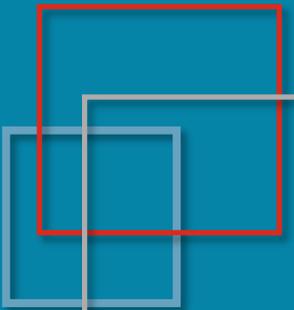


メールとAIシリーズ

プライバシー保護のための次世代技術 ～NLPの未来と大規模言語モデル活用術

平野 善隆 (Vade Japan株式会社)

Nuwan Senevirathne (株式会社クオリティア)



自己紹介

名前 平野 善隆

所属 Vade Japan 株式会社
Principal Messaging Engineer

NLPとの関わり

奈良先端科学技術大学院大学
情報科学研究科 自然言語処理学
形態素解析器ChaSenの開発

趣味 世界の長距離の自転車大会(1,200kmとか、2,000kmとか)
バンド演奏

主な活動 M³AAWG, JPAAWG, 迷惑メール対策推進協議会
Audax Randonneurs Nihonbashi



自己紹介

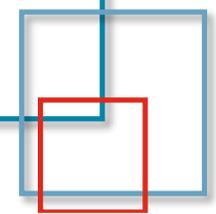
名前 ヌワン セネビラツナ

所属 株式会社クオリティア
AI Engineer

学歴 筑波大学 - Spatial Information Science (MSc)
University of Colombo - Computer Science (BSc)

趣味 Gaming
Singing





もくじ

- 昨年のAIセッションのおさらい
- メール匿名化の手法
- 固有表現抽出について
- 小規模言語モデルを用いた固有表現抽出
- 大規模言語モデルを用いた固有表現抽出
 - プロンプトエンジニアリング
 - LLMアダプター

昨年のJPAAWGのおさらい

メールサービスを面白くするための AIと自然言語処理の基本と応用

- トランスフォーマーモデルの紹介
- メールへの応用
 - ➔ メール本文から「件名」の自動生成

13:00-13:45 AI-3

メールサービスを面白くするためのAIと自然言語処理の基本と応用

デジタル時代において電子メールの重要性は高まっており、AI技術の活用が注目されています。本セッションでは、AIが電子メールの改善とセキュリティ強化にどのように貢献するかを探ります。AIは魅力的なメール件名の提案から、誤送信の防止、正確な配信の実現まで多岐にわたります。特にTransformerと呼ばれるAIモデルは、言語タスクにおいて優れた成果を上げ、電子メールシステムの向上に貢献しています。AI技術が電子メールの効率性と信頼性向上にどのように貢献するかについて議論します。

<講演資料: Nuwan Senevirathne>



Nuwan Senevirathne
株式会社クオリティア

メールの件名自動生成



qualityia.co.jp

インフルエンザ予防接種 詳細

各位

お疲れ様です。人事総務部の[]です。
標記の件、日程が決まりましたのでご連絡いたします。

期間：11月5日(火)～11月29日(金)

- ※毎週水曜日は担当医師1名のみでの営業になるため非常に混雑します。
- 水曜日の受診はなるべく避けてください。
- 上記期間内で必ず受診してください。

受付時間：

AM9: 00～12: 30

PM2: 00～4: 30

- ※毎週水曜日のみ午前中受付開始がAM9: 30～となります。
- ※上記時間内に必ず受付を完了してください。

クリニック名：[]ドクリニック

住所：〒103-[]
東京都中央区日本橋茅場町[]

持ち物：・添付ファイル「インフルエンザ予防接種 予診票」
・(今回のクリニックに掛かったことのある方は)診察券
※こちら忘れると接種できませんので事前に記入し持参してください。

費用：全額当社負担(個別での支払いは不要です。)

その他注意点：卵アレルギーのある方は接種できません。
また、一般の患者さんで込み合う時間帯もあるため、
待ち時間が長くなる場合がありますのでご注意ください。
待合スペースは非常に狭いです。
5名、10名単位で一緒に受診することは避けてください。

添付の「インフルエンザ予防接種 注意点」も必ずご一読をお願いします。

ご質問、ご不明点は人事総務部[]までご連絡くださいませ。

皆様のご協力、何卒よろしくお願い申し上げます。

人事総務部 []

メールの件名自動生成

qualityia.co.jp

インフルエンザ予防接種 詳細

住所: 〒103-0000

各位

お疲れ様です。人事総務部の
標記の件、日程が決まりました。

期間: 11月5日(火)~11月29日(金)

*毎週水曜日は担当医師1名の
水曜日の受診はなるべく避
上記期間内で必ず受診して

受付時間:

AM9: 00~12: 30

PM2: 00~4: 30

*毎週水曜日のみ午前中受付開

*上記時間内に必ず受付を完了

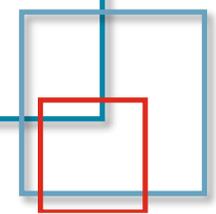
クリニック名:

Candidate subject

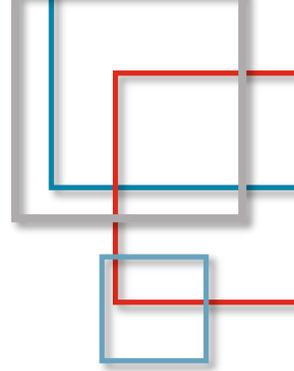
- インフルエンザ予防接種日程のご連絡
- 【人事総務部】インフルエンザ予防接種日程のご
連絡
- インフルエンザ予防接種日程のご連絡(人事総務部)

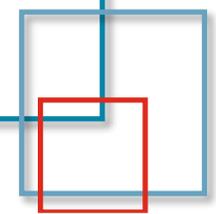
Cancel

OK

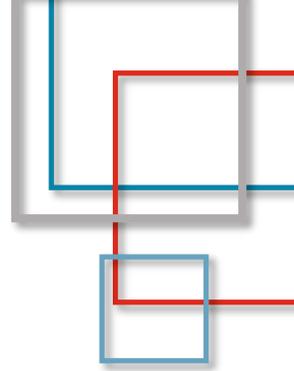


件名生成の課題

- 学習データに企業名や個人名などのデータが含まれる
 - 件名生成時に関係のない企業名などが生成されることがある
- ➔ メールデータの匿名化の必要性
- 



メールアドレスの匿名化の必要性



- 学習するメールアドレスに個人情報等が入らないようにする
- 個人情報を匿名化したメールアドレスを保存する
- 個人情報で検索できないようにする
- 外部のシステムを使用するときに、個人情報を匿名化して渡す
- などなど

個人情報を匿名化する方法



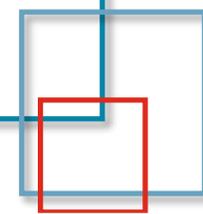
Redaction

My name is John Doe.



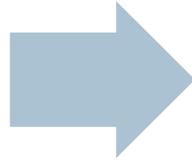
My name is .

- Definition:
 - 個人情報を削除する
- 目的:
 - 機密情報の保護
- 特徴:
 - データが完全に削除される
 - 削除した場所がわかる
- Use Cases:
 - 法的な書類、健康診断の結果、機密情報 など

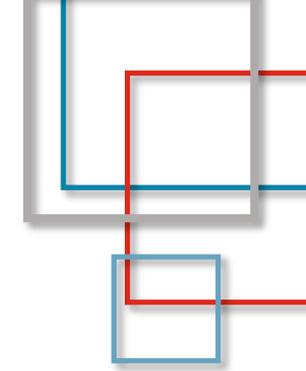


Replacement

John Doe



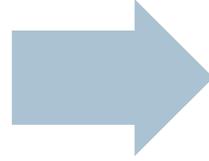
[REDACTED]



- Definition:
 - 個人情報を固定文字列に置き換える
- Purpose:
 - 機密情報をはっきりマスクする
- Key Features:
 - 削除された部分が明確になる
 - 効果的に個人情報を隠せる
- Use Cases:
 - 法的な書類、機密情報など

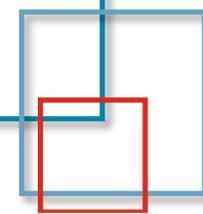
Truncation

John Doe



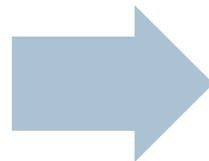
J*** D**

- Definition:
 - 機密情報の一部を隠す
- Purpose:
 - 文脈は残したまま個人情報を保護する
- Key Features:
 - 置き換え文字列を使用 (*やXなど)
 - プライバシーと利便性のバランス
- Use Cases:
 - 氏名、クレジットカード番号、メールアドレス

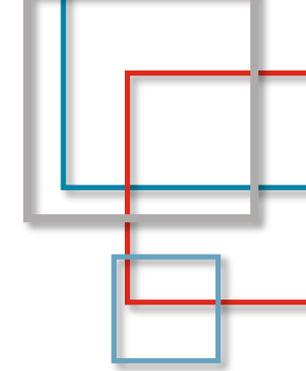


Randomization

John Doe



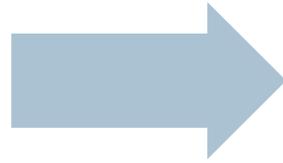
Bxqu Lmn



- Definition:
 - ランダムな文字列と置き換える
- Purpose:
 - 不可逆に匿名化する
- Key Features:
 - 元のデータと関連がない
 - 元のデータを推測できない
- Use Cases:
 - 高機密なデータ、匿名化されたテストデータ

Named/numbered replacement

John Doe

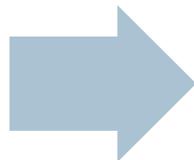


[名前][名字]

- Definition:
 - 個人情報をラベルに変換する
- Purpose:
 - 識別情報を残したままデータを匿名化する
- Key Features:
 - 一貫性のあるラベリング
 - 匿名化された情報の追跡が可能

Pseudonymization

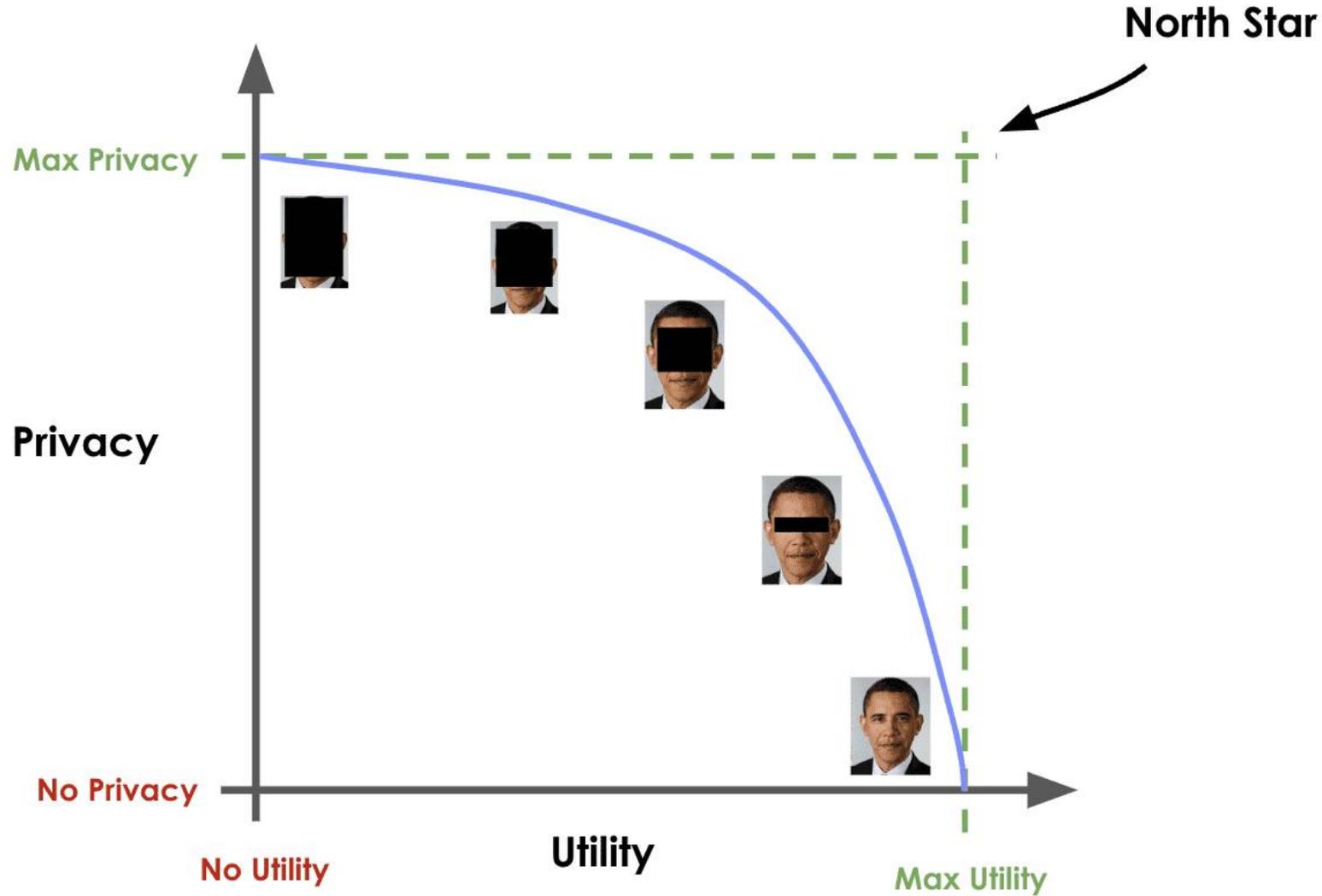
John Doe



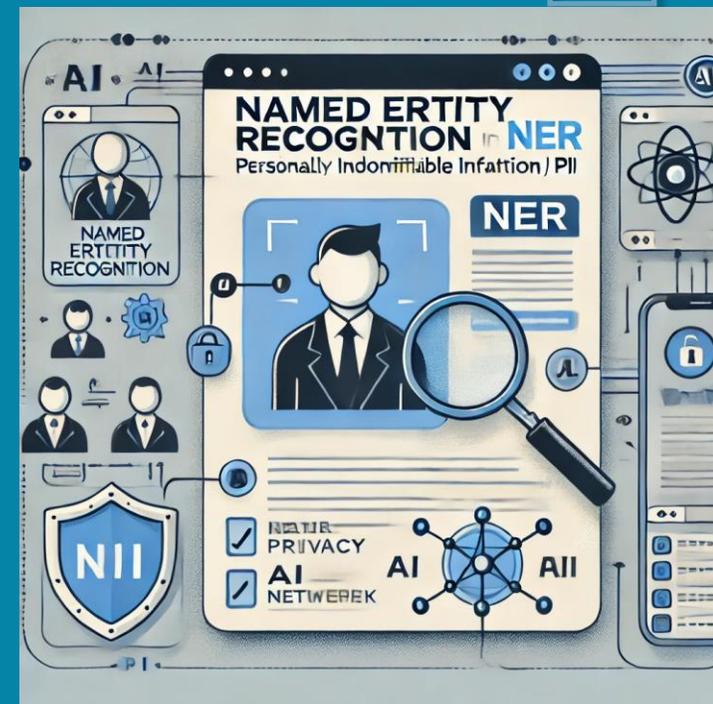
Mike Watson

- Definition:
 - 本物のデータをそれっぽいダミーデータに置き換え
- Purpose:
 - プライバシーと利便性を両立
- Key Features:
 - 一貫性があり、文脈が理解可能

プライバシーと利便性の関係



Named Entity Recognition (NER) 固有表現抽出



固有表現抽出とは

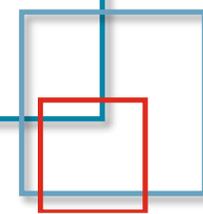
Taylor Swift is singing in a concert in Tokyo this coming December.

Person

Location

Time

- メール的重要情報を抽出・分類する
 - 氏名、メールアドレス、住所、etc..
- メール個人情報保護に不可欠
- 固有表現抽出の手法
 - Rule-based.
 - Machine Learning.
 - Deep Learning.

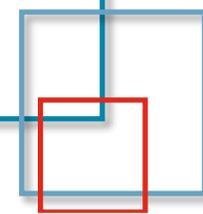


メール内の個人情報

- Examples of Key Entities:
 - 氏名
 - メールアドレス
 - 住所
 - 電話番号
 - 会社名
 - クレジットカード番号
 - 誕生日

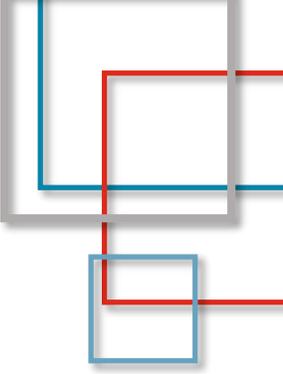
Small Language Model を用いる方法





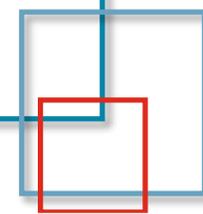
Small Language Modelとは?

- 特定のタスクに特化したモデル
- 少ないパラメーター:
 - 計算コストが少なくて済む
- Use Cases:
 - メール分類、リアルタイム処理
- Examples:
 - Bert, GPT-2 Small, DistilBERT, XLM-RoBERTa



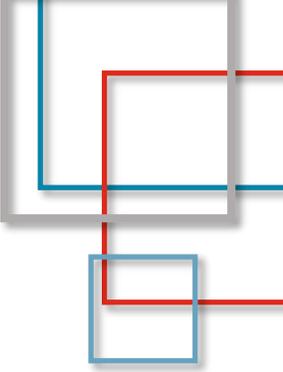
Small Language Modelを使用した固有表現抽出

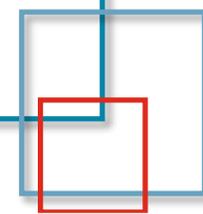
- Definition:
 - 名前、住所、日付などの固有表現を抽出する
- Purpose:
 - 小さなモデルで高速に個人情報の抽出と匿名化を行う
- Approach:
 - 小さなモデルを固有表現抽出タスク用にFine-tuneする
 - 個人情報を識別できるようにモデルを学習する
 - メール匿名化のモデルとして適用



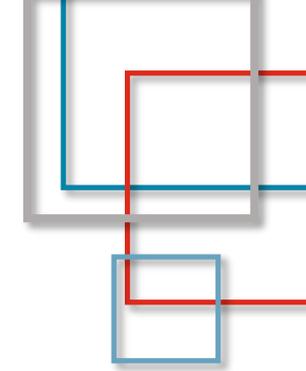
Small Language Modelのいいところ

- 処理が速い:
 - 軽量でリアルタイムな匿名化処理に最適
- 計算コストが低い:
 - 小さなデバイスや少ないリソースで動作可能
- 学習がタスクに特化しやすい:
 - メール処理に特化したfine-tunedがやりやすい

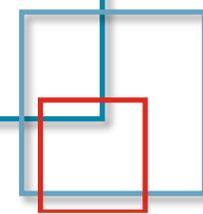




Small Language Modelの限界



- 容量が限られる:
 - 複雑な構造を理解するのは苦手
- 精度の高いラベル付データが必要:
 - 効果を発揮するにはきちんとラベル付けされたデータが必要。
人手でやるのはほんとに大変。
- 特定のタスクへの特化:
 - 特定のタスクにFine-tuneされているので、汎用性は低い
- Examples:
 - 長文のメールでは情報や文脈を取りこぼすかもしれない

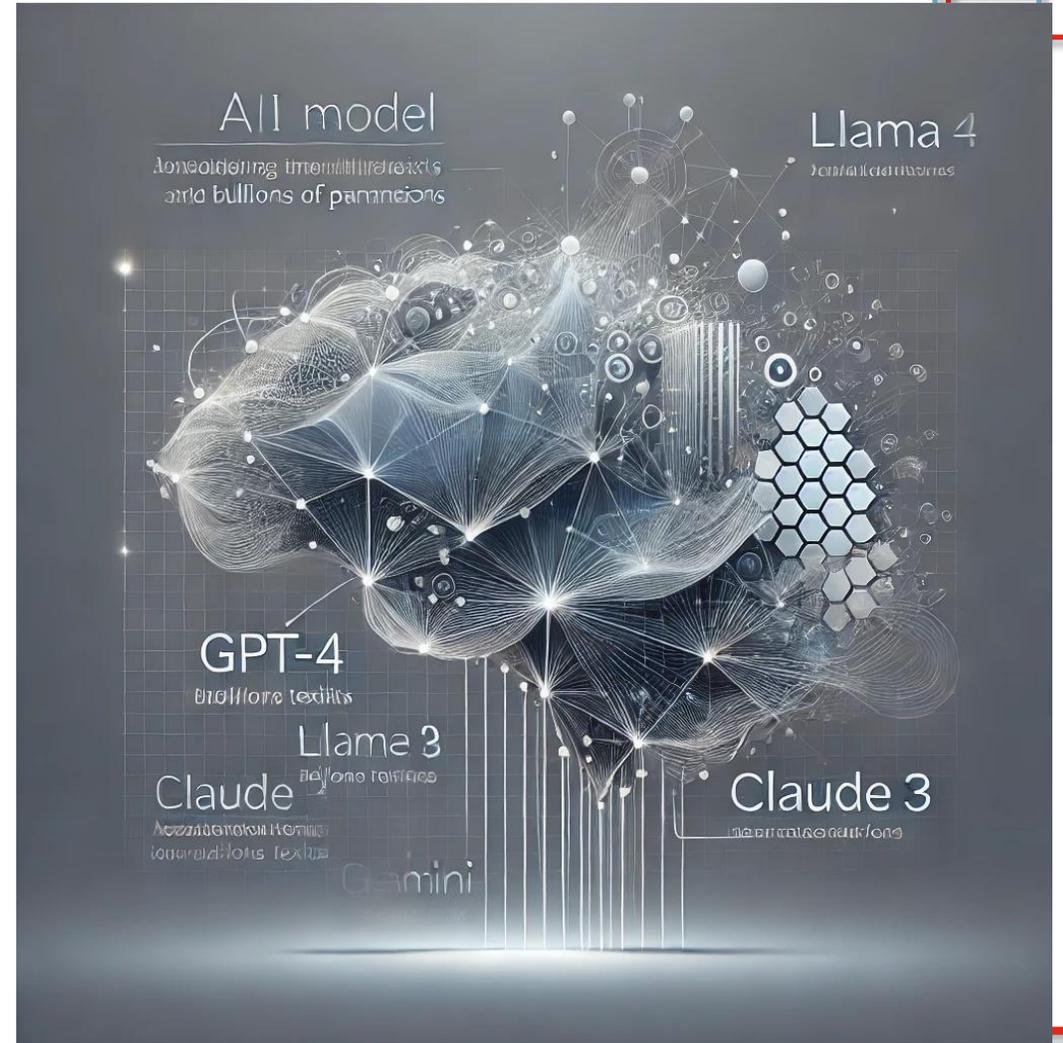


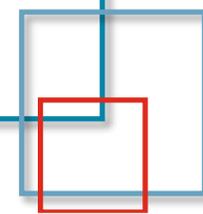
Small Language Modelを使う場面

- Real-Time Processing:
 - リアルタイムの匿名化や高速にデータを抽出するタスクに最適
- Resource Constraints:
 - ハードウェアや予算に制限がある場合に有効
- High-Volume Data:
 - あまり複雑ではないが大規模なデータを処理する場合に有効

Large Language Modelとは

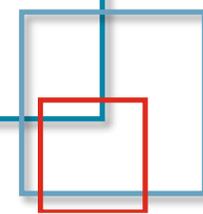
- 数十億のパラメーターで構成
- Understand:
 - 複雑な文章
 - 微妙なニュアンス
- Examples:
 - GPT-4
 - LLama
 - Claude 3
 - Gemini





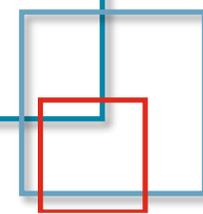
メール匿名化におけるLLM

- 複雑な文脈の理解:
 - LLMは機密情報に対して間接参照や微妙な表現でも判定できる
- 複雑な言語の理解:
 - 長い文章や複雑な表現のあるメールの匿名化に有効
- 様々なフォーマットへの適応力:
 - 構造、言語などに依らず様々なメールを効率的に処理可能



Large Language Modelの限界

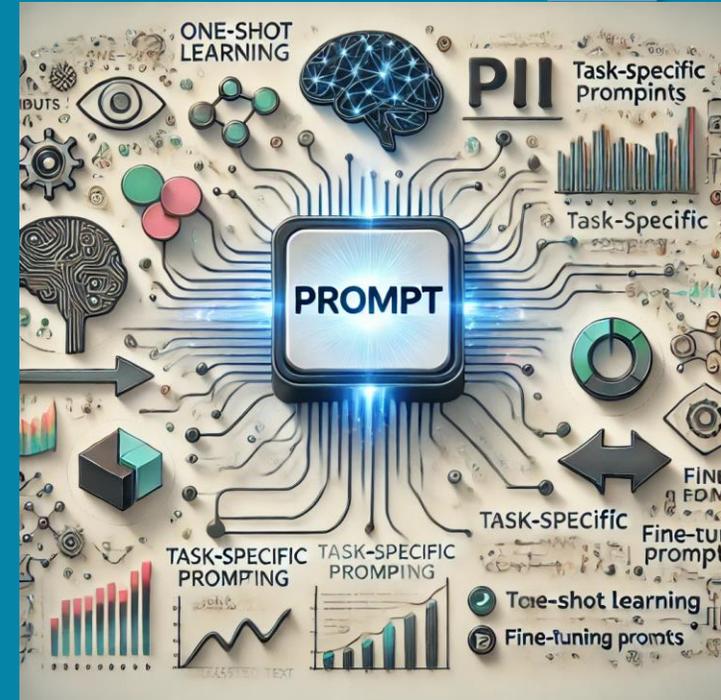
- High Computational Costs:
 - 効果的に運用するには大規模なハードウェアとリソースが必要
- Slower Processing Times:
 - 大規模なデータの匿名化には処理時間が長くなる
- Risk of Overfitting:
 - 新しいデータや多様な形式に対応するために慎重なfine-tuningが必要
- Risk of Hallucination:
 - ハルシネーションを起こさないような慎重なプロンプトの作成が必要
 - LLMはすぐにそれっぽい嘘をつく

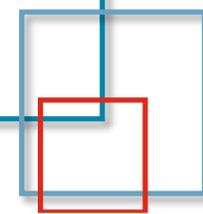


LLMを使う場面

- 複雑なコンテンツの匿名化:
 - 微妙な表現や間接参照を含む場合に有効
- 長い文章:
 - 長いメールや文書に対して有効
- 学習データが少ない場合:
 - 十分なラベル付けデータを用意できない場合に有効

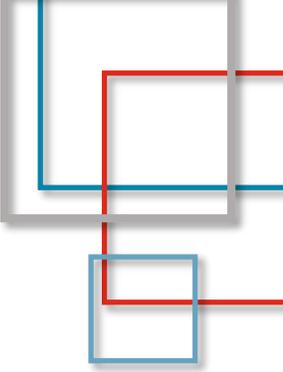
Prompt Engineering

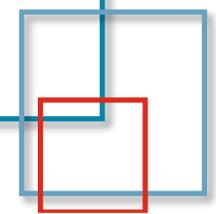




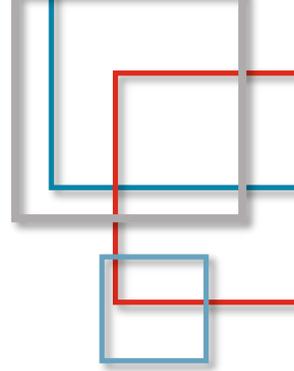
Prompt Engineeringとは

- Definition:
 - 欲しい結果が出力されるようにAIモデルに対しての入力を設計する
- Purpose:
 - モデルの出力と精度を向上
- Key Elements:
 - 明確な指示
 - 文脈と例示
- Benefits:
 - 精度の向上、出力のカスタマイズ

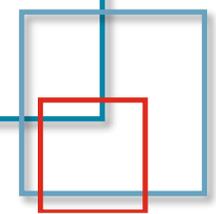




Prompt Engineeringの手法



- Basic Prompt
- One-Shot Learning
- Few-Shot Learning
- Task-Specific Prompting
- Role Prompting
- Chain of Thought (COT)



Basic Prompt

- Definition:
 - 単純な指示をモデルに与える
- Purpose:
 - タスクを完了させるための明確で簡潔なガイダンスを与える
- Example:
 - 「以下の文章から個人情報を抽出してください」
- Key Elements:
 - 直接的で簡潔
 - ひとつのタスクに集中させる

Basic Prompt 例

以下の文章から個人情報を抽出してください

--

平野 善隆様

ふるさと納税の寄付申込み、ありがとうございます。担当の田原俊彦と申します。

下記住所に神戸牛40kgを送付させていただきます。

〒100-0004

東京都千代田区大手町一丁目9番2号

Vade Japan株式会社

担当泉田様

11月13日午前中の到着となります。

よろしく願いいたします。

以下が抽出された個人情報です：

氏名

- 平野 善隆
- 田原 俊彦
- 泉田

住所

- 〒100-0004
- 東京都千代田区大手町一丁目9番2号

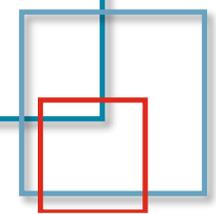
企業名

- Vade Japan株式会社

配送予定日

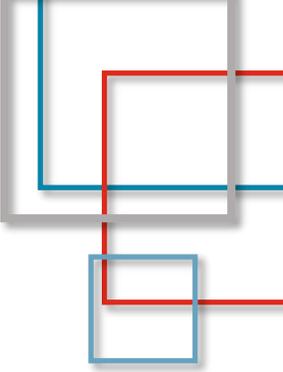
- 11月13日午前中

これらの情報は個人名、住所、会社名、日時に関するものであり、個人情報やプライバシーに配慮が必要な内容です。



One-Shot Learning

- Definition:
 - モデルに対して1つだけ例を与える
- Purpose:
 - 最小限のデータからモデルを汎用化させる



One-Shot Learning 例

以下の文章の個人情報を匿名化してください: '山田太郎の住所は東京都練馬区です' → '[秘密]の住所は[秘密]です'

--

平野 善隆様

ふるさと納税の寄付申込み、ありがとうございます。
担当の田原俊彦と申します。

下記住所に神戸牛40kgを送付させていただきます。

〒100-0004

東京都千代田区大手町一丁目9番2号

Vade Japan株式会社

担当泉田様

11月13日午前中の到着となります。

よろしく願いいたします。

[秘密]様

ふるさと納税の寄付申込み、ありがとうございます。
担当の[秘密]と申します。

下記住所に神戸牛40kgを送付させていただきます。

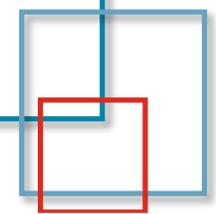
〒[秘密]

[秘密]株式会社

担当[秘密]様

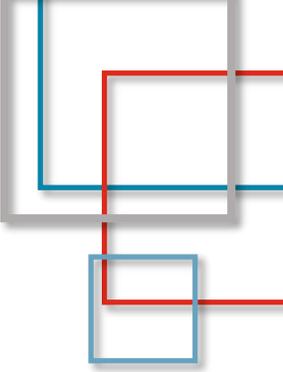
11月13日午前中の到着となります。

よろしく願いいたします。



Few-Shot Learning

- Definition:
 - モデルに対して複数の例を学習させる
- Purpose:
 - 複雑で多様なタスクが動作するようにする



Few-Shot Learning 例

以下の文章の個人情報を匿名化してください:

例:

- '山田太郎の住所は東京都練馬区です'
→ '[秘密]の住所は[秘密]です'
- '〒100-1111' → '〒[秘密]'

--

平野 善隆様

ふるさと納税の寄付申込み、ありがとうございます。担当の田原俊彦と申します。

下記住所に神戸牛40kgを送付させていただきます。

〒100-0004

東京都千代田区大手町一丁目9番2号

Vade Japan株式会社

担当泉田様

11月13日午前中の到着となります。

よろしくお願いいたします。

[秘密]様

ふるさと納税の寄付申込み、ありがとうございます。担当の[秘密]と申します。

下記住所に神戸牛40kgを送付させていただきます。

〒[秘密]

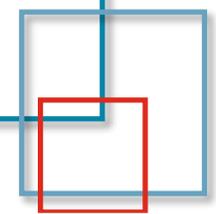
[秘密][秘密][秘密]

[秘密]株式会社

担当[秘密]様

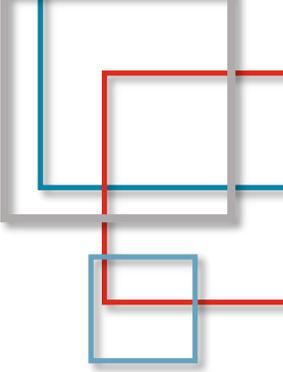
11月13日午前中の到着となります。

よろしくお願いいたします。



Task-Specific Prompting: Q&A Type

- Definition:
 - 特定の質疑応答に合わせてカスタマイズしたプロンプト
- Purpose:
 - 個人情報抽出に特化した答えを導き出す



Task-Specific Prompting 例

You are expert in Japanese Language and expert in PII detection.

Instructions:

- Your task is extracting the PII information.
- Only use the information provided in the text, do not make any assumptions or add additional details.
- What are the **names of the people** mentioned in the text
- What are the **addresses** mentioned in the text?
- What are the **prefectures** mentioned?

平野 善隆様

ふるさと納税の寄付申込み、ありがとうございます。担当の田原俊彦と申します。

下記住所に神戸牛40kgを送付させていただきます。

〒100-0004

東京都千代田区大手町一丁目9番2号

Vade Japan株式会社

担当泉田様

11月13日午前中の到着となります。

よろしくお願いたします。

Names of People:

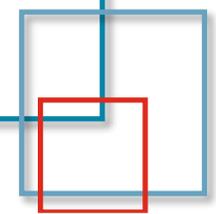
- 平野 善隆
- 田原 俊彦
- 泉田

Addresses:

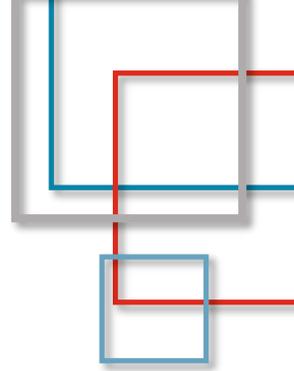
- 東京都千代田区大手町一丁目9番2号

Prefectures:

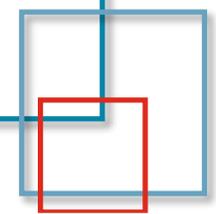
- 東京都



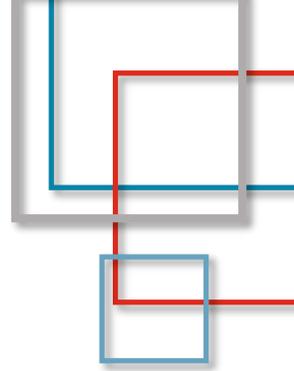
Role Prompting



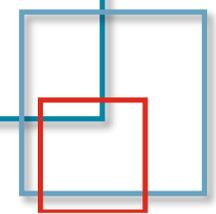
- Definition:
 - AIに役割や視点を与えるテクニック
- Purpose:
 - AIが文脈に合った結果を生成するようにする
- Examples:
 - **Imagine you are lawyer**. Explain Basic steps involved in filing lawsuit.
 - **You are a financial advisor**. Provide advice on budgeting.
- Benefits:
 - 専門的なタスクやシナリオに合った結果を返す。



Chain of Thought (CoT)

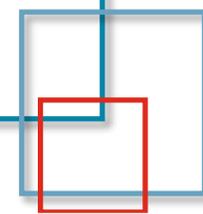


- Definition:
 - 推論を段階的に行わせるテクニック
- Purpose:
 - 特に複雑なタスクに対して問題を明確にし、正答率を上げる
- Benefits:
 - 構造的に推論させることでエラーが減り正答率が上がる
 - 中間のステップを理解できる
 - 推論の間違いを見つけやすくなり、プロンプトを修正しやすくなる



CoT – 例

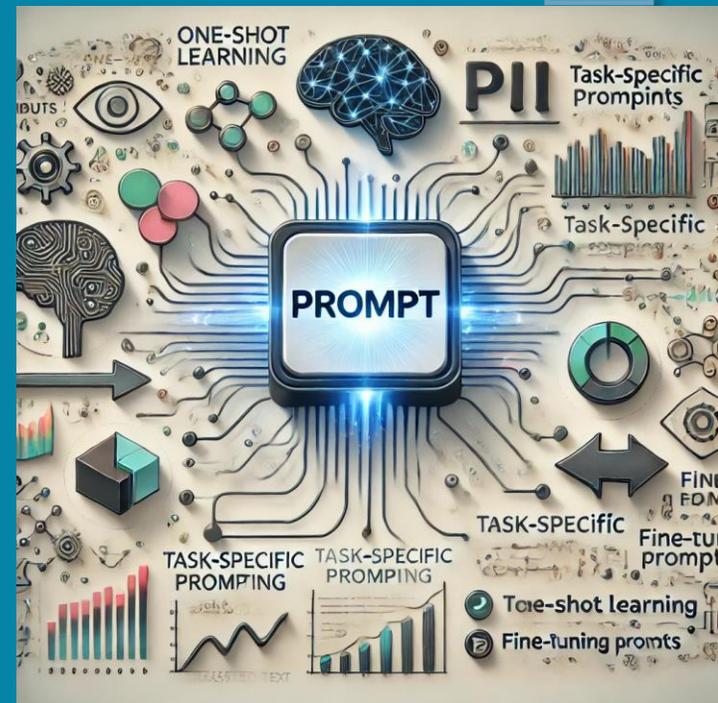
- CoTのない場合:
 - If 5 people each have 7 apples, how many apples are there in total?
 - Output: "There are 35 apples."
- CoTがある場合:
 - If 5 people each have 7 apples, how many apples are there in total?
Let's think through this step by step.
 - Output: "Each person has 7 apples. There are 5 people, so $5 \times 7 = 35$. Therefore, there are 35 apples in total."



Why Use Prompt Engineering?

- Guides the Model:
 - 特定のタスクにフォーカスさせることでよりよい結果が出る
- Flexible and Adaptable:
 - 特定の個人情報を抽出するなど様々なタスクに適応できる
- Improves Results:
 - 要求が明らかになりエラーが減る

Demo



Example Email

平野 善隆様

ふるさと商店をご利用いただき、誠にありがとうございます。
す。

ふるさと納税の寄付申込みを受付けました。
お申込みいただき、ありがとうございました。

申込番号: 943320567164

以下、お客様の入力内容です。

▼ 寄付者情報

メールアドレス: yoshitaka.hirano@hornetsecurity.com

氏名: 平野 善隆

ふりがな: ひらの よしたか

郵便番号: 103-0025

都道府県: 東京都

住所: 中央区日本橋茅場町3-11-10

電話番号: 03-5623-2530

▼ お礼の品と寄付金額

お礼の品: 松坂牛 40キロ 三重県 御浜町 1個 12,000円

合計寄付金額: 12,000円

▼ お支払い方法

お支払い方法: クレジットカード払い

カードの種類: VISA

カード番号: 4012 8888 8888 1881

カード名義: Yoshitaka Hirano

【お問合せ先】

御浜町ふるさと納税サポート室

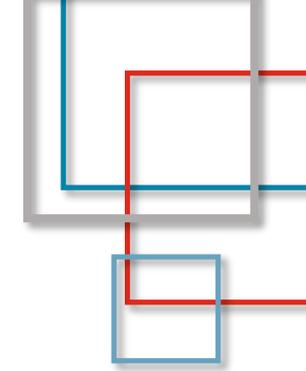
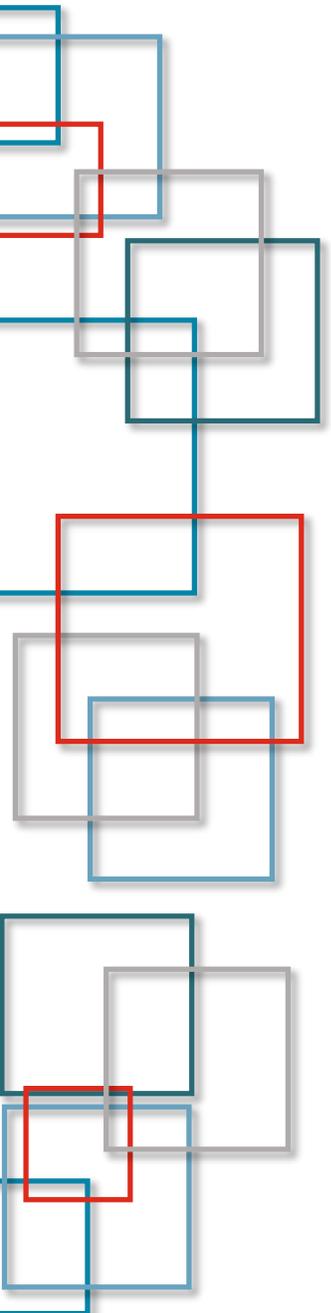
TEL : 050-3100-1730

受付時間 9 : 00~18:00 (土曜日・日曜日・祝日及び12月29日
~1月3日を除く)

<メールでのお問合せ>

support@mihama.furusato-lg.jp

Demo



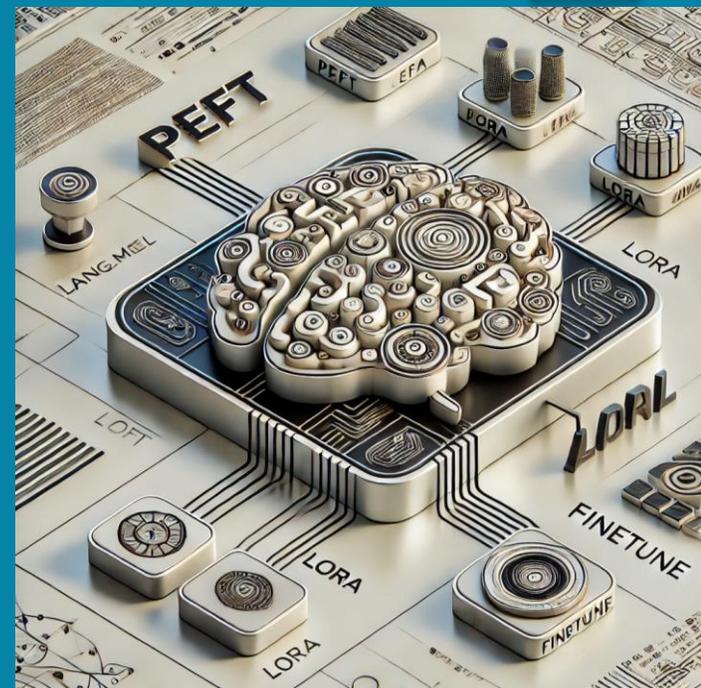
Example

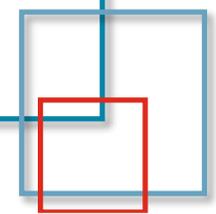
Entity	Basic Prompt	Role + Q&A Prompt	Role + Q&A + 1-Shot Prompt
平野 善隆	Detected	Detected	Detected
申込番号: 943320567164	Not Detected	Not Detected	Detected
yoshitaka.hirano@hornetsecurity.com	Detected	Detected	Detected
平野 善隆	Detected	Detected	Detected
ひらの よしたか	Not Detected	Not Detected	Detected
103-0025	Detected	Detected	Detected
東京都	Detected	Detected	Detected
中央区日本橋茅場町3-11-10	Detected	Not Detected	Detected
03-5623-2530	Detected	Detected	Detected
三重県	Not Detected	Not Detected	Detected

Example

Entity	Basic Prompt	Role + Q&A Prompt	Role + Q&A + 1-Shot Prompt
御浜町	Not Detected	Detected	Detected
4012 8888 8888 1881	Detected	Detected	Detected
Yoshitaka Hirano	Not Detected	Detected	Detected
御浜町ふるさと納税サポート室	Not Detected	Detected	Detected
050-3100-1730	Not Detected	Not Detected	Detected
support@mihama.furusato-lg.jp	Not Detected	Not Detected	Detected

LLMを改善する



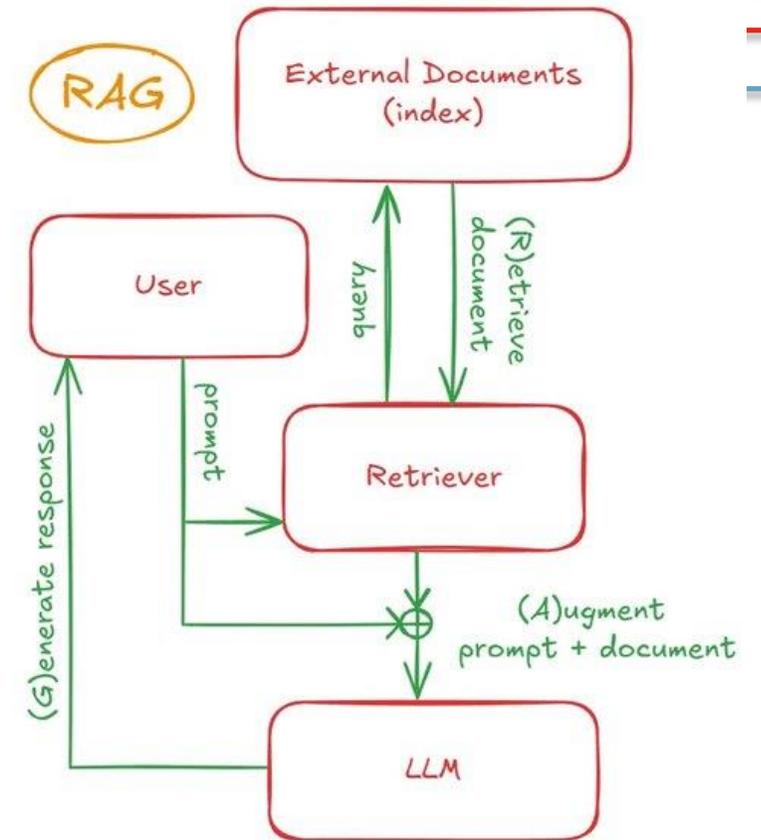


Improving LLMs

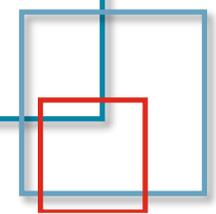
- RAG (Retrieval Augmented Generation)
 - 外部から取得した情報を使ってモデルの精度を上げる
- LLM Adapters
 - タスクに特化した効率的にfine-tuningするための軽量モジュール

Retrieval Augmented Generation

- What is RAG?
 - Definition:
 - LLMを外部の知識と組み合わせて強化する手法
 - Core Concept:
 - 外部データを保存 → データを取り込み → 応答を生成し返す
 - Key Benefit:
 - 静的LLMの学習当時の知識と現在の知識とのギャップや特定の知識とのギャップを埋める
 - Popular



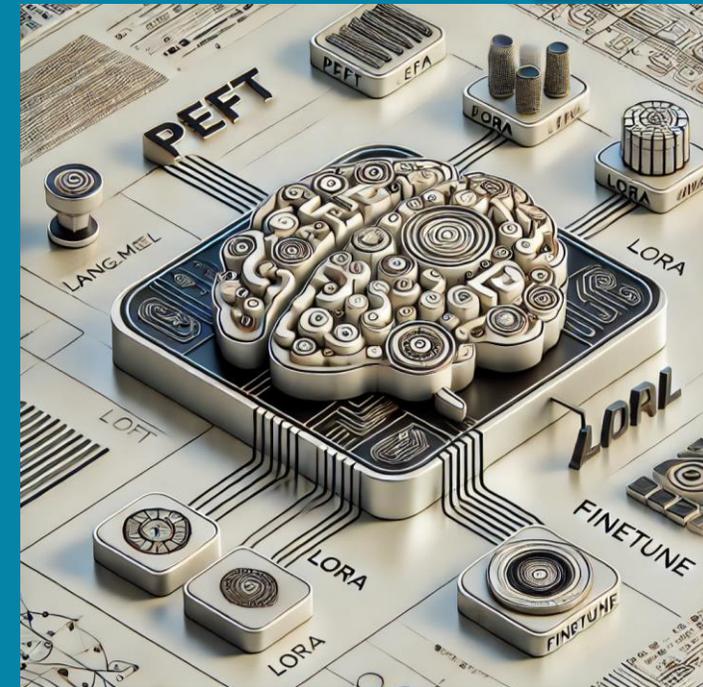
<https://levelup.gitconnected.com/the-best-rag-technique-yet-anthropics-contextual-retrieval-and-hybrid-search-62320d99004e>



Safer RAG

- 機密情報の匿名化でRAGを安全に使う:
 - Pre-processing:
 - RAGの知識ベースに追加する文書から機密情報を削除する
 - Runtime checking:
 - コンテンツに機密情報が結果に含まれていないことを確認する
 - Output filtering:
 - 念のため、結果に機密情報が含まれていないかを確認する

LLM Adapters



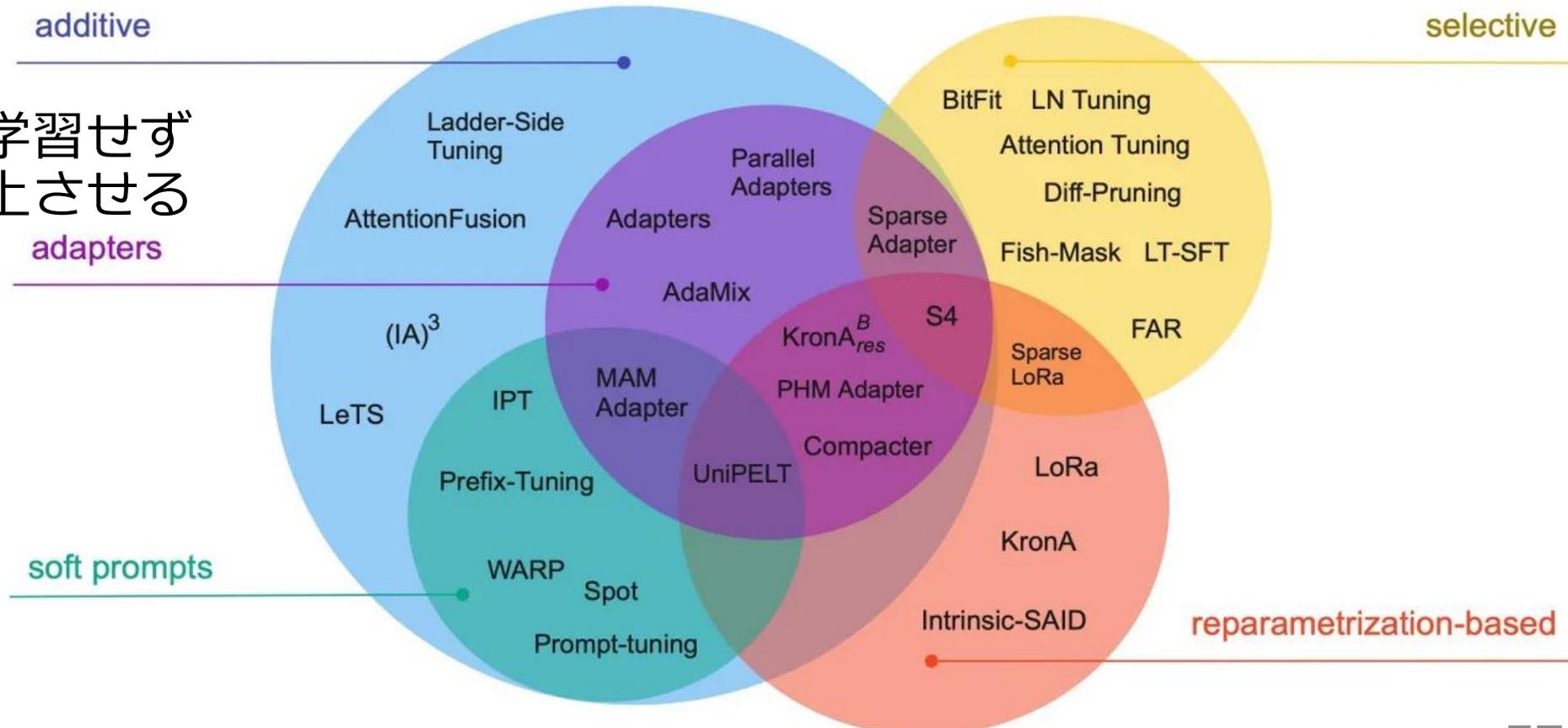
LLM Adapterとは

- **Definition:**

- 特定のタスク用に大規模言語モデルをfine-tuneする軽量モジュール

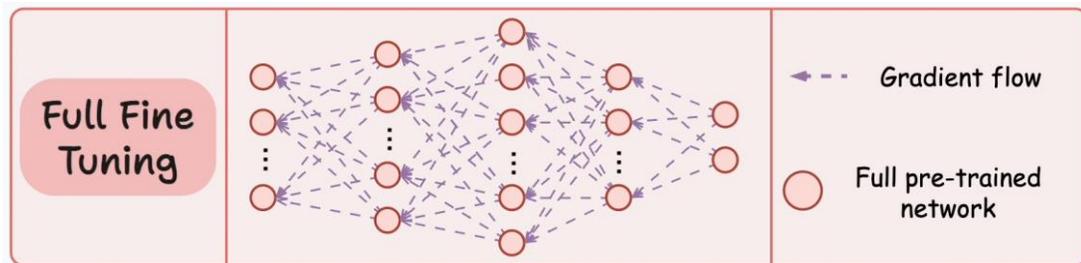
- **Purpose:**

- モデル全体を再学習せず
特定の能力を向上させる

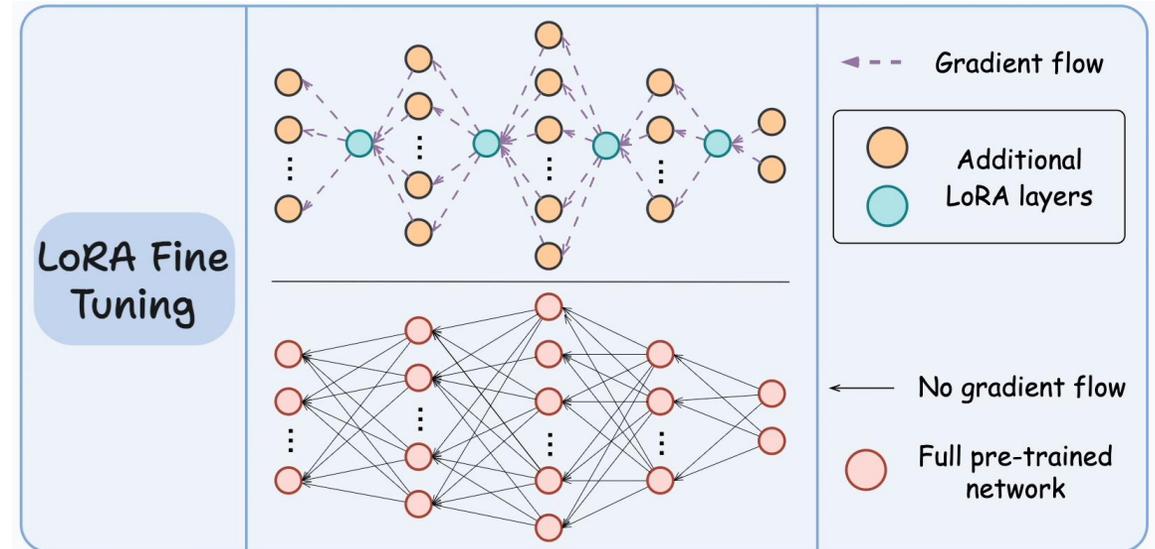


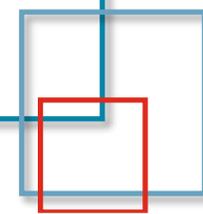
PEFT (Parameter-Efficient Fine-Tuning)

- PEFT (Parameter-Efficient Fine-Tuning):
 - モデルのパラメーターの一部だけを更新する
 - メモリや計算のリソースが少ない
- LoRA (Low-Rank Adaptation):
 - モデルの重みを調整するためにlow-rank行列を使用
 - 計算のオーバーヘッドを最小にする

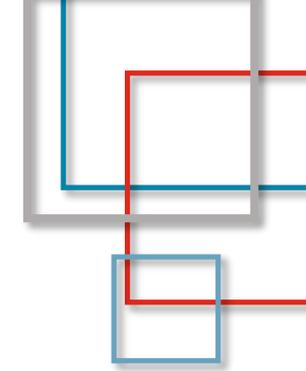


<https://blog.dailydoseofds.com/p/full-model-fine-tuning-vs-lora-vs>

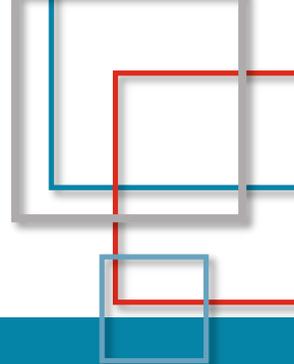
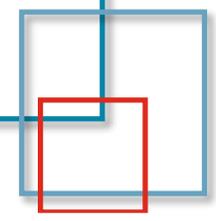




メールの匿名化でのLLM Adapterの利用

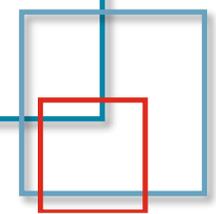


- Purpose:
 - メールの個人情報を匿名化するためにカスタマイズする
- Benefits:
 - 最小限の学習で素早く精度を上げることができる
 - オンプレ環境で学習させることで学習データ内の個人情報をセキュアに扱える

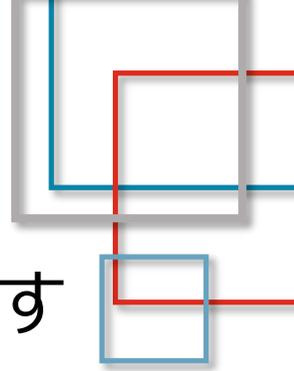


まとめ





まとめ

- メールから機密情報を取り除くことでメールデータを扱いやすくなる
 - 固有表現抽出 の技術で情報抽出できる
 - Small Language Modelは速い、安いが、準備が大変
 - Large Language Modelは遅い、高いが、準備は楽で柔軟
 - LLMの精度を上げる技術
 - プロンプトエンジニアリング
 - RAG
 - LLM Adapter
- 

Thank you

