allard, Logan Kilpatrick, Steven Heide. "GPT-3.5 Turbo fine-tuning and API updates". OpenAI blog. 2023-08-22. <https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates>, (参照 2023-08-24).

「LLM は、何でも知ってるな」



“


何でもは知らないわよ。  
知ってることだけ

— 〈物語〉シリーズ 羽川翼

# LLM を使いこなす「プロンプトエンジニアリング」

## LLM は「LLM が知っている事を基に予測を生成する」 知っていることを引き出す手法が、プロンプトエンジニアリング


DE 結婚式には何を持っていけば良いか？

 結婚式に招待された場合、持って行くべきものや心がけるべきことはいくつかあります。以下は一般的なものを挙げてみました。

1. **祝儀（しゅうぎ）**: 結婚式には、祝儀としてお金を包んで持参します。金額は、関係の深さや地域、文化により異なりますが、一般的な額や慣習を参考にすると良いでしょう。また、祝儀袋には自分の名前を記入します。
2. **招待状**: 結婚式や披露宴の場所、時間などの詳細が書かれているため、持っていくと確認に便利です。
3. **正装**: 結婚式は非常にフォーマルな場なので、ドレスコードを守って適切な服装を選びましょう。
4. **小物**: 手鏡、ティッシュ、メイクアップ用品など、小物を持って行くと便利です。
5. **お祝いのカードやメッセージカード**: カードには、新郎新婦へのお祝いの言葉を書くといいでしょう。
6. **携帯電話**: 緊急の際や連絡が必要な場合に役立ちますが、式中はマナーモードにしておくことを忘れずに。
7. **足元に合わせた靴**: 特に女性の場合、ハイヒールなど長時間履いていても痛くならないものを選ぶといいでしょう。
8. **披露宴の際の飲食に関する注意**: アレルギーや宗教上の制約などがある場合、事前に新郎新婦や結婚式場に伝えておくこととスムーズです。

参加する結婚式や披露宴の形式、招待された側の立場などによって、持参するものは異なる場合がありますので、事前に確認することをおすすめします。

DE あなたは冠婚葬祭プランナーです。  
10年ぶりに会う友人の結婚式に招待されました。友人との思い出を振り返りつつ、祝福を伝えることができる企画や持ち物を提案を3つ挙げてください。

 冠婚葬祭プランナーとして、10年ぶりに再会する友人の結婚式に招待された際に特別なサプライズやアイデアを考えるのは楽しいものです。以下に、友人との思い出を振り返りながら、祝福を伝えるための3つの提案をします。

### 1. 思い出の写真アルバム or スライドショー

- 10年という長い間の思い出や、その前の友人との共通の思い出をまとめて写真アルバムやデジタルスライドショーを制作します。その中には、学生時代の写真、旅行の写真、ランダムな日常の一幕、そして手紙やメモなどのスキャンも加えることができます。結婚式の披露宴の時に、このアルバムやスライドショーをサプライズとして上映することで、新郎新婦はもちろん、ゲストもその感動的な瞬間を共有することができます。

### 1. 手紙のタイムカプセル

- 結婚式の前に、友人や家族、共通の友人たちに未来の新郎新婦への手紙を書いてもらいます。それをタイムカプセルとして特別な箱や容器に入れて保存します。新郎新婦には、特定の記念日（例：結婚5周年）にそのタイムカプセルを開けて読むように伝えます。これにより、未来の自分たちへのメッセージやアドバイス、祝福を受け取ることができます。

# プロンプトエンジニアリングを学ぶ



## Prompt Engineering Guide

<https://www.promptingguide.ai/>

日本語: <https://www.promptingguide.ai/jp>

Prompt Engineering Guide

Prompt Engineering Course Services About Search document

Prompt Engineering

Introduction

- LLM Settings
- Basics of Prompting
- Prompt Elements
- General Tips for Designing Prompts
- Examples of Prompts

Techniques

- Zero-shot Prompting
- Few-shot Prompting
- Chain-of-Thought Prompting
- Self-Consistency
- Generate Knowledge Prompting
- Tree of Thoughts

### Prompt Engineering Guide

Prompt engineering is a relatively new discipline for developing and optimizing prompts to efficiently use language models (LMs) for a wide variety of applications and research topics. Prompt engineering skills help to better understand the capabilities and limitations of large language models (LLMs).

Researchers use prompt engineering to improve the capacity of LLMs on a wide range of common and complex tasks such as question answering and arithmetic reasoning. Developers use prompt engineering to design robust and effective prompting techniques that interface with LLMs and other tools.

Prompt engineering is not just about designing and developing prompts. It encompasses a wide range of skills and techniques that are useful for interacting and developing with LLMs. It's an important skill to interface, build with, and understand capabilities of LLMs. You can use prompt engineering to improve safety of LLMs and build new capabilities like augmenting LLMs with domain knowledge and external tools.

Motivated by the high interest in developing with LLMs, we have created this new prompt engineering guide that contains all the latest papers, learning guides, models, lectures, references, new LLM capabilities, and tools related to prompt engineering.

## Best practices for prompt engineering with OpenAI API

<https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api>

### Best practices for prompt engineering with OpenAI API

How to give clear and effective instructions to GPT-3 and Codex

Written by Jessica Shieh  
Updated over a week ago

If you're just getting started with OpenAI API, we recommend reading the [Introduction](#) and [Quickstart](#) tutorials first.

#### How prompt engineering works

Due to the way the instruction-following [models](#) are trained or the data they are trained on, there are specific prompt formats that work particularly well and align better with the tasks at hand. Below we present a number of prompt formats we find work reliably well, but feel free to explore different formats, which may fit your task best.

プロンプトエンジニアリング(PE)自体の基礎から、LLM 全般に対するテクニックを解説  
※英語版が最新

OpenAI が提供する API を使いこなす 8 つの Tips だが、他の LLM 向けにも通ずる内容

# ChatGPT のユースケース

## ChatGPTができること | 自然言語におけるユースケース



	自然言語	ユースケース		
オリジナルが 人間のアウトプット  ↑  ↓  オリジナルが AIのアウトプット	文章の 添削、校正	文法ミスだけでなく 内容の改善点まで 指摘		
	文章や 概念の要約	文字数を指定して 要約		
	壁打ち、 ブレインストーミング	「～～」という考えの うち抜けている点を 指摘	AIと人間で会話しな がらアイディエーション	
	リサーチ、 論点の洗い出し	「XX」について要点 だけ説明	ディベート風に論点 を洗い出し	表形式で比較表を 作成
	アイデアの提案	SEO対策したタイト ルやキーワード形成	広告文の自動生成	何かにな りきって答える
結婚式のスピーチの アウトライン作成		人気の出るブログの 内容案の提案	例題や乱数などの 生成	
		⋮	⋮	⋮

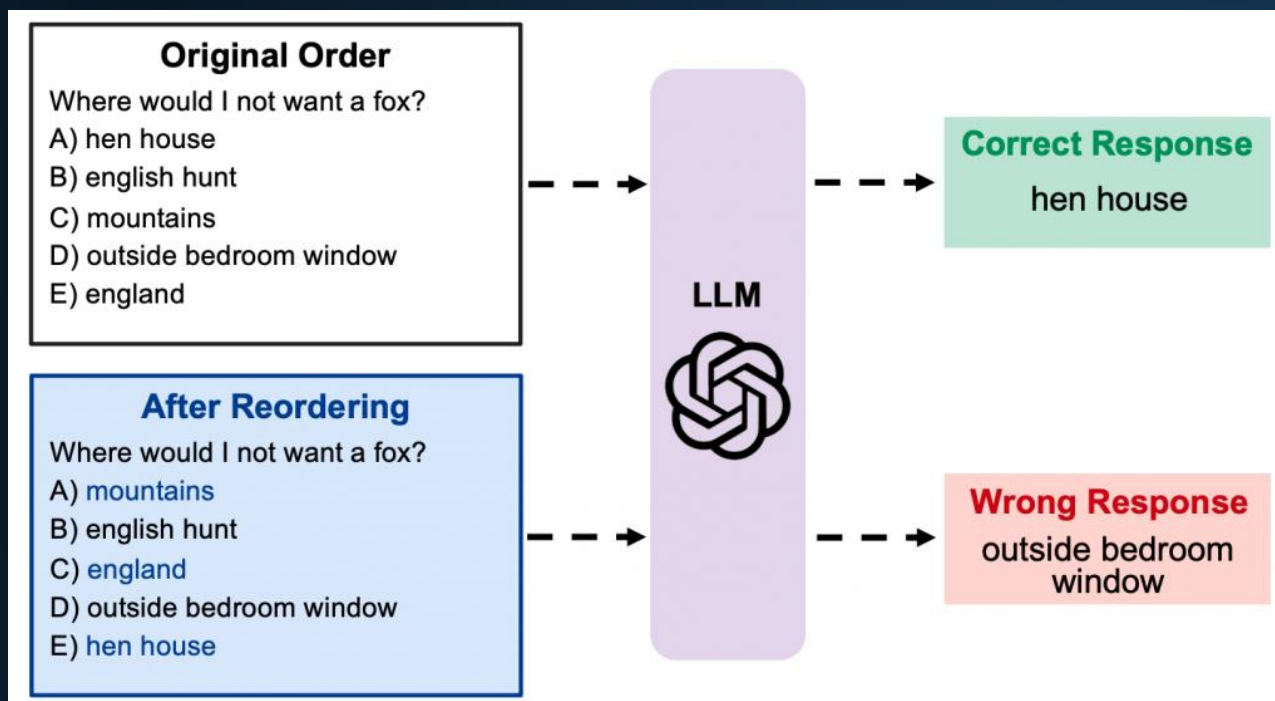
従来と同じタスクだが、  
精度が大幅に向上

ChatGPTの卓越したコ  
ミュニケーション精度と  
圧倒的な知識量(デー  
タ量, パラメータ数)か  
ら、以前では決してでき  
なかつたことまでできる  
ようになっている。



# Update since 8/20: プロンプトエンジニアリングは発展期

GPT-4に選択肢を与えるとき、順序を入れ替えるだけで性能に大きな変化があることが明らかに



“

選択肢の順序が変更されることで、モデルの性能に13%から75%もの大きな変動が生じたという点

丁寧な精査も必要だがコンピューターの原則

## GIGO

Garbage In, Garbage Out

を意識することで、より良い結果が得られる



# 70分間の過ごし方 (アジェンダ)

- GenAI の振り返りと、プロンプトエンジニアリング  
《 QA 》
- IoT と GenAI の関係性  
《 QA 》
- ハードウェアからの GenAI 利用の実態
- GenAI 利用と通信の活用 — 適した通信技術
- このあと、どうしたらよいのか？  
— GenAI との付き合い方  
《 QA 》

# IoTとは？

センサー/デバイス  
“モノ”

ネットワーク

クラウド



現場をデジタル化

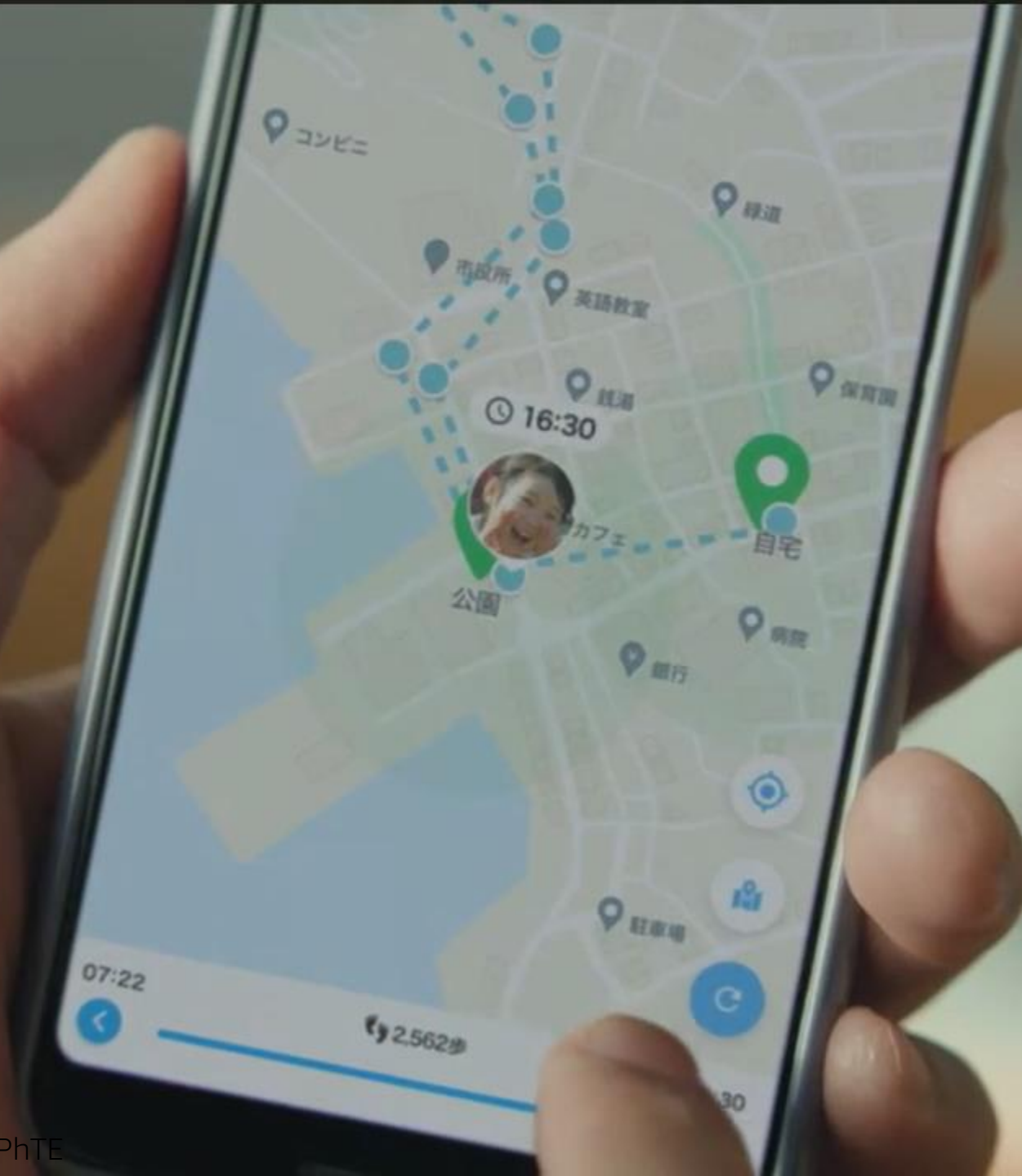
現場とクラウドをつなげる

デジタルデータの活用

## モノやコトをデジタル化

人手に頼らずデータを集める、活かす、現場を動かす





画面はイメージです。

## 導入事例

# 株式会社ミクシィ

小さなお子さまをもつ親御さんのために、子供が携帯しやすい「みてね」の子ども向けGPSを開発。

高精度な位置情報とAI学習で自動で出発・到着をお知らせ。お子さまの1日の歩数や移動履歴も。



利用したSORACOMサービス：  
SORACOM Air, SORACOM Beam



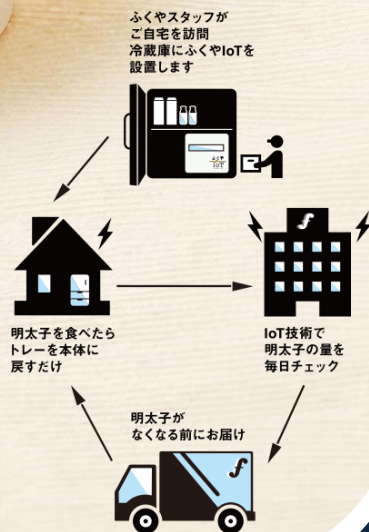
協力パートナー：JENESIS株式会社



# お客様事例：ふくや様



味の明太子  
博多中洲 **ふくや**



## 専用機器で毎日明太子の 消費量を送信、 なくなる前に自動でお届け

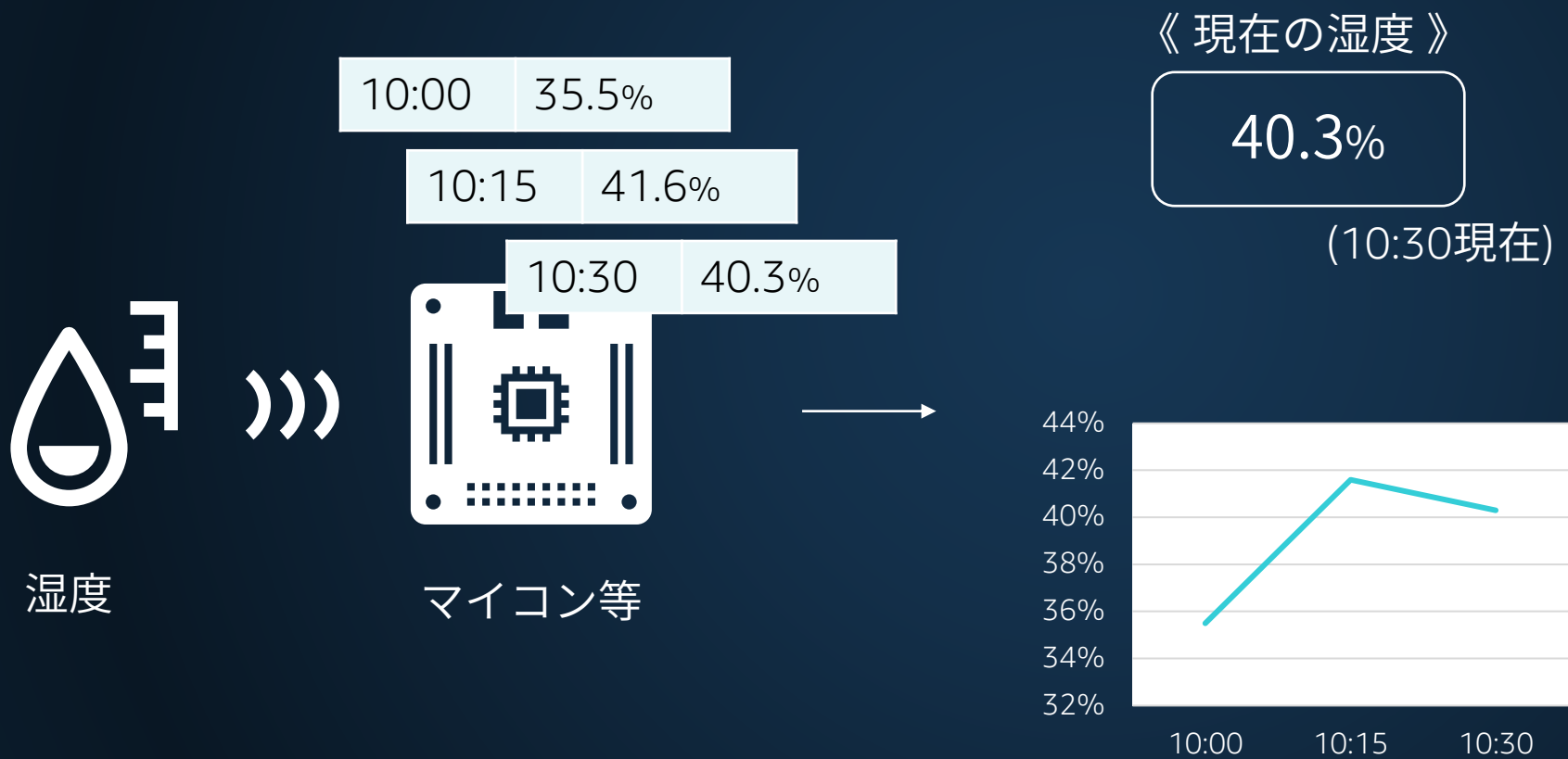
※福岡限定でサービス開始

## SORACOM Harvest や SORACOM Lagoon で サーバの準備なくデータの 可視化・通知を実現

協力パートナー：メカトラックス



# IoT が産み出すデータは2系統 「状態」と「時系列」

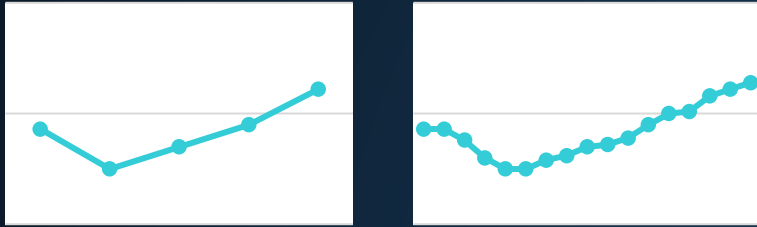


今の湿度は？  
= 状態

過去の記録は？  
= 時系列

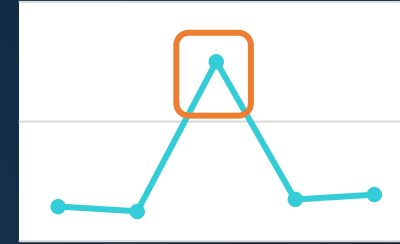
# 時系列データの扱いで考慮すべきこと

## 記録頻度とデータ量



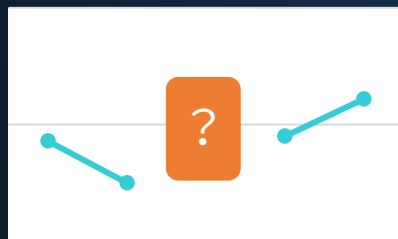
特徴が判別可能となる  
記録頻度(解像度)と必要なデータ領域

## ノイズと外れ値



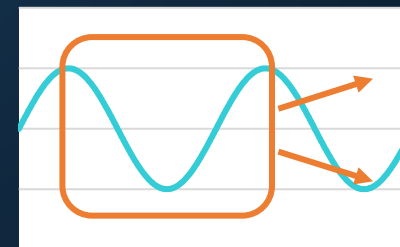
センサーの誤作動などで生じた  
不正確・不要な値の処理戦略

## 欠損値



デバイスの故障や通信途絶などで生じた  
欠損値の補完戦略

## 周期性とパターン



短期的(季節変動)と長期的(トレンド)の  
パターン認識とパターン外の識別

# 時系列データから「得られる情報」と、 私たちが「知りたい事」は異なる

過去

現在

未来



35.5%

41.6%

40.3%



時系列データ

本当に知りたい事

# 時系列データを基に ChatGPT に 分析や未来予測をしてもらう

過去

現在

未来



35.5%

41.6%

40.3%

時系列データ



ChatGPT





# 分析や未来予測に ChatGPT を用いる

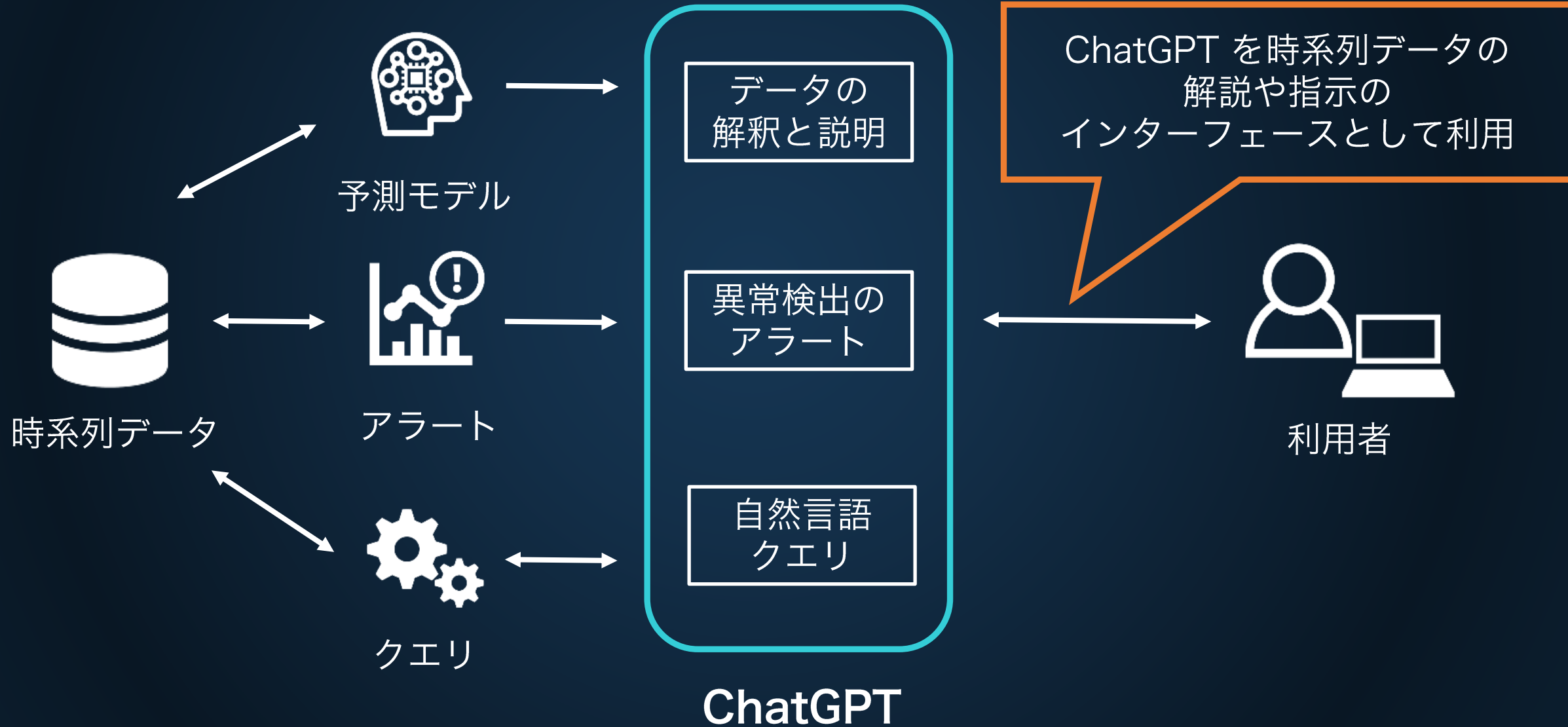
## 利点

- 事前学習が不要
- 理由を回答させることで、妥当性を(人間が)検証可能

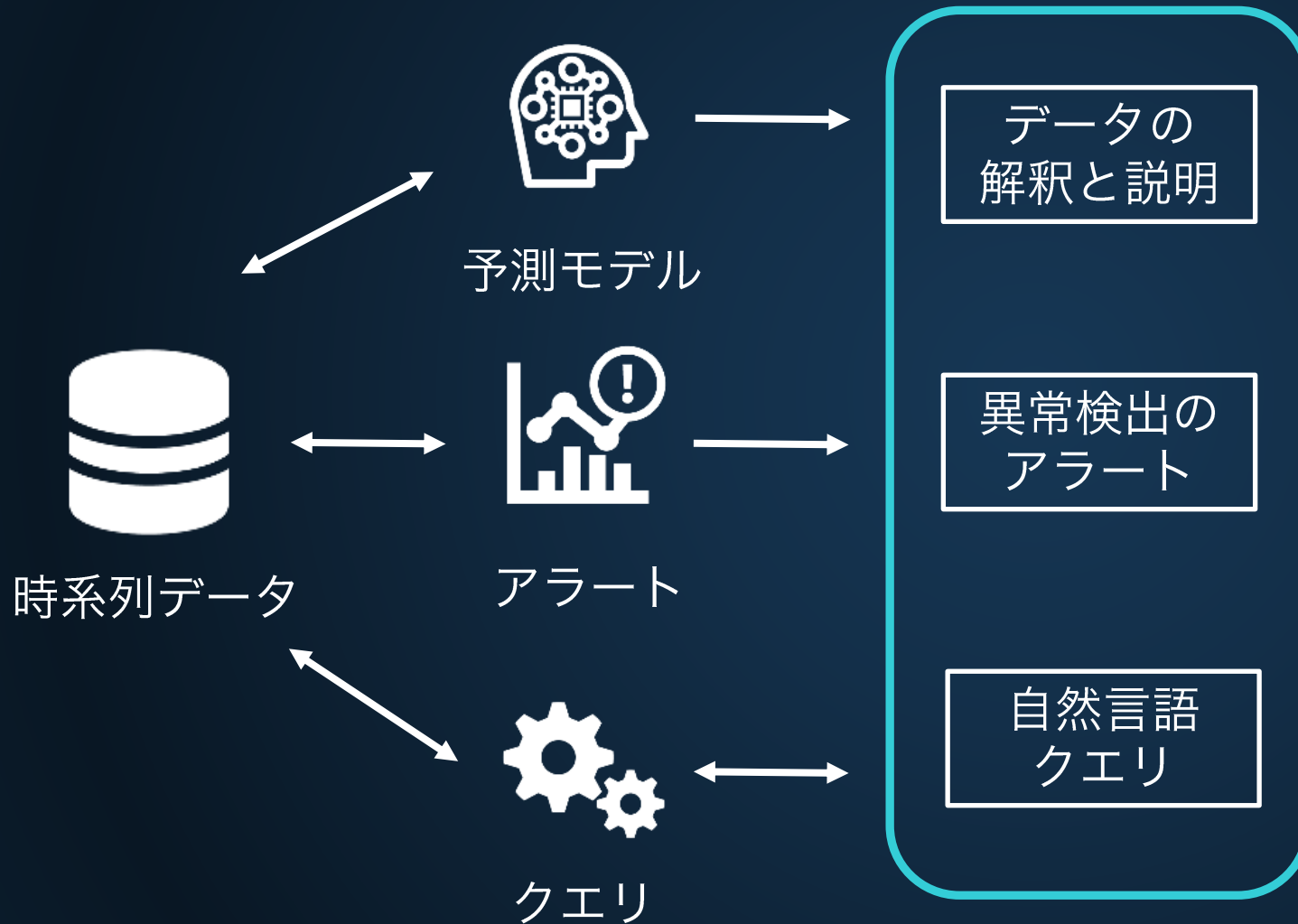
## 気を付けたい点

- 与えるデータサイズの調整
- 冪等性が無い (同じデータから別の予測が導き出される可能性がある)

# 時系列データ と ChatGPT



# 時系列データ と ChatGPT



ChatGPT

- 実績と要因の洞察
- その理由や検証  
例) 売上と相関がある数値の発見
- 周期やトレンドを考慮したアラート
- 欠損自体の認識
- 自然言語によるクエリ
- クエリ言語 (SQL、GraphQL) 自体の生成

# ChatGPT API 利用時の考慮ポイント

## 1. トークン制限

- API リクエストは、入力と出力の合計で利用可能なトークン数に制限があります。これは、非常に長いテキストを処理することができないことを意味します。例えば、数千語の長い文章を一度に生成することはできません。

## 2. データのプライバシー

- API を通じて送信されるすべてのデータは、OpenAI のプライバシーポリシーに従います。したがって、機密情報や個人を特定可能な情報を含むテキストを与えることは避けるべきです。

## 3. 知識の更新

- ChatGPT の知識はある時点の訓練データに基づいており、訓練データの収集時点以降の情報については知識を持っていません。最新の情報を利用する場合は、ユーザー自身で プラグインによるインプット や、必要な情報での Fine-tuning が必要です。



## 4. 情報の正確性

- ChatGPT は時々、正確性が不明な情報を生成する可能性があります。これは、モデルが訓練データから学習したパターンに基づいてテキストを生成するためです。一方で、生成した内容の理由や経緯を一緒に出力 できるので、そこを含めて内容を確認することで、検証の助けとなります。

## 5. 数値計算が苦手

- ChatGPT は数値計算を行う能力がありますが、複雑な計算や特定の種類の計算では、計算を上手く行えないことがあります。

# トークンのカウント

## Tokenizer

The GPT family of models process text using **tokens**, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

You can use the tool below to understand how a piece of text would be tokenized by the API, and the total count of tokens in that piece of text.

GPT-3 Codex

```
'''CSV
time,temp,humi
2023-05-29 08:48:45,27.3,62.8
2023-05-29 08:47:45,27.4,62.6
2023-05-29 08:46:44,27.3,61.3
2023-05-29 08:45:43,27.0,66.7
2023-05-29 08:44:43,27.2,61.6
2023-05-29 08:43:43,27.6,61.5
```

Clear Show example

Tokens 1,214  
Characters 1825

```
'''CSV
time,temp,humi
2023-05-29 08:48:45,27.3,62.8
2023-05-29 08:47:45,27.4,62.6
2023-05-29 08:46:44,27.3,61.3
2023-05-29 08:45:43,27.0,66.7
2023-05-29 08:44:43,27.2,61.6
2023-05-29 08:43:43,27.6,61.5
2023-05-29 08:42:43,27.1,61.3
2023-05-29 08:41:43,27.8,60.8
2023-05-29 08:40:43,27.7,63.2
2023-05-29 08:39:43,27.9,61.3
2023-05-29 08:38:43,27.9,62.4
2023-05-29 08:37:48,27.7,61.1
2023-05-29 08:36:44,27.9,61.6
2023-05-29 08:35:42,27.0,61.7
2023-05-29 08:34:48,27.9,64.0
2023-05-29 08:33:43,27.6,62.1
2023-05-29 08:32:43,27.9,61.7
TEXT TOKEN IDS :43,27.8,57.6
```

1 時間分の編集済みデータでも多くの  
トークンを消費します

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

A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly 3/4 of a word (so 100 tokens ~ 75 words).

# 70分間の過ごし方 (アジェンダ)

- GenAI の振り返りと、プロンプトエンジニアリング  
《 QA 》
- IoT と GenAI の関係性  
《 QA 》
- ハードウェアからの GenAI 利用の実態
- GenAI 利用と通信の活用 — 適した通信技術
- このあと、どうしたらよいのか？  
— GenAI との付き合い方  
《 QA 》

# 組み込み系ハードウェアにおける LLM の活用は？



開発支援に  
LLM を活かす



LLM の機能を  
組み込む



# ハードウェア開発支援に LLM を活かす

## ChatGPT のユースケース

ChatGPTができること | 自然言語におけるユースケース

	自然言語	ユースケース			
オリジナルが 人間のアウトプット	文章の 添削、校正	文法ミスだけでなく 内容の改善点まで 指摘			
	文章や 概念の要約	文字数を指定して 要約			
	壁打ち、 ブレインストーミング	「〜」という考えの うち抜けている点を 指摘	AIと人間で会話しな がらアイディエーション		
	リサーチ、 論点の洗い出し	「XX」について要点 だけ説明	ディベート風に論点 を洗い出し	表形式で比較表を 作成	
	オリジナルが AIのアウトプット	アイデアの提案	SEO対策したタイト ルやキーワード形成	広告文の自動生成	何かに なりきって答える
結婚式のスピーチの アウトライン作成			人気が出るブログの 内容案の提案	例題や乱数などの 生成	
		⋮	⋮	⋮	

従来と同じタスクだが、  
精度が大幅に向上

ChatGPTの卓越したコ  
ミュニケーション精度と  
圧倒的な知識量(デー  
タ量、パラメータ数)か  
ら、以前では決してでき  
なかつたことまでできる  
ようになっている。

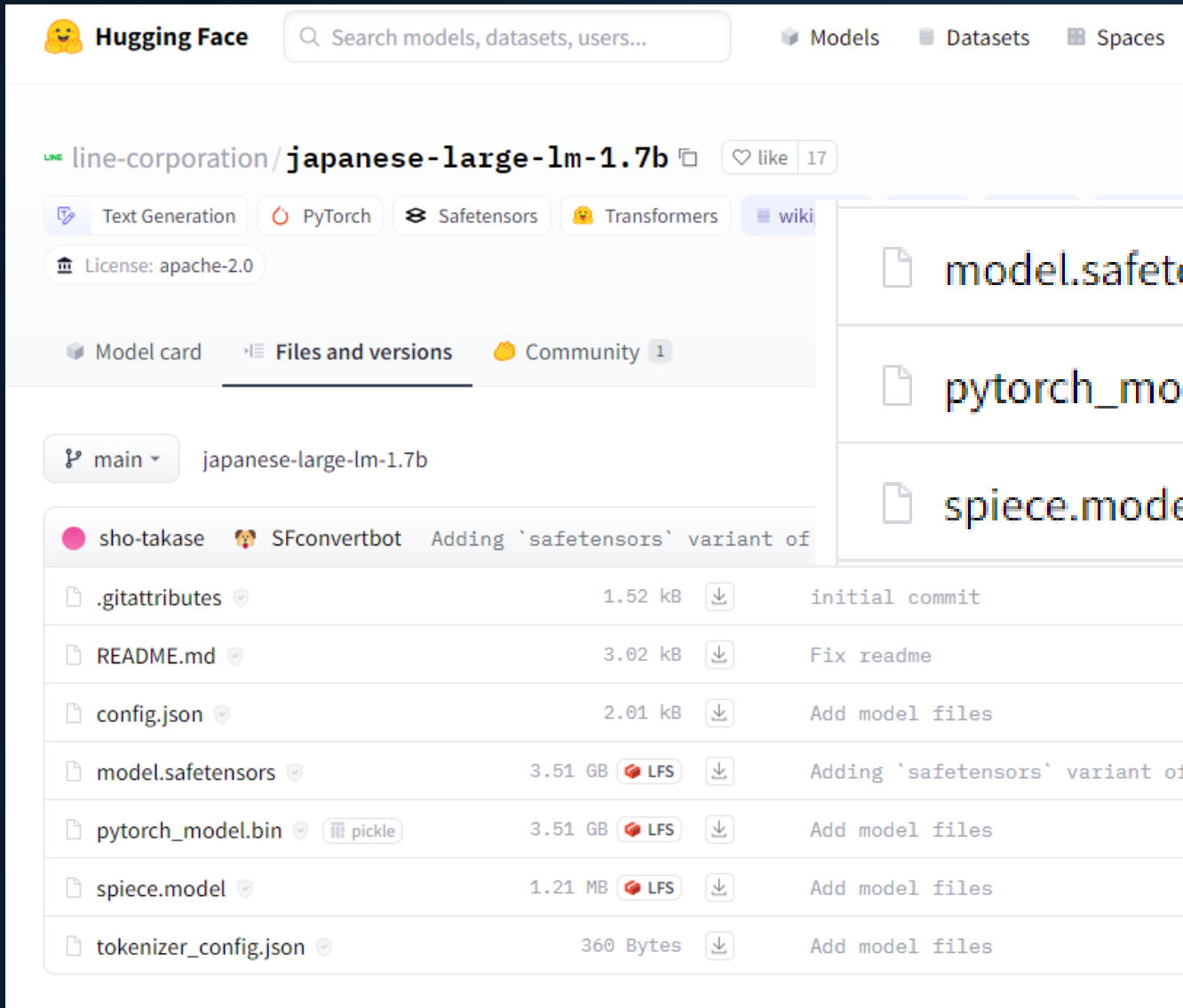
参考: ChatGPT (OpenAI提供)

©MATSUO LAB, THE UNIVERSITY OF TOKYO

20

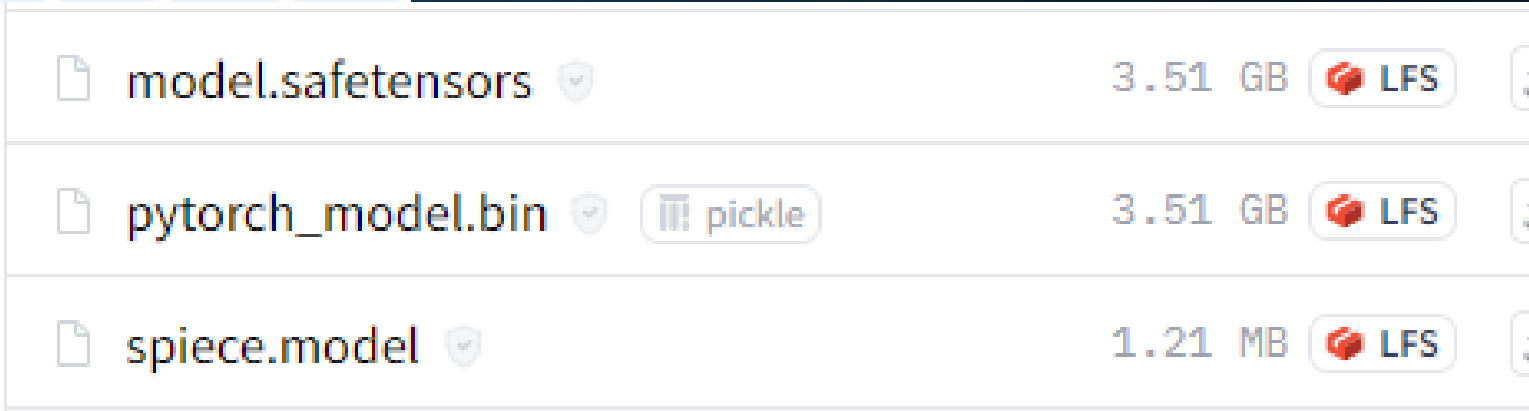
- コードや回路の設計、生成
- デバッグの計画立案
- ドキュメント生成やレビュー

# LLM のファイルサイズは数 GB 以上



The screenshot shows the Hugging Face interface for the model `line-corporation/japanese-large-lm-1.7b`. The page includes a search bar, navigation tabs for Models, Datasets, and Spaces, and a list of tags such as Text Generation, PyTorch, Safetensors, Transformers, and wiki. The license is listed as apache-2.0. The main content area shows the file list for the `main` branch of the repository. The files and their sizes are:

File Name	Size	Format	Commit Message
<code>.gitattributes</code>	1.52 kB		initial commit
<code>README.md</code>	3.02 kB		Fix readme
<code>config.json</code>	2.01 kB		Add model files
<code>model.safetensors</code>	3.51 GB	LFS	Adding `safetensors` variant of
<code>pytorch_model.bin</code>	3.51 GB	LFS, pickle	Add model files
<code>spiece.model</code>	1.21 MB	LFS	Add model files
<code>tokenizer_config.json</code>	360 Bytes		Add model files



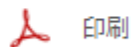
<code>model.safetensors</code>	3.51 GB	LFS
<code>pytorch_model.bin</code>	3.51 GB	LFS, pickle
<code>spiece.model</code>	1.21 MB	LFS

# Raspberry Pi で動かす試みはされている

## ラズパイで動く大規模言語モデルがGitHubで公開 性能は“GPT-3相当”、Metaの「LLaMA」派生

🕒 2023年03月18日 14時04分 公開

[松浦立樹, ITmedia]



印刷



見る



Share



247

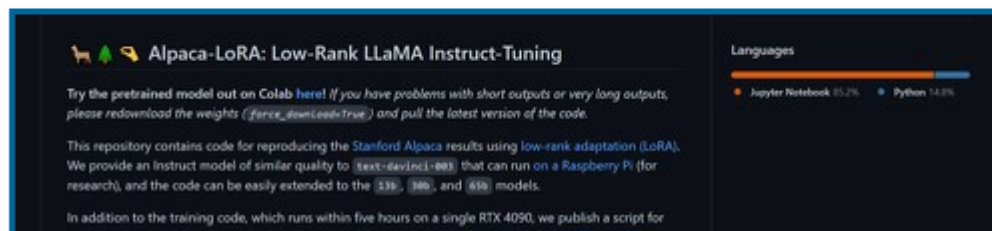


1

PR エンジニア必見の工数見積もり術！無理ゲー進行を脱する

PR 重さ約1.08kgで最大23.5時間駆動!! 高性能モバイルPC「MousePro G4」

Raspberry Pi上で動かせる、GPT-3相当の大規模言語モデル（LLM）「Alpaca LoRA」がGitHub上で公開された。米MetaのLLM「LLaMA」の派生モデル「Stanford Alpaca」を再現できるという。



LLaMAの7Bモデルを「Raspberry Pi 4」で動かせるか検証したところ、超低速ながら稼働したと報告。

## 1トークン当たり10秒

# LLM のローカル実装は「見極め」

## ChatGPT のユースケース

ChatGPTができること   自然言語におけるユースケース	
自然言語	ユースケース
オリジナルが人間のアウトプット	
文章の添削、校正	文法ミスだけでなく内容の改善点まで指摘
文章や概念の要約	文字数を指定して要約
壁打ち、ブレインストーミング	「〜」という考えのうち抜けている点を指摘 AIと人間で会話しながらアイディエーション
リサーチ、論点の洗い出し	「XX」について要点だけ説明 ディベート風に論点を洗い出し 表形式で比較表を作成
オリジナルがAIのアウトプット	アイデアの提案 SEO対策したタイトルやキーワード形成 広告文の自動生成 何かになりきって答える 結婚式のスピーチのアウトライン作成 人気の出るブログの内容案の提案 例題や乱数などの生成

従来と同じタスクだが、精度が大幅に向上

ChatGPTの卓越したコミュニケーション精度と圧倒的な知識量(データ量, パラメータ数)から、以前では決してできなかったことまでできるようになっている。

参考: ChatGPT (OpenAI提供)

©MATSUO LAB, THE UNIVERSITY OF TOKYO

20

出所: 松尾豊. "AIの進化と日本の戦略". 自民党AIの進化と実装に関するプロジェクトチーム(第2回). 2020-02-17. [https://note.com/akihisa\\_shiozaki/n/n4c126c27fd3d](https://note.com/akihisa_shiozaki/n/n4c126c27fd3d), (参照 2023-08-10).

- 閉じたハードウェア上で、何を実現したいのか？
- 労力が見合うのか？

# 参考: LLM のローカル動作を試みるなら

Docswell ログイン 新規登録

## オープンなLLMをローカルで動かす

TechFeed Experts Night #23  
2023/07/19



<https://techfeed.io/events/techfeed-experts-night-23>



からあげ @karaage0703

1 / 51

目次 < > 全画面 シェア



# ハードウェアからの GenAI 利用

- LLM の本質は「知っている事だけ」モデルサイズが価値とも言える
- ユースケースを限定「目的特化」とすれば、価値が出せる可能性もある
- GenAI 自体に特定課題向け教師データを生成させ、マイコン動作可能な ML モデル(Not LLM モデル)を作る手段も考えられる

注意: GenAI による教師データ生成が規約で NG な場合もあるため  
許諾範囲は要チェック

それでも LLM を使いたい？



導入事例

# ポケットーク株式会社 「POCKETALK<sup>®</sup>」

外国語での双方向コミュニケーションを実現するAI通訳機。

グローバル通信搭載で、電源を入れれば、世界中ですぐに利用可能。

\* 対応言語の詳細は製品ウェブサイトをご確認ください

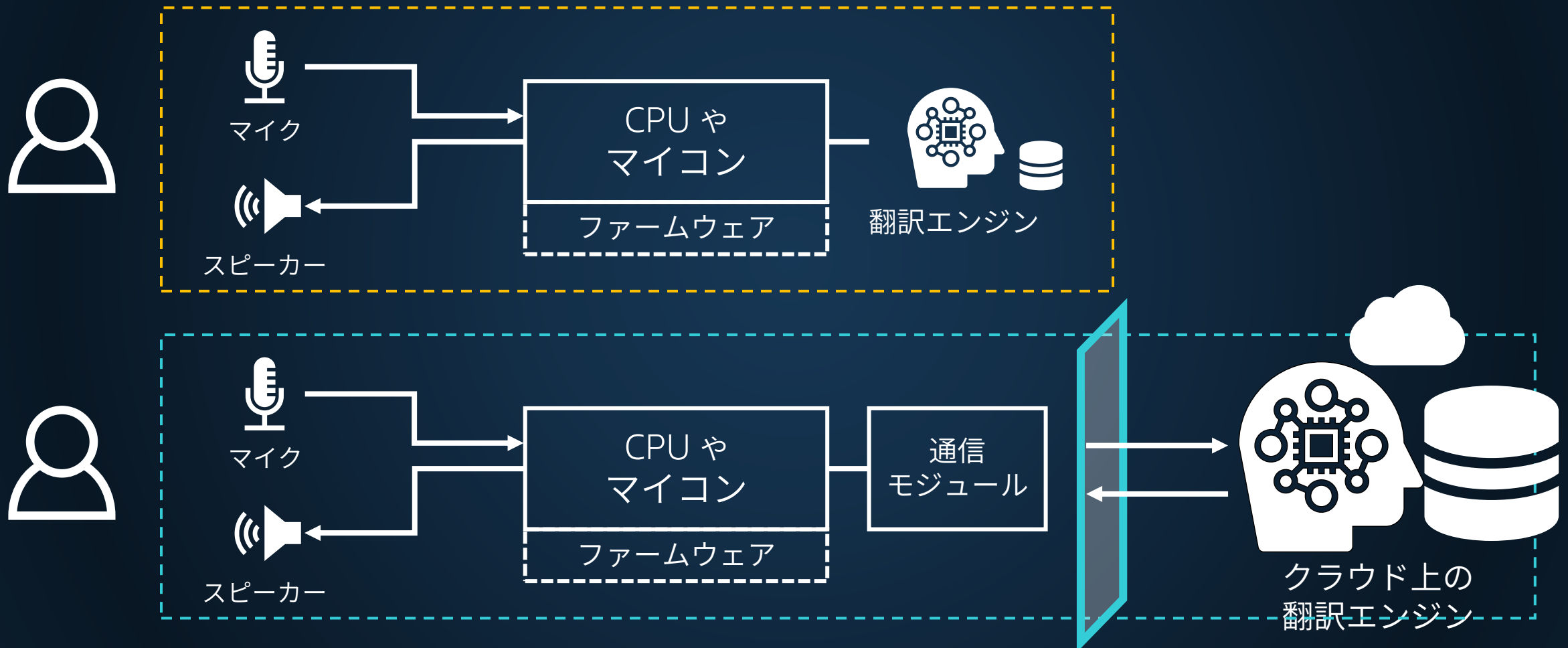


利用したSORACOMサービス：SORACOM Air  
(plan01s / planX1) チップ型SIM (eSIM) 搭載



協力パートナー：JENESIS株式会社

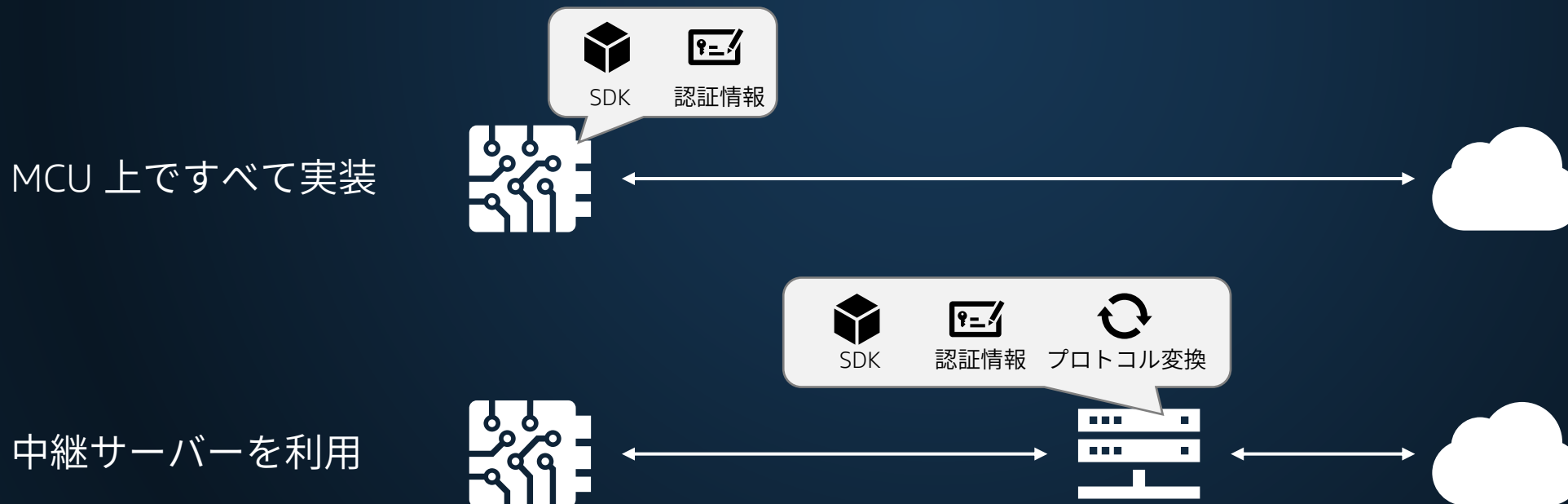
# 通信の利用でハードウェアの機能を増やす、 価値を上げる



クラウドと一体化して「製品」

# GenAI を利用するための通信

- 結局のところ、クラウド API プログラミング
- JSON での送受信、IP 通信が不可欠  
Wi-Fi、有線、セルラー(LTE / 5G / LTE-M) 等が候補

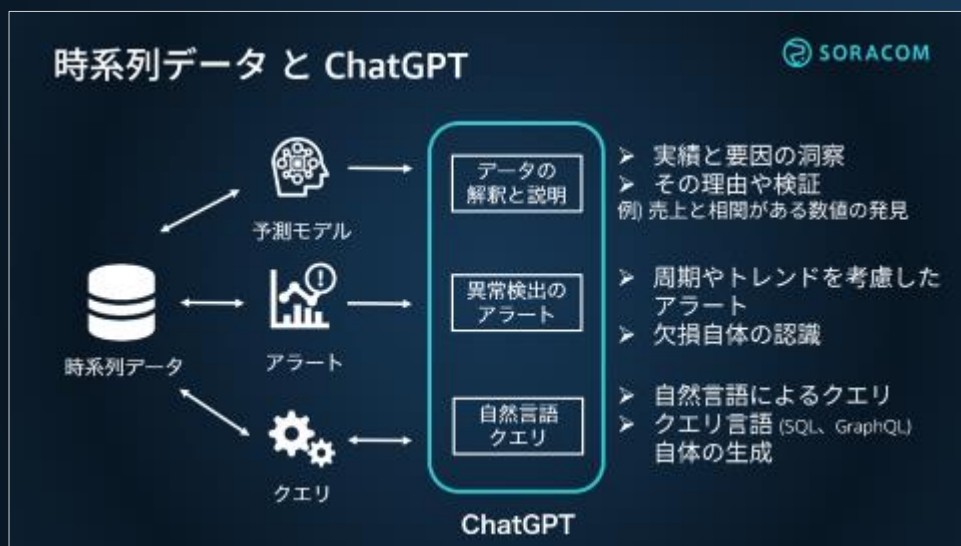


例: [nginx](#) を MQTT ↔ MQTTS 変換機にして AWS IoT Core へ接続

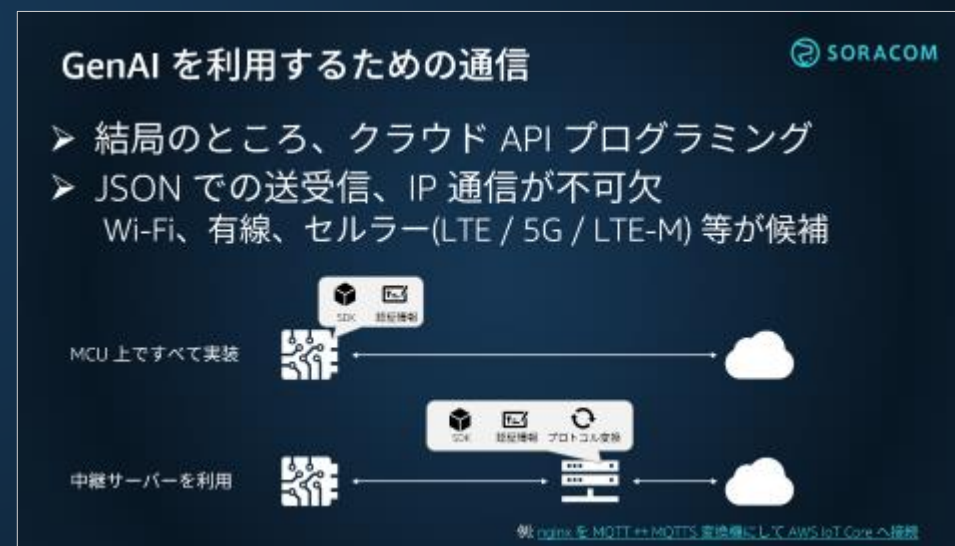


# Conclusion

- GenAI は、生成モデルに着目した ML/DL の一形態
- LLM ができることは「次の単語を予測して生成」
- プロンプトエンジニアリングで LLM の活用が大きく変わる



IoT における LLM 活用の 1 つは  
時系列データの分析



ハードウェアからの LLM 利用は  
通信を用いるアーキテクチャが容易

# GenAI との、これからの付き合い方

## 共通して知っておきたい事

### プロンプトエンジニアリング

- 資料内「[プロンプトエンジニアリングを学ぶ](#)」

### システム間連携

(例: API レスポンスを JSON に固定したい)

- [Function calling](#) (OpenAI)

### 自分好みにする

(例: FAQ を基に回答するようにさせたい)

- Fine-tuning  
(e.g., GPT-3.5 Turbo + fine-tuning)

## GenAI のこれから

- AI について「作るか」「使うか」が明確になり“使う”環境が整ってきた
- 取り扱いデータのマルチモーダル(文字/画像/音声/動画を同様に扱う)化が進んでいる
- 汎用 LLM と目的特化 LLM、どちらが主流になるのかは専門家でも意見が分かれている状況

**「開発のお供“copilot”」なのか「製品の1機能」にしたいのか、  
まず私たちのスタンスが大切です**

IoTの「つなぐ」を簡単に

You Create. We Connect.



**SORACOM**