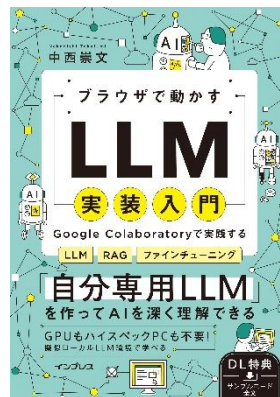


各 位

2026 年 8 月 4 日
株式会社インプレス

LLM をカスタマイズして仕組みがわかる！
『ブラウザで動かす LLM 実装入門 Google Colaboratory で実践する LLM・RAG・ファインチューニング』を 8 月 4 日（火）に発売

インプレスグループで IT・デザイン関連メディア事業を展開する株式会社インプレス（本社：東京都千代田区、代表取締役社長：高橋隆志）は、ブラウザだけで本格的な LLM（大規模言語モデル）実装を学べる書籍『ブラウザで動かす LLM 実装入門 Google Colaboratory で実践する LLM・RAG・ファインチューニング』を 2026 年 8 月 4 日（火）に発売いたします。



■ブラウザだけで本格的な LLM 実装を学べる

AI 活用のセキュリティ強化や、AI を目的に応じてカスタマイズする観点からローカル LLM の技術への注目が高まっています。しかし、本格的なローカル LLM 実装には高価な GPU 搭載 PC が必要という制約から、学習に踏み出せずにいる方も少なくありません。本書では、ブラウザで使える無料の実行環境「Google Colaboratory」を活用して、ローカル LLM に準じた環境でプログラムを実行しながらハンズオン形式で学べます。高性能な GPU 搭載 PC を持たない学習者でも、LLM の実装・カスタマイズを行いながら LLM の理解が深めることができます。

■公開 LLM モデルを自分専用カスタマイズし、LLM の仕組みを理解する

本書では、LLM の実装を学ぶメリットを示したうえで、ブラウザを用いた Google Colaboratory の操作方法を示し、Microsoft の Phi など公開 LLM モデルの利用法をステップ・バイ・ステップで解説します。RAG（検索拡張生成）による知識検索システムの構築からファインチューニング（追加学習）による LLM の文体の変更、小規模モデルの自作、Tool Calling を用いた AI エージェントの実装まで、実践的な事例で学べます。読者が自身の業務や研究課題に合わせて LLM をカスタマイズし、オリジナルな LLM を開発・提案できる応用力を養えます。本書で扱うソースコードはすべて読者特典としてダウンロード可能なので、コード実行のエラーに悩まされることなく実践できます。

■本書は以下のような方におすすめです

- ・ LLM を自分の手で動かして理解したい情報科学・データサイエンス系の学生

- ・ RAG などを活用した技術の習得や検証を試したいエンジニア
- ・ AI 教育・研究・企業研修に携わっており、LLM の仕組みについてハンズオンで教える必要がある立場の人
- ・ ローカル LLM 導入や LLM のカスタマイズに興味があるが、ハイスペックな PC が手元にない人

■紙面イメージ

新規ノートブックを開くと、以下のような画面が表示される。中央にあるセル（cell）と呼ばれる入力単位であり、ここにプログラムや説明文を記述する。Colab では、このセルを単位としてプログラムを分割し、順に実行していく構造を採用している。セルには2種類がある。1つは Python のコードを入力するための「コードセル」、もう1つは説明文や図表などを記述するための「テキストセル」である。



コードセルでは Python のプログラムを実行できる。テキストセルでは Markdown 記法を使ってタイトルや形式、番号書きをすることができ、画面上部のメニューにある「フォーマット」>「テキスト」ボタンを押すことで、どちらのセルも自由に追加できる。実行時は Python のコードを実行してみよう。

●コードセルをクリック



●「Enter」と入力し「Shift+Enter」を押す



●入力した式が実行され、計算の結果「2」が出力された



Colab では、「Shift+Enter」を押すと現在のセルを実行し、次のセルへ移動する。セルを追加して連続して実行することで、最終的にプログラムも構築できる。次に、もう少し複雑な Python プログラムを実行してみよう。以下のコードは、数列を渡した数値を返す関数と数列の総和である NumPy という数値計算ライブラリを使用し、配列の要素を1つずつ取り出して数列に出力している。

コード 2-1-1 繰り返し処理のサンプルプログラム

```
1 import numpy as np
2 a = np.array([1,2,3,4,5])
3 for i in a:
4     print("NumPy: ", i)
```

NumPy という数値計算ライブラリを使用し、配列の要素を1つずつ取り出して数列に出力している。

見やすい手順で、Google Colaboratory を迷わず操作して公開 LLM モデルを利用できる

6-2-1 「クジ」の類似度の計算

```
1 import torch
2 import torch.nn.functional as F
3 from transformers import AutoTokenizer, AutoModel
4
5 bert_t5 = "google-bert/bert-base-multispan-clsml"
6 bert_tokenizer = AutoTokenizer.from_pretrained(bert_t5)
7 bert_model = AutoModel.from_pretrained(bert_t5)
8 bert_model.eval()
9
10 @torch.no_grad()
11 def sentence_embedding(sentence: str) -> torch.Tensor:
12     tokens = bert_tokenizer.tokenize(sentence)
13     sentence = bert_tokenizer.tokenize(sentence)
14     tokens = torch.tensor(tokens)
15     tokens = tokens.unsqueeze(0)
16     tokens = tokens.to(device)
17     output = bert_model(tokens)
18     tokens = output.last_hidden_state
19     mean = torch.mean(tokens, dim=-1)
20     return mean
21
22 # 類似度の計算
23 sentences = [
24     "会社をクジにしよう",
25     "会社をクジにしよう",
26     "会社をクジにしよう",
27 ]
28 vectors = [sentence_embedding(s) for s in sentences]
29
30 for i in range(len(sentences)):
31     for j in range(i+1, len(sentences)):
32         score = F.cosine_similarity(vectors[i], vectors[j], dim=-1)
33         print(f"Similarity: {score}")
34
35 print("結果: この結果は、この結果は、この結果は、この結果は、この結果は、")
36
37 # 結果の平均値を計算して類似度を算出する
38 import gc
39 del bert_model
40 gc.collect()
```

このコードを実行すると、おおよそ次のような結果が得られる。

```
1 Similarity: 0.9999999999999999
2 Similarity: 0.9999999999999999
3 Similarity: 0.9999999999999999
```

結果から、「会社をクジにしよう」と「会社をクジにしよう」はコサイン類似度が非常に高く（0.9999999999999999）、この実行例では、「会社をクジにしよう」と「会社をクジにしよう」の文のベクトルが、ほぼ完全に一致していることが分かる。一方で、「会社をクジにしよう」と「会社をクジにしよう」を含む組み合わせでは類似度が低くなった。このBERT の実験は「クジ」という単語そのものではなく、文全体の Embedding 空間になっているので、BERT が「クジ」という単語の意味を正確に捉えているわけではない。BERT に基づく文 Embedding を用いると、「会社をクジにしよう」と「会社をクジにしよう」が正確に捉えられる。BERT は単語を単語で理解するのではなく、その文脈を考慮して意味を判断していることが確認できる。

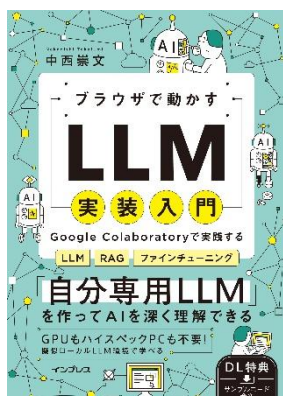
2-4 BERT のアルゴリズムを直感的に理解する

ここまで、BERT が文脈を理解して意味の異なる「クジ」を正しく区別できることを確認した。しかしながら、なぜ BERT がそのような高度な理解を実現できるのだろうか。BERT のアルゴリズムの仕組みをできるだけ直感的に、かつ学術的な観点から見てみる。BERT の構築にあるのは、Transformer というアーキテクチャである。Transformer は、入力された単語の順序を無視して処理するのではなく、すべての単語同士の間隔を同時に考慮する。この仕組みは「自己注意機構（Self-Attention）」と呼ばれる。これによって、BERT は文全体の単語がどの単語とどの単語と関係しているかを学習できる。単語で（例えば）「会社」と「クジ」を比較すると、BERT は「会社」の単語が「クジ」の単語とどの単語と関係しているかを学習している。人間が文章を読むとき、特定の単語を理解するために前後の文脈を思い出しながら意味を導く。たとえば、「会社をクジにしよう」という文章では

コードは全文ダウンロード可能。LLM の性質やカスタマイズ方法についての解説も充実

■本書の構成

- 第1章 いま、なぜ LLM の実装を学ぶのか——「使う」から「創る」へ
- 第2章 Google Colaboratory のセットアップをする
- 第3章 公開 LLM モデルを使う準備をする
- 第4章 Gradio で Web インターフェースを簡単につくる
- 第5章 ブラウザで LLM を動かし、その仕組みを理解する
- 第6章 知識を拡張する —— RAG でつくる「自分専用」検索システム
- 第7章 モデルを改造する —— ファインチューニングで好みの文体を覚えさせる
- 第8章 ブラックボックスの中身を知る —— Transformer で理解する AI の仕組み
- 第9章 AI エージェントにつながる Tool Calling を実践する



■書誌情報

書名：ブラウザで動かす LLM 実装入門 Google Colaboratory で実践する LLM・RAG・ファインチューニング

著者：中西崇文

発売日：2026年8月4日（火）

ページ数：256 ページ

サイズ：B5 判

定価：2,640 円（本体 2,400 円＋税 10%）

電子版価格：2,640 円（本体 2,400 円＋税 10%）※インプレス直販価格

ISBN：978-4-295-02481-1

◇Amazon の書籍情報ページ：<https://www.amazon.co.jp/dp/4295024813>

◇インプレスの書籍情報ページ：<https://book.impress.co.jp/books/1125101145>

■著者プロフィール

中西崇文（なかにしたかふみ）

東京工科大学コンピュータサイエンス学部教授。早稲田大学ナノ・ライフ創新研究機構 規範科学総合研究所 客員主任研究員（研究院客員准教授）、デジタルハリウッド大学大学院客員教授。専門は説明可能な AI (XAI)、データマイニング、感性情報処理。機械学習モデルの逆写像に基づく説明技術「AIME (Approximate Inverse Model Explanations)」の提案者として、医療、経済、教育、音楽など幅広い分野への応用研究を進めている。筑波大学大学院博士課程システム情報工学研究科修了後、情報通信研究機構 (NICT) 研究員、国際大学 GLOCOM 准教授・主任研究員、武蔵野大学データサイエンス学部准教授を経て現職。

著書に『ChatGPT はどのように動いているのか？』（翔泳社）、『シンギュラリティは怖くない』（草思社）、『Python ハンズオンによるはじめての線形代数』（森北出版）などがある。AI リテラシ教育や生成 AI 実装教育にも力を入れ、理論と実践をつなぐ研究・教育活動を展開している。

以上

【株式会社インプレス】 <https://www.impress.co.jp/>

シリーズ累計 8,000 万部突破のパソコン解説書「できる」シリーズ、デザイン・グラフィック関連の書籍ブランド「MdN」、「デジタルカメラマガジン」等の定期雑誌、IT 関連の専門メディアとして国内最大級のアクセスを誇るデジタル総合ニュースサービス「Impress Watch シリーズ」等のコンシューマ向けメディア、「IT Leaders」をはじめとする企業向け IT 関連メディアなどを総合的に展開・運営する事業会社です。IT・デザイン関連出版メディア事業、およびデジタルメディア&サービス事業を幅広く展開しています。

【インプレスグループ】 <https://www.impressholdings.com/>

株式会社インプレスホールディングス（本社：東京都千代田区、代表取締役：塚本由紀）を持株会社とするメディアグループ。「IT・デザイン」「音楽」「山岳・自然」「航空・鉄道」「モバイルサービス」「学術・理工学」を主要テーマに専門性の高いメディア&サービスおよびソリューション事業を展開しています。さらに、コンテンツビジネスのプラットフォーム開発・運営も手がけています。

【本件に関するお問合せ先】

株式会社インプレス 広報担当：丸山

E-mail: pr-info@impress.co.jp URL: <https://www.impress.co.jp/>