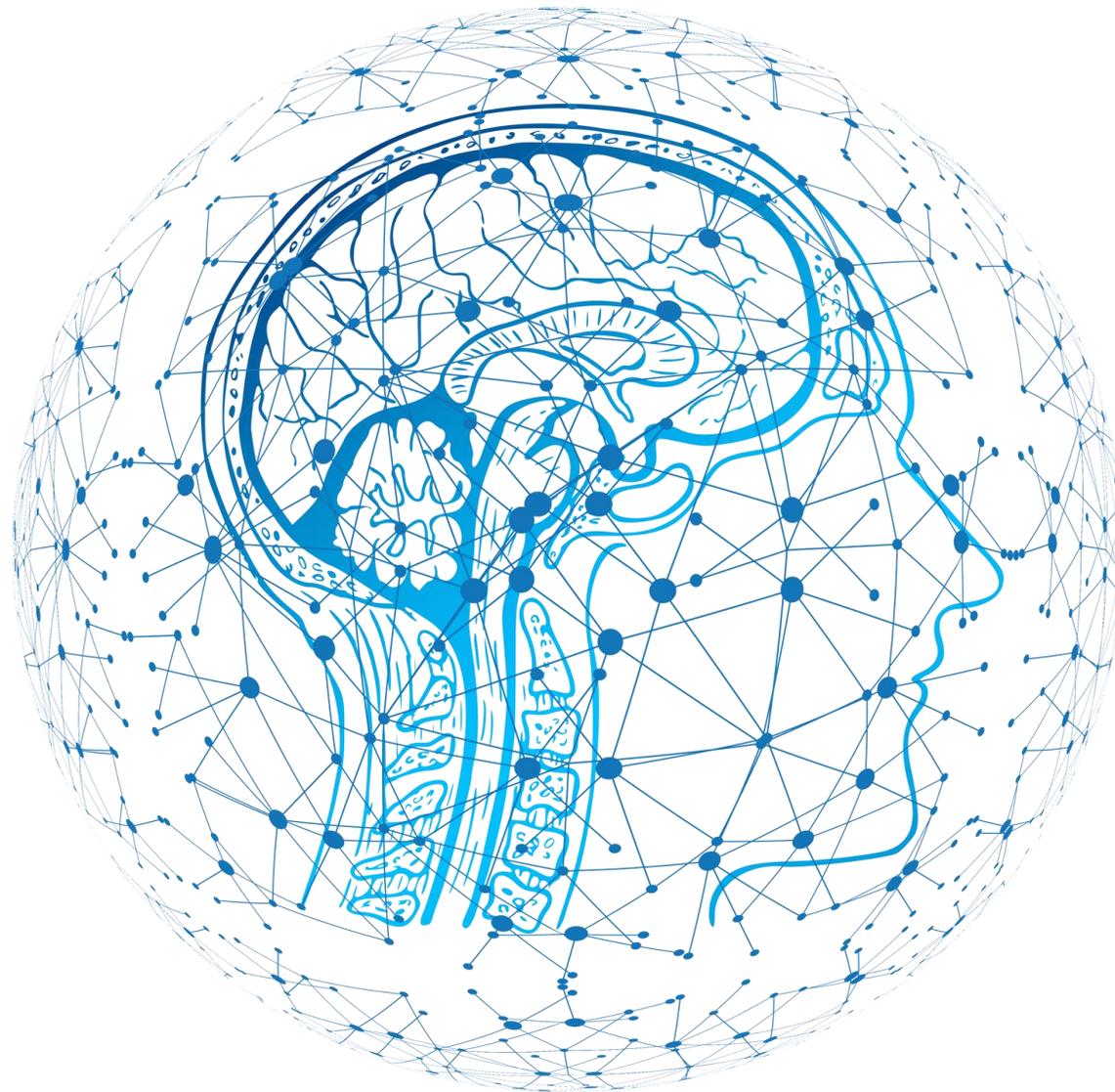




3分でわかる

大規模言語モデル

LLMの精度改善 ハンドブック



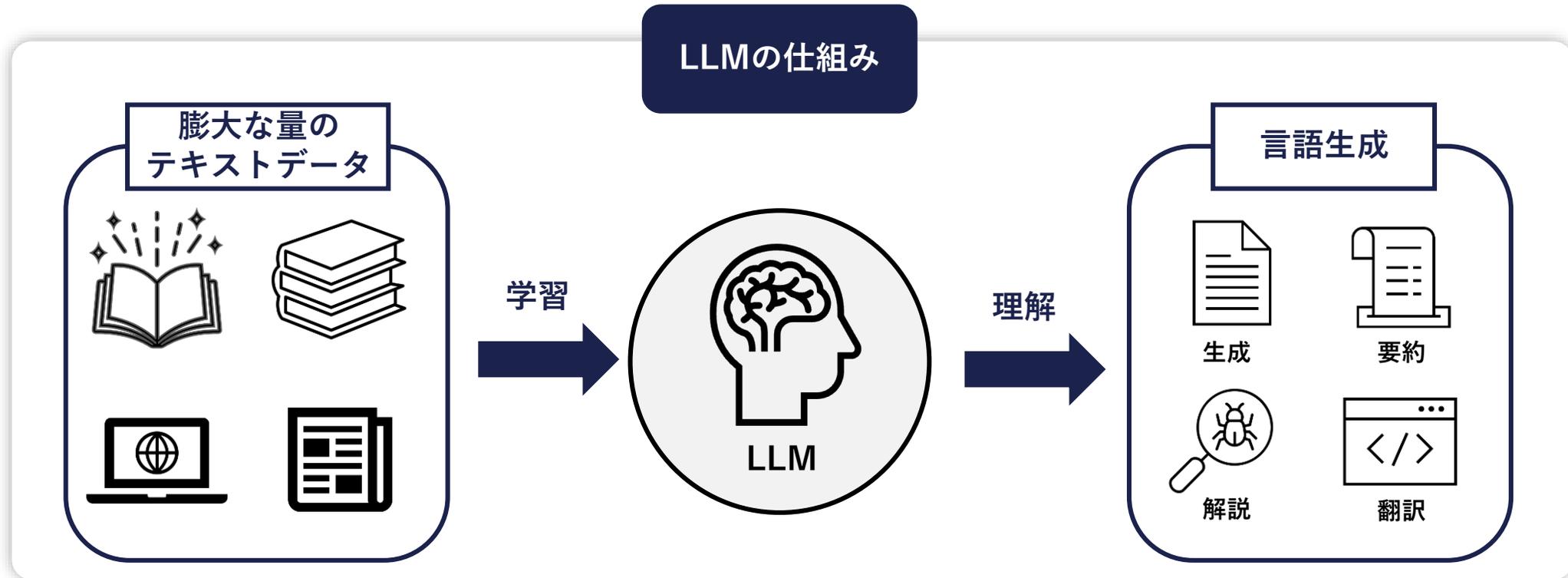
LLMの精度とは

LLM（大規模言語モデル）の開発においてその精度は非常に重要です
本資料では以下の順序でLLMの精度を左右するデータについて解説させていただきます

1.	LLMとは	3
2.	LLMの精度とは.....	4~5
3.	LLMの精度を改善するには.....	6
4.	事前学習について	7
5.	ファインチューニングについて	8
6.	アノテーションについて	9~10
7.	弊社LLM開発支援事業について	11

LLMとは

LLMは生成AIの一種で日本語では大規模言語モデルと言われており、
巨大なデータセットとAIのディープラーニングの技術を用いて構築されます
大量のテキストデータを学習することで人間の言語を理解し、
自然かつ流暢で論理的な言語生成を実現するものです
有名なGPT-4はこのLLMの一種となります



LLMの精度とは

LLM（大規模言語モデル）の開発において出力の精度は非常に重要です
LLMの精度が悪いとどんな文章が生成がされると思いますか？
以下に具体例を挙げさせていただきます

LLMの精度が悪い場合の生成例

例1. 不正確な応答

Q 東京タワーはどのくらいの高さですか？

A 東京タワーは約830メートルです。

例2. 文脈の誤解

Q ジャガイモを植えた後、どのくらいで収穫できますか？

A ジャガイモは一般に1年中いつでも収穫できます。

例3. 不適切なテキスト生成

Q ストレス管理に役立つ方法を教えてください。

A ストレスを感じたら、もっと頑張ってください。

例4. 言語理解の不足

Q AIの倫理的な問題について教えてください

A AIはコンピュータプログラムです

LLMの精度とは

前ページの具体例のような生成がなされてはユーザーは不便ですよね
では良いモデルの生成例はどのようなのか？
以下に具体例を挙げさせていただきます

LLMの精度が良い場合の生成例

例1. 不正確な応答

Q 東京タワーはどのくらいの高さですか？

A 東京タワーは約333メートルです。

例2. 文脈の誤解

Q ジャガイモを植えた後、どのくらいで収穫できますか？

A 植え付けから約70日から120日で収穫が可能です
が、品種や気候条件により異なります。

例3. 不適切なテキスト生成

Q ストレス管理に役立つ方法を教えてください。

A ストレス管理には、深呼吸、瞑想、適度な運動、
趣味への時間投資などが有効です。

例4. 言語理解の不足

Q AIの倫理的な問題について教えてください

A プライバシーの侵害、バイアスの問題、意思決定プロセスの透明性の欠如などがあります。

LLMの精度を改善するには

LLMは学習したデータを元に生成がなされます
よってその生成精度には元となるデータの量と質が非常に重要です

大規模な言語モデルの学習には基本的に「事前学習」と「ファインチューニング」の二つの段階に分けられ、それぞれのフェーズで異なる種類のデータセットが使われます

事前学習 (Pre-Training)

大規模言語モデルの学習の第一段階では、大量のデータを使用して大規模な自然言語コーパスを学習し、言語の基本的な言葉やパターン、文脈の理解の方法を学びます。

ファインチューニング (Fine-Tuning)

事前学習済みのモデルが特定のタスクや目的に合わせて微調整され、さらに学習を深める段階です。これにより、モデルは特定のタスクやドメインに対する性能を向上させることができます。使用用途に応じて、一部の学習済みデータと新たに追加したモデルの一部を利用して調整を行います。

例えば、社内情報に答えるための社内データのデータセットや医療業界のように専門的な用語が使われる場合はそのためのデータセットが使用されます。

事前学習について

LLMの精度を上げるにはデータセットにおいてあらゆる種類のテキストが含まれていることが理想的です。
その中で、重要なデータセットの一つとして事前学習において用いられるコーパスが挙げられます

コーパスとはテキストデータが大量に含まれるデータセットであり、
例えば以下のように様々な文書やテキスト、媒体から収集されます

コーパスはモデルがさまざまな言語表現を学習できるようにするために
多様なトピックやジャンル、スタイルの文章を含むことが重要です



Webページ



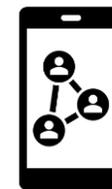
ニュース記事



書籍



Wikipedia



会話ログ・テキスト



他言語コーパス

ファインチューニングについて

事前学習が完了すると基本的なベースモデルが完成したと言えます
ただ、**人の指示を聞いたり対話を可能にするにはファインチューニングが必要**になります
ファインチューニングの段階においては、モデルにおける特定のタスクや領域における
性能を向上させるためことを目的にそれに特化したデータセットが必要となります

例えば、「高品質な翻訳」「特定トピックに関する質問応答」「テキスト上に現れる感情や意図の理解」
などの専門的なデータセットを場合にに応じて学習させる必要があります

また、**LLMの精度を上げるためにはファインチューニングの段階において
高品質で多様なアノテーションデータが必要不可欠**となります

ex. プライベートLLMに社内給与規定の情報を生成させたい際の
ファインチューニング



社内給与規定



事前学習済みモデル



追加学習済みモデル

アノテーションについて

LLM開発においてはデータの量と質がその結果を大きく変えますが、
様々な文書から収集されたコーパス、特定のタスクに関連する専門的なデータセットに並び、
ラベル付けされたデータが、モデルの質向上には不可欠です。

データにラベル付けを行うのは実際には基本的には人が行いますが、その工程をアノテーションといい
アノテーションされたデータのことを教師データと呼びます

そんなアノテーションの作業ですが大きく分けると以下の2つが挙げられます



人の手によるアノテーション

- ・ 自社でアノテーションを行う
- ・ クラウドワーカーのアノテーターに依頼
- ・ AIベンダーなどに委託する
等の方法があります



自動アノテーション

コストは安く抑えられる反面、
人の手が入らないため品質が不安定になっ
てしまう懸念があります

アノテーションについて（よくある課題）

LLMの精度向上には人の手によるアノテーションが不可欠ですが
アノテーション作業を自社で内製化する際は以下のような面で
課題を抱える企業様が多く散見されます。

そんな課題をお持ちの場合にはそれに割かれるリソースやデメリットを考えると
一定の費用はかかるものの外注するメリットは少なくありません



工数が膨大で多くの時
間を取られる



アノテーション作業の
リソースが足りない



モデルの改善が
うまくいかない

弊社のLLM開発支援事業について

APTOではharBestを活用したアノテーションプラットフォーム事業のみならず
LLM開発用のデータセット事業も行っております

当社のデータセット事業は、AI開発者向けに以下のサービスを提供しております
(APTO LLMデータセット事業 : <https://harbest.io/news/144>)



プライバシー保護に 力を入れたデータ提供

提供するデータセットプライバシー保護に対して細心の注意を払っており、開発者が法的なリスクを可能な限り回避できるようにします



最適な開発を支援する コンサルティングサービス

AI開発者が最適なデータセットを選択し、開発できるようにデータ使用に関する質問や疑問に対するサポートを提供します



データセットの マネジメント

新たなデータ要件に合わせてデータセットを継続的に更新し、開発者が最新のデータを利用できるようにします

株式会社オルツ様の事例

al+

IT・インターネット

<https://harbest.io/case/162>

日本発、大規模言語処理モデルOriginal LLM
「LHTM-2」の開発に挑む。



事例概要

自然言語の処理において、精度の高いデータ提供を "harBest" を通して実現。

「パーソナルAI」を掲げているだけに、対話コミュニケーションと自然言語処理系のご依頼が多いが

「パラメーターをいかに増やすか」や「変な日本語とかそのあたりも含めて直してください」等の

"わがままを聞いてもらえるデータ屋" としてオルツ様のLLM開発をご支援。

会社概要

AI開発の8割を占めるデータ収集・アノテーションに
特化したサービスを提供しております

社名	株式会社APTO
代表	高品 良
事業内容	① アノテーションツール提供 ② アノテーション受託 / AI開発受託 ③ データセット販売 ④ ノーコード / ローコードプラットフォーム提供
受賞歴	Tech Crunch japan 2021、Tribus、Gstartup
設立	2020年1月20日
拠点	本社：東京都渋谷区神南1-5-14 三船ビル403
Webサイト	https://harbest.io/



Best Venture 100

ベンチャー通信が「これから成長が期待される100社」をご紹介

2024年度ベストベンチャー100に選出

無料相談受付中

AI開発・アノテーション等に関するお悩みをお持ちの方は
下記からお気軽にご相談ください

webフォームから

click

無料お問い合わせ
AI開発相談はこちら

メールでのお問い合わせ

info@apto.co.jp

会社名・氏名・メールアドレス・電話番号を
ご記入の上、お問い合わせください

※当社のホームページからでも資料請求・お問い合わせが可能です。

<https://harbest.io/>