（IBM Blog「Generative AI that's tailored for your business needs with watsonx.ai」から
リンクされている、2023 年 11 月 30 日に更新された Granite Technical Paper の和訳）

Granite 基盤モデル

IBM Research

Abstract

企業での利用が可能な、生成 AI（人工知能）タスクのためのデコーダーのみ基盤モデル「Granite」シリーズを紹介する。アーキテクチャー、機能、基礎データとデータ・ガバナンス、学習アルゴリズム、計算インフラ、エネルギーとカーボン・フットプリント、テストと評価、社会技術的な弊害と緩和策、使用ポリシーについて報告する。

キーワード：基盤モデル、大規模言語モデル、LLM、生成 AI、データ・ガバナンス、対照ファイン・チューニング、エネルギー消費、評価、社会技術的弊害、利用ガバナンス、透明性文書

I. はじめに

本テクニカル・レポートでは、生成 AI タスクのためのデコーダーのみ基盤モデル Granite シリーズを紹介する。このシリーズで最初にリリースされた granite.13b は、英語専用の大規模言語モデル（LLM）である。ベースとなるこのモデルは、自己教師あり学習を使用して、セクション II で説明する IBM が作成した事前学習データセットで学習された。IBM は、エンタープライズ・グレードの基盤モデルを開発するために、社内のエンド・ツー・エンドのデータおよび AI モデルのライフサイクル・ガバナンス・プロセスと能力を活用している。そして同様の能力を watsonx プラットフォームを利用するお客様にも提供している。

granite.13b の最初のバージョン（v1）は 1 兆トークンで学習されたベースモデルを出発点にしていた。granite.13b の第 2 のバージョン（v2）のベースモデルは 2.5 兆トークンで学習された。どちらのバージョンにおいてもベースモデル（granite.13b）が、granite.13b.instruct と granite.13b.chat という 2 つのバリエーションへの出発点である。最初のバリエーションである granite.13b. instruct は、プロンプト・エンジニアリングによってエンタープライズ・タスクを実行するのに使用できるように、命令追従性[1]を改善する教師ありファイン・チューニングが行われた。第 2 のバリエーションである granite.13b.chat は、モデルの命令追従性をさらに向上させ、ある種の有害な出力を減らし、モデル出力が一定の社会規範に従うことを促し、有用性の概念を持つように、教師ありファイン・チューニングの後に新しい対照ファイン・チューニングを施してある[2]-[4]。我々は、これらの概念が普遍的なものではないことを強調し、社会技術的弊害とリスクに関するセクション VI でこの点についてより詳細に議論する。

granite.13b.instruct および granite.13b.chat モデルは、watsonx プラットフォーム [5] を通じて IBM が提供する。IBM は、watsonx プラットフォーム上でのこれらのモデルのお客様による使用に対して補償し、IBM 標準契約条件に従って、IBM のすべての製品と同じ、契約上の知的所有権保護を IBM が開発した AI モデルに提供する。

A. 能力の概要

名前の 13b は、モデルが 130 億個のパラメータを持つことを示す。さらに、ベースとなるデコーダーのみモデル granite.13b は、学習された位置埋め込みによるマルチクエリー・アテンション機構を持ち、 GPT-NeoX 20B トークナイザー[6]で作成された 1 兆個のトークンで学習され、8 千トークンのコンテキスト長を持つ。granite.13b で最初にリリースされたモデルはどれも、1 兆トークンで学習されたベースモデルの初期のチェックポイントに基づ

いて学習された。それに続くモデル（granite.13b.instruct.v2 と granite.13b.chat.v2）は、その後に追加で 1.5 兆トークンの学習データを与えられた granite.13b のチェックポイントに基づいて学習された。つまり granite.13b.v2 のモデルはどれも合計で 2.5 兆トークンを事前学習に利用している。セクション V で説明するように、Granite のモデルは、ベンチマーク評価において同サイズでの比較で競争力がある一方で、ガバナンスの点ではエンタープライズ対応である。

Granite モデルが使用される可能性のある主要なエンタープライズ・タスク（分野横断的に共通）には、検索拡張生成（RAG）、要約、コンテンツ生成、固有表現抽出、洞察抽出、および分類がある。Granite モデルは、watsonx プラットフォームのプロンプト・エンジニアリングによって、特定のエンタープライズ・アプリケーションで発生する特定のタスクに適応させることができる（図 1）。

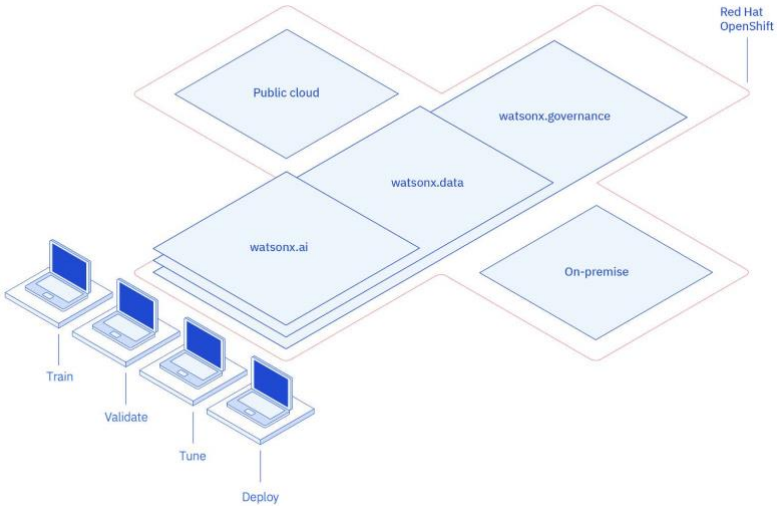


図 1 ：watsonx のコンセプト図

B. Granite 事前学習データセットの概要

granite.13b など大規模なエンタープライズ・グレードの基盤モデルの学習を実現するために、IBM は、アカデミア、インターネット、企業（金融、法律など）、コードなどにわたるソースから、関連する非構造化言語データの大規模なデータセットをキュレーションした（セクション III に詳述）。IBM はセクション II で学習データセットの説明を公開することで、透明性と責任ある AI へのコミットメントを示している。これはプロプライエタリな LLM を提供する大手企業としては珍しい。

Granite の事前学習データセットは、"The Pile" [7]や "C4" [8]のような LLM 学習用に一般的に使用されているオープンソースのデータセットに代わるものとして作成された。企業の自然言語処理にとって重要ないくつかのドメインは、これらのデータセットでは比較的十分に代表されていない。さらに、これらのデータセットは、有害、攻撃的、または海賊版のコンテンツを含んでいるとして批判されている[9]。独自の事前学習データ・コーパスをキュレーションすることで、IBM はこれらの問題やその他の問題に対処するための重要な一歩を踏み出す。

IBM がキュレーションした事前学習データセットは、コーパスに追加する追加データを定期的に審査・検討し、継続的に成長・進化している。事前学習データのサイズと範囲を拡大するだけでなく、フィルタリング機能の強化（重複排除、ヘイトスピーチや下品な表現の検出など）やツールの改善を反映して、これらのデータセットの新しいバージョンが定期的に生成され、維持される。

C. 報告書の構成

本レポートの残りの部分は以下のように構成されている。セクション II では、granite.13b の事前学習で使用したデータソースについて述べる。セクション III では、私たちが行ったデータ処理ステップについて、私たちが従うガバナンス・ステップを中心に説明する。セクション IV では、事前学習とファイン・チューニング・アルゴリズム、用いられた計算処理、および推定したエネルギー消費についてさらに詳しく説明する。セクション V では、他のモデルとの定量的比較とともに、テストと評価の枠組みを示す。セクション VI では、Granite モデルによる社会技術的弊害を理解し、軽減するためのアプローチについて述べる。セクション VII では、Granite モデルの使用ポリシーと社会技術的文書化について簡単に論じる。最後にセクション VIII で、今後の課題と考察を交え結論を述べる。

II. データソース

granite.13b の事前学習の初期フェーズを始めるまでに、IBM は前処理をする以前のサイズで 6.48 TB、前処理後のサイズで 2.07 TB のデータをキュレーションした（詳細はセクション III に記載）。すべてのデータセットはフィルタリングされた英文とコードの非構造化データファイルである。事前に定義されたラベルやターゲットはない。テキスト以外のアーティファクト（画像、HTML タグなど）はすべて除去した。

具体的には、ベースモデルの第 1 バージョンである granite.13b.v1 の学習のために、合計 14 のデータセットから 1 兆個のトークンが生成された。学習に使われた個々のデータセットは以下の通り。

1) arXiv: 180 万を超える科学論文のプレプリントが arXiv に投稿されている。
2) Common Crawl：ウェブ・クロール・データのオープン・リポジトリ。
3) DeepMind Mathematics：数学的な質問と回答のペアデータ。
4) Free Law：米国の連邦裁判所と州裁判所の公開法律意見。
5) GitHub Clean：CodeParrot による、様々なコーディング言語をカバーするコード・データ。
6) Hacker News：2007 年から 2018 年にかけてのコンピューター・サイエンスと起業に関するニュース。
7) OpenWeb Text：2019 年までのウェブページを含む OpenAI のウェブテキスト・コーパスのオープンソース版。
8) Project Gutenberg（PG-19）：米国の著作権が切れた古い作品を中心とした無料の電子書籍のリポジトリ。
9) Pubmed Central：生物医学・生命科学の論文。
10) SEC Filings：1934 年から 2022 年までの米国証券取引委員会（SEC）による 10-K/Q 提出書類。
11) Stack Exchange：Stack Exchange ネットワーク上のすべてのユーザー投稿コンテンツの匿名化されたセット。
12) USPTO：1975 年から 2023 年 5 月までに付与された意匠特許を除く米国特許。
13) Webhose：IBM が取得した非構造化ウェブコンテンツを機械可読データフィードに変換したもの。
14) Wikimedia：8 つの英語 Wikimedia プロジェクト（enwiki、enwikibooks、enwikinews、enwikiquote、enwikisource、enwikiversity、enwikivoyage、enwiktionary）。ページや記事から抽出されたプレーンテキストを含む。

ベース・モデルの第 2 バージョンである granite.13b.v2 は、新しくキュレーションされた 1.5 兆トークンの学習データを追加して granite.13b.v1 の事前学習を継続したもので、合計 2.5 兆トークンを事前学習に利用している。この学習トークンの第 2 部分に使われたデータセットは、granite.13b.v1 用と同じ 14 データセット（ただし Common Crawl からスナップショットを追加）と以下に列挙した 6 つの新しいデータセットの混合である。全てのデータセットとスナップショットはセクション III に記述された処理を施して使用した。

15) Earnings Call Transcripts: 企業が投資家に対して行う四半期決算報告会の書き起こし。このデータセットは決算報告会の書き起こしと、関係のある株価、業種別指数を集めた報告となっている。

16) EDGAR Filings: 25 年以上にわたる米国での全ての上場企業のアニュアルレポート。
17) FDIC: 連邦預金保険公社（FDIC）の年次報告書に基づいたデータ。
18) Finance Text Books: ミネソタ大学（UMN）の Open Textbook Library に基づいたコーパスで、財務のタグが付けられた全てのテキストブックが含まれる。
19) Financial Research Papers: 公開されているフィナンシャル・リサーチ・ペーパーのコーパス。
20) IBM Documentation: IBM のレッドブックや製品関連ドキュメント。

III. データ・ガバナンス

IBM は、Granite モデルをお客様自身のアプリケーションに適応できるようにするため、IBM の標準的なデータ・クリアランス・プロセス、文書品質チェック、その他の基準を含む、ガバナンス、リスク、コンプライアンス（GRC）基準に対してデータセットを評価するデータ・ガバナンス・プロセスに多大な投資を行ってきた。IBM は、IBM の AI 倫理の原則に合致し、IBM コーポレート・リーガル・チームによって指導される、LLM 事前学習データセットのガバナンス手順を開発した。LLM 開発に関するベスト・プラクティスは、AI モデルやその使用方法、規制要件の変化などに対する理解がますます深まるにつれて、継続的に進化している。

データの GRC 基準への対応は、データの要求からトークン化まで、学習データのライフサイクル全体に及ぶ。IBM にとって重要な目的は、学習された基盤モデルから、そのモデルの学習に利用したデータセットの特定のバージョンまでたどることができるような内部監査可能なリンクを確立することであり、これには学習前に実行された各処理ステップに関する情報も含まれる。IBM がキュレーションした事前学習用データセットに関する統計概要は、図 2 に示されている。

図 2： IBM のキュレーションした事前学習データセットのガバナンス状況統計量（granite.13b 学習時点）

データガバナンスは、モデル学習前のデータ・ライフサイクルの各フェーズに対応する以下のプロセスに整理される：

A. データのクリアランスと収集
B. 前処理
C. トークン化

各プロセスは、特定のガバナンスの側面に焦点を当てたサブ・プロセスで構成されている。本節の残りの部分では、各フェーズについて詳しく説明する。

A. データのクリアランスと収集

データ・クリアランス・プロセスは、Granite シリーズを含む IBM の基盤モデルの学習に、いかなるデータセット

も、慎重に検討されないで使用されることはないことを保証する。あるデータが、IBM のキュレーションした事前学習データセットに追加される前には、そのデータはデータ・クリアランス・プロセスに提出され、技術、ビジネス、およびガバナンスの観点でのレビューの対象となる。クリアランス要求は、データセットの詳細な説明、データ所有者、使用目的、地理的位置、データ分類、ライセンス情報（あれば）、使用制限、機密性（個人情報など）など、データセットに関する包括的な情報を取得する。追加情報には、誰がデータにアクセスできるのか、データはどのように取得されるのかなどが含まれる。

データセットが審査プロセスを完了すると、利用候補としてタグづけられ、そのデータセットのメタデータは承認済みデータセットのカタログに移され、そしてデータセットそのものがダウンロードされ、後続の前処理段階に準備される。

備考：IBM の事前学習用データセットは現在、海賊版（違法にアップロードされたコンテンツ）を発信していることが知られているウェブサイトの URL ブロックリストを選択に使用することで、著作権で保護された素材に対処している。ブロックされているデータセットの例としては、Books3 データセットがあり、これはデータの著作権の状態やモデル学習での使用に関する懸念のため、特に使用から除外されている。

B. 前処理パイプライン

データ利用条件がクリアされダウンロードされると、前処理パイプラインと総称されているさまざまなステップを経て、モデル学習の準備が行われる。今回のリリースの Granite モデルの場合で、前処理パイプラインの概要を図 3 に示す：

1) テキスト抽出
2) 重複排除
3) 言語の識別
4) 文の分割
5) ヘイトスピーチ、暴力的表現、下品な表現へのアノテーション（タグづけ）
6) 文書品質のアノテーション
7) URL ブロックリストのアノテーション
8) フィルタリング
9) トークン化

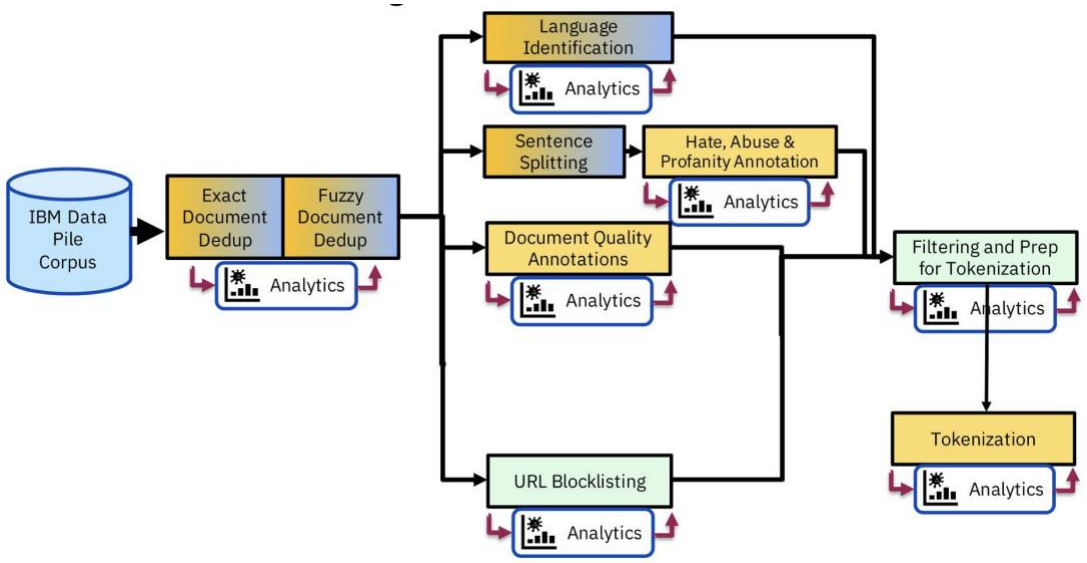


図 3：IBM のデータ前処理パイプライン

いくつかの前処理ステップは、アノテーション／フィルタリングのやり方で処理される。すなわち、ドキュメントや文がまずアノテーションされ、後でフィルタリング・タスクのステップで閾値と比較してフィルタリングされる。パイプラインの各ステップの完了はログに記録される。ログは、そのデータセットに対して実行された前処理ステップをメタデータに正確に記録するために使用され、モデル・ライフサイクルのエンド・ツー・エンドのトレーサビリティーの土台となる。

ここで、前処理パイプラインの各ステップをより詳細に説明する。

1) テキスト抽出：テキスト抽出はパイプラインの最初のステップであり、さまざまな文書から言語を抽出し、その後の処理のために標準化された形式にするために使用される。
2) 重複排除：重複排除の目的は、重複する文書を特定し、削除することである。重複排除はデータセット単位で行われ、学習済みモデルがデータセット内の繰り返しデータによって不自然な言語パターンを学習しないようにするために不可欠である。

厳密な重複排除とファジーな重複排除の 2 つの手法を使うが、どちらもハッシュに基づいた手法を用いる。その名が示すように、厳密な重複排除はデータセット中の文書間の厳密な重複を除去する。各文書はハッシュ化され、同じハッシュを持つ文書は一つに融合される。例えば、データセット中の 50 の文書が同じハッシュを持つ場合、1 つの文書が使われる。ファジーな重複排除は局所性を考慮したハッシュで文書間の Jaccard 類似度を求める。データセットの複数の更新スナップショットがダウンロードされた場合、正確な重複排除はすべてのスナップショットにわたって実行される。

3) 言語識別：言語の識別は、Watson の自然言語処理（NLP）ライブラリ[10]を使用して、支配的な言語を検出するために文書レベルで実行される。

このタスクの出力は、パルケ・ファイル（parquet ファイル）に対する追加列で、そこには 2 文字の ISO 言語コードが記入される。Common Crawl データセットの場合、言語はフォルダ名を通じてすでに提供されている。しかしそれでも、Watson NLP 言語識別アルゴリズムは Common Crawl 文書に対しても実行され、これらの文書に対して、Common Crawl と Watson NLP の 2 つの言語分類が得られることになる。

4) 文の分割：文の分割では、各文書を、構成する文に分解する。HAP アノテーションは文レベルで行われるため、文の分割は（後述する）ヘイトスピーチ、暴力的表現、下品な表現（HAP）アノテーションの鍵となる。そのため、HAP アノテーションを開始する前に、文の分割を行う必要がある。英語の文分割は、Watson NLP を使用して行われる。
5) ヘイトスピーチ、暴力的表現、下品な表現（HAP）アノテーション：Common Crawl のようなオープン・インターネットからのデータ・ソースには、悪い言語表現が必然的に含まれる。Granite モデルがそのようなコンテンツを生成する可能性を減らすため、各文書の各文は HAP コンテンツのレベルについて評価され、スコアリングされる。HAP 検出器はそれ自体、IBM によって訓練された言語モデルであり、内部モデルや，OffensEval [11]、AbusEval [12]、HatEval [13]などの公開モデルに対してベンチマークされている。IBM HAP 検出器は、HateBERT [14]と同等の性能を発揮する。文書内の各文章にスコアが割り当てられた後、文とスコアに対して分析が実行され、HAP アノテーションを持つ各文書のアノテーションの分布が調べられる。これは、文書内の HAP 文の割合を決定するためと、後でフィルタリング時に使用する閾値を決定するための両方の役割を果たす。
6) 文書の品質：品質アノテーションの目的は、ヒューリスティクスと分類器の両方を用いて、言語的価値の低い文書を特定することである。ヒューリスティクスは Gopher の品質フィルタリング基準[15]に由来する：
   - 総語数：50〜100,000 語の範囲外；
   - 平均単語長：1 単語あたり 3〜10 文字の範囲外；
   - 単語に対する記号の比率：10%以上
   - 箇条書きの比率：90%以上
   - 3 点省略記号（...）の比率：30%以上；

- アルファベット単語の比率：80%未満；
- 一般的な英単語：{the、be、to、of、and、that、have、with}から少なくとも 2 つを含まない。

分類器の方は、ウィキペディアの文書で事前に訓練された KenLM 線形分類器[16]、[17]を使って、パープレキシティー・スコアを割り当てる。どの文書に対しても、モデルは学習コーパス（すなわち Wikipedia）に対する文書の類似度スコアを提供する。これらのヒューリスティクスと分類器は、パルケ・ファイルに品質スコアの列を追加で出力する。これらのアノテーションは、フィルタリング・ステップにおける品質フィルタリングの要素となる。

7) URL ブロック・リスティング：あらかじめ決めたブロック・リストに載っているサイトの文書は、IBM のキュレーションした事前学習データセットに追加しないようブロックする。継続的に維持されるブロックリストには、著作権で保護されていることがわかっている資料の URL や、2022 Review of Notorious Markets for Counterfeiting and Piracy [18]に含まれるようなブロックリストのサイトが含まれる。

8) フィルタリング：フィルタリングは文書レベルで行われ、トークン化の前の最後のステップである。ここでは、以前の前処理ステップで作成されたアノテーションに基づいて、問題のある文書をトークン化に使用されないようにはじく。例えば、HAP の閾値を超える文書や、定義された品質を満たさない文書は除外される。現在の英語のみの Granite モデルでは、言語識別アノテーションは英語以外の文書を除外するために使用される。

C. トークン化

トークン化は、モデル学習前の最後の前処理ステップである。granite.13b では、クリーニングおよびフィルタリングされたテキストは、GPT-NeoX 20B トークナイザー [6]を使用して、文字列からトークンのベクトルに変換される。

IV. 学習

このセクションでは、事前学習とファイン・チューニングのアルゴリズムの詳細、必要な計算、そしてカーボン・フットプリントの見積もりも含めて、デコーダーのみ Granite モデルの学習プロセスについて詳述する。

A. アルゴリズムの詳細

1) 事前学習：私たちは[19]の事前学習設定をほとんどすべて採用している。具体的には、標準的なデコーダーのみトランスフォーマー・アーキテクチャー[20]、ガウス誤差線形ユニット(GELU)活性化関数[21]、推論効率のための MultiQuery-Attention[22]、学習された絶対位置埋め込みを用いる。また、学習を高速化し、メモリ・フットプリントを削減するために FlashAttention [23] を採用することで、コンテキスト長を多くの既存 LLM で使用されている 2,048 から 8,192 に増やした。

granite.13b.v1 ベースモデルは、バッチサイズ 4M トークン、合計 1 兆トークンを用い、300K イテレーションで学習されている。granite.13b.v2 ベースモデルはその granite.13b.v1 チェックポイントに追加して事前学習を継続し、300K 追加イテレーションと合計 2.5 兆トークンを使用した。
最適化アルゴリズムは Adam [24]を使い、$\beta 1 = 0.9$, $\beta 2 = 0.95$, $\varepsilon = 10\text{e-}8$, ウェイト・ディケイ 0.1 で学習した。2,000 ステップのウォームアップで、コサイン・ラーニング・レート・スケジューラーを用い、最終的にラーニング・レートを $3 \times 10\text{e-}4$ から $3 \times 10\text{e-}5$ まで減衰させる。8K のコンテキスト長での学習を可能にするために、シーケンス並列も含めたパイプライン並列とテンソル並列の両方を使用した 3 次元並列レイアウトでモデルを事前学習する。さらに、granite.13b.v2 モデルの学習には FlashAttention-2 [25]を利用して、8k コンテキスト長を持つモデルの学習と同じコストで、はるかに長いコンテキスト長（例：16k）への対応を可能にした。

2) Granite.13b.instruct のアライメント（AI の応答を良い方向に方向づけること）：事前学習では、LLM が入力テキストに後続する単語を生成し続けるように教示する。しかしユーザーは実際には、入力テキストを、従うべき命

令として扱うことを LLM に期待することが多い。そのような命令追従性を可能にするために、私たちは異なるソースからのデータセットを混合して教師ありファイン・チューニング（SFT）を実行する。各サンプルはプロンプトと回答から構成される。ラーニング・レートの初期値 2×10e-5、ウェイト・ディケイ 0.1、バッチサイズ 128、シーケンス長 8192 トークンのコサイン・ラーニング・レート・スケジュールを用いる。SFT を 3 エポック行い、granite.13b.instruct.v1 モデルを得る。

SFT データには、Flan コレクション[26]のサブセット、Dolly[2]の 15K サンプル、Anthropic の有用性と無害性に関する人間のプリファレンス・データ[3]、Instructv3[27]、および要約と対話タスク用に特別に設計された、IBM 独自の合成データセットが含まれる。

granite.13b.instruct.v2 の学習を、2.5 兆トークンに基づいたベースモデルの新しいチェックポイントに対して行った際には、同じ SFT データを用いたが、空白文字や特殊文字へのロバスト性を向上するためにノイズ・オーグメンテーション（ノイズを追加することによるデータ水増し）を利用した。さらに、対話タスクでのパフォーマンスを改善するために（データや計算量の追加はなしで）学習中に埋め込みベクトルにノイズを追加する NEFTune [28]を採用した。

3) Granite.13b.chat.v1 のアライメント：対照ファイン・チューニング（Contrastive Fine-tuning：CFT）は、非尤度学習（unlikelihood based training）[29]に基づくインストラクション・ファイン・チューニングのアプローチであり、granite.13b.chat.v1 のアライメントのために使用した。これは負例のデータ分布で求めた確率にしたがってペナルティを与えると同時に、正例のデータ分布で求める確率を増加させる（図 4 参照）。言い換えれば、LLM が毎回の学習プロンプトに対して、人間の価値観に合わない悪い応答（たとえば有害な応答）を生成することを抑制する一方で、価値観に合った良い応答（たとえば有用な応答）を奨励する。

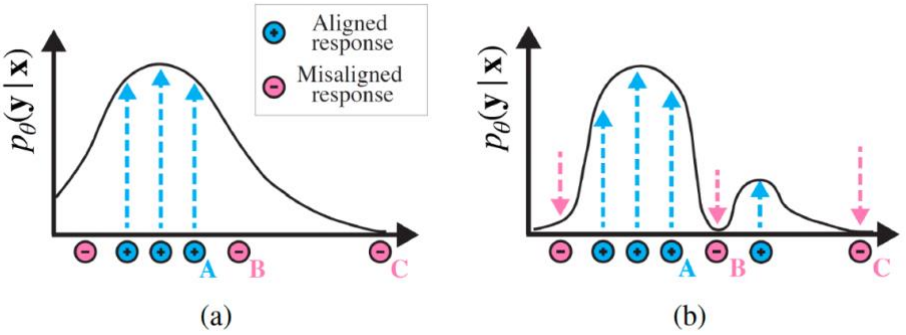


図 4：与えられたプロンプト x に対する LLM の回答 y の分布。(a) SFT は、良い回答の尤度を高めるが、悪い回答に対するコントロールには欠ける。その結果、良い回答とよく似た悪い回答が高い尤度を持つことがある。例 A と B はこれを示しており、A は良い回答、B は悪い回答だが、両者は類似している。(b)私たちのアプローチである対照ファイン・チューニングは、悪い応答に低い尤度を明示的に割り当てることにより、これを緩和する。

CFT では、モデルが 2 つの応答のうちどちらが悪いかを判断するために、両方の応答が同じプロンプトとペアになっている必要がある。しかし、一般に公開されている人間のデモ（実演）・データセットの多くには、同じプロンプトに対して、良い回答と悪い回答がペアになってはいない。そのため、人は疑問に思うかもしれない：「どのようにして良い回答と悪い回答の両方を得ることができるのだろうか？」。負例のデータ分布を得るための簡単なアプローチは、各プロンプトに対して悪い回答を人間に書かせることである。しかし、このようなアプローチはコスト高になる可能性がある。そこで、私たちの研究では、「ネガティブなペルソナ」として振る舞う別の LLM を使用することを提案する。すなわち、この LLM は、悪い（たとえば有害な、あるいは真実でない）回答をする傾向のある架空の人間を模倣するように作る。私たちは、公開されている、悪い人間のデモ・データセットを使って別の LLM をファイン・チューニングすることにより、これを実現する。その結果、プロンプトと人間の（良い）デモ・データセッ

トが与えられれば、このネガティブなペルソナの LLM を用いてプロンプトに対して生成した回答が、悪い回答となり、人間のデモそのものが、良い回答としてペアを作ることができる。
granite.13b.chat.v1 では、granite.13b.instruct.v1 の初期バージョンを、その（ファイン・チューニング元の）別 LLM として使用する。CFT のためのデータセットは、OpenAssist 報酬モデル[3]、Dolly[2]、ProsocialDialog[4]を使ってフィルタリングされた、Anthropic の有用性と無害性に関する人間のプリファレンス・データのペアサンプルである。

CFT ステップの一部として、granite.13b.chat.v1 は、人とエージェントの対話をサポートするために、以下のシステム・プロンプト[3]で動作するように学習された：

> *Below are a series of dialogues between various people and an AI assistant. The AI tries to be helpful, polite, honest, sophisticated, emotionally aware, and humble-but-knowledgeable. The assistant is happy to help with almost anything, and will do its best to understand exactly what is needed. It also tries to avoid giving false or misleading information, and it caveats when it isn't entirely sure about the right answer. Moreover, the assistant prioritizes caution over usefulness, refusing to answer questions that it considers unsafe, immoral, unethical or dangerous.*
>
> *Human:<prompt>*
> *Assistant:*

そのシステム・プロンプトの和訳は以下の通り：

> *以下は、さまざまな人と AI アシスタントとの一連の対話である。AI は、親切で、丁寧で、正直で、洗練されていて、感情に敏感で、謙虚だが知識豊富であろうとする。アシスタントはほとんど何でも喜んで手助けし、何が必要かを正確に理解するために最善を尽くす。また、嘘や誤解を招くような情報を与えないようにし、正しい答えが完全にわからないときには注意を促す。さらに、アシスタントは有用性よりも注意深さを優先し、安全でない、不道徳、非倫理的、もしくは危険であると考えられる質問には答えない。*
>
> *人間：<プロンプト*
> *アシスタント：*

4) Granite.13b.chat.v2 のアライメント：granite.13b.chat.v2 の最新のバージョンでは、生成タスク、それも特に Retrieval-Augmented Generation（RAG）や要約といった、フォーマットが長い生成タスクの品質を改善することにフォーカスした。
生成品質を改善するための我々の戦略には、FLAN データセットといった一般に公開されているインストラクション・チューニング用データセットを使ってファインチューニングすることに伴う限界への対処が含まれていた。FLAN は多様性に富んでいるが、その応答文は簡潔であることが特徴で、実験的にはこのことがモデルのパフォーマンスに改善の余地が生まれる原因であると考えられている [30, 31]。この問題を緩和するために、我々は IBM の f-ORCA 手法で生成された新しい合成データセットを granite.13b.chat.v2 のインストラクション・チューニングに導入した。
f-ORCA 手法は、新しい、例示されたガイドに基づくインコンテキスト・ラーニング（ICL）手法であり、この手法は LLM のインストラクション・チューニングにおいて人間の専門家によるデモンストレーション（模範回答）とよく適合した、高品質で多様性に富んだ応答文を生成することに特化して設計されている。この f-ORCA 手法は、FLAN のようなデータセットに存在する多様なプロンプトに対して優れた応答文を生成する能力においては特に、最近開発された ORCA 手法と類似点を持っているが、いくつかの重要なポイントで異なってもいる。第一に、f-ORCA のアプローチでは ChatGPT や GPT-4 のような、アライメントがとれていて、インストラクション・チューニング

がなされている高コストなブラックボックス・モデルを利用する必要がない。f-ORCA はその代わりに事前学習されていて高品質な応答が出力可能な Falcon-180b モデルだけを利用する。さらに、f-ORCA は、強力な品質コントロール機構として、追加されたフィルター処理（f-UMPBACK）を組み込む。この処理で Falcon-180b モデルは、例示されたガイドに基づく ICL 手法の 1 バリエーションを用いて、監査役としての役割を果たす。これによってモデルがその出力を自分自身で評価・キュレーションし、正確さ、事実に基づく正しさ、自然さと安全さについての高い水準を保つことを可能にする。このプロセス全体の結果としておよそ 30 万サンプルのデータセットが得られる。そこにさらに granite.13b.chat.v1 の学習において有用性を証明した一部のデータを組み合わせて、最終的なデータセットとし、granite.13b.chat.v2 のインストラクション・チューニングに使用される。

granite.13b.chat.v2 では、モデルの安全性をさらに改善するために最近 IBM が発表した RLAIF アルゴリズムである SALMON [32]を適用した。SALMON の目的は、セルフ・アライメント[33]である。すなわち、モデル自体が生成した批評 [34]や自分で改良した出力 [35]を使ってブートストラッピングをするなどして AI モデルにそれ自体を改善させることを主眼としている。SALMON の中心をなしているのは、ガイドに従った報酬モデルである。合成プリファレンス・データで学習したこの報酬モデルは、人間が定義した任意のガイド（規則）に基づいて報酬スコアを計算する。強化学習の学習フェーズにおいてこのガイドを調整するだけで、報酬モデルを用いて挙動の完全なコントロールを得ることができて、その結果、強化学習で学習された方策（ポリシー）の振る舞いに影響を与えることができ、人間のプリファレンス・データをオンラインで収集する必要がなくなる。

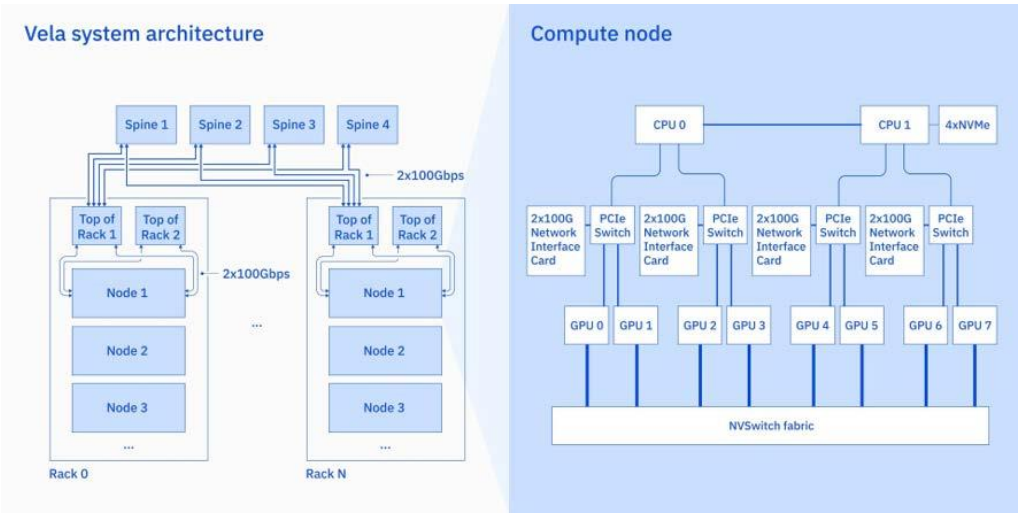
我々は文献 [32]で提案されている手法に従い、40B パラメータのベースモデルから、人間によるプリファレンス・データセットと、ガイドに従う合成プリファレンス・データセットを用いて、インストラクションに従う報酬モデルを学習する。次に、この報酬と、IBM の研究員が定義したガイドを用いた PPO 学習で granite.13b.chat.v2 のアライメントを行う。このガイドの全体は Appendix B に付録した。

granite.13b.chat.v2 の学習における f-ORCA と SALMON の処理ステップで使用したデータセットには、IBM が生成した合成データである HHRLHF [36], sharegpt データセット [37]から取り出した人が作成したプロンプト（chatGPT の応答文は使用しなかった）、FLAN 2022 インストラクション・チューニング・データ・コレクションのうちのフィルターした一部と、OASST [38] が含まれている。

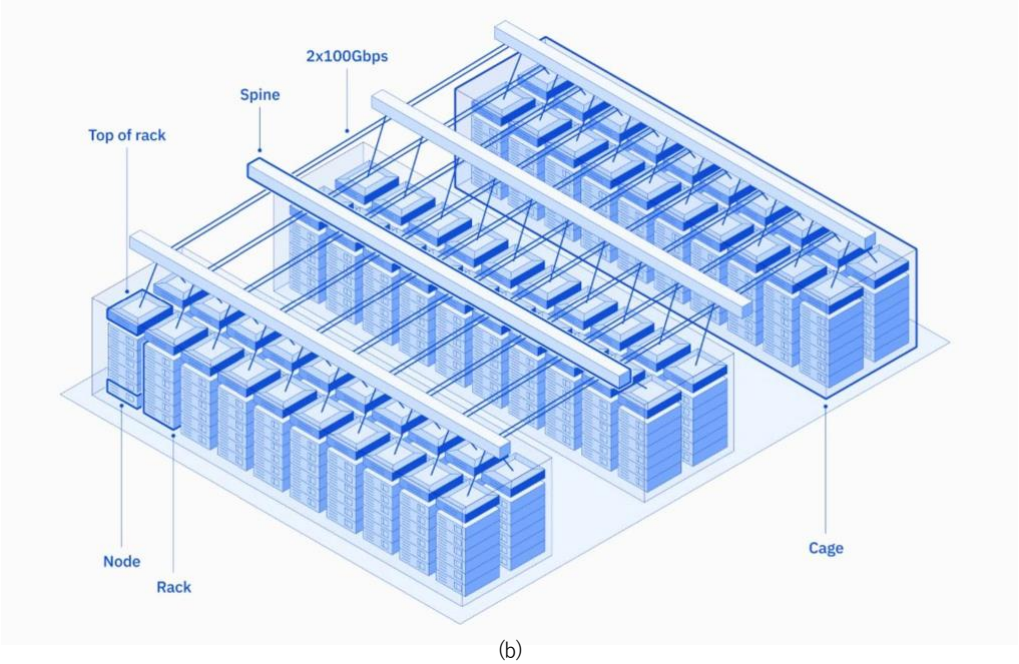
granite.13b.chat.v1 とは異なり、granite.13b.chat.v2 は推論時にシステムプロンプトを必要としない。granite.13b.chat.v2 の完全な評価を予定している本レポートの将来のバージョンにおいては、特定のユースケースのためにシステムプロンプトを使用することがあるかもしれない。

B. 計算

基盤モデル学習のための IBM の主要な計算インフラストラクチャーは、AI スーパーコンピューター Vela[39]である（図 5 を参照）。Vela は、リソースの割り当てに柔軟性を持たせるために、仮想マシンをベースにしたアプローチを採用している。様々な最適化により、いわゆる「仮想マシン税」、すなわち、仮想マシンを用いることによるオーバーヘッドは 5%未満である。各 AI ノードは、8 つの Nvidia A100 GPU カード、96 の vCPU、1.5 TB の DRAM、4×3.2 TB の NVMe ドライブを搭載している。ノードはイーサネットで相互接続されている。各ノードには 2 個の 100 Gbps イーサネット・リンクがある。現在モデルの学習に使用されている Vela インスタンスは、IBM Cloud のワシントン D.C.データセンターに設置されている。今後の Granite モデルは Vela を使用して学習される予定だが、granite.13b のベース・モデルは、Vela インスタンスが完全に立ち上がる前の古いインフラで学習された。granite.13b.v1 は、256 個の A100 GPU を 1,056 時間、120 TFLOPs を使用した。granite.13b.v2 は同じインフラストラクチャーで 120TFLOPS、1,152 時間を使用し、合計 2,208 時間使用ということになった。



(a)



(b)

図 5：AI スーパーコンピューターVela の(a)アーキテクチャー図と、(b)インフラストラクチャー図

C. エネルギー消費と炭素排出量

granite.13b ベース・モデルのエネルギー消費と炭素排出量の推定に使用した方法は以下の通りである。特定の場所 L における、モデル M に関連する炭素排出量 Carbon は次式で与えられる：

$$\mathrm{Carbon}(M,L) = E(M) \times \mathrm{PUE}(L) \times \mathrm{CEF}(L), \tag{1}$$

ここで、E(M)はモデル M の電力消費量、PUE(L)は場所 L における電力使用効率、CEF(L)は場所 L に適用される炭素排出係数である。

情報技術（IT）消費電力 E（M）は、すべての GPU の平均 GPU 使用率を使用して推定される。これは、図 6 に示すように、GPU 使用率は一般的にノード電力と高い相関があるため、AI モデル M の学習に使用される電力を推定

するための近似となる。次に、推定されたノード電力に、学習時間と使用された GPU の数を乗じて、総計算エネルギー消費量 E を計算する。

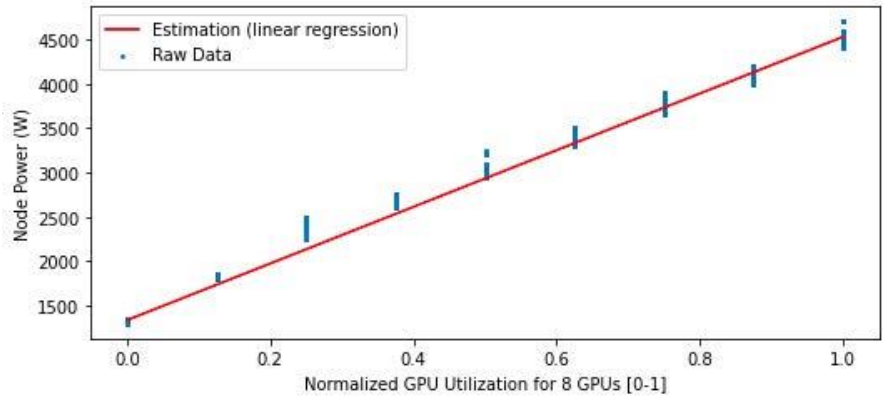


図 6：サーバー（ノード）の電力と正規化された GPU 利用率の関係

電力使用効率 PUE(L)は、データセンターで消費される総電力量（IT とサポート・オーバーヘッド・インフラによる総消費量）の、IT インフラで消費される電力量に対する比率で与えられる。GHG プロトコルのスコープ 2 ガイダンス[40]に従って、ロケーション・ベースの炭素排出係数 CEF(L)を計算する。

この推定方法を granite.13b.v1 のベースモデルに適用すると、153,074.3767kWh のエネルギー消費 E(M)、0.12kg kWh の炭素排出係数 CEF(L)、これは 22.2263995 トンの CO2 換算 Carbon(M,L)と推定され、これは二酸化炭素と、メタンや亜酸化窒素などの他のすべての温室効果ガスを考慮したものである。
granite.13b.v2 ベースモデルのエネルギー消費と炭素排出量はまだ評価中であり、後日このレポートの更新版で公表する。

エネルギーとカーボン・フットプリントを削減するために、多くの緩和戦略を用いることができる。例えば、学習で使用する資源の量は、再生可能エネルギーの利用可能性の関数として調整することができる。あるいは、資源の使用量は一定のエネルギー使用量や排出量の上限値を超えないように制限することができる。

V. テストと評価

このセクションでは、Granite モデルのテストと評価のために取られたアプローチについて説明する。また、同程度の能力レベルの他のいくつかのモデルとの比較とともに、実証的な結果を示す。

A. 基盤モデルの評価フレームワーク

私たちは、モデルの開発ライフサイクルを通じて、包括的な基盤モデル評価フレームワーク（FM-eval）を使用している。FM-eval は、評価ベンチマークを効率的に実行するために、GPU をサポートした RedHat OpenShift（https://www.redhat.com/en/technologies/cloud-computing/openshift）クラスタ上で複数のモデルに対して並列に実行されている。自動化フレームワークは、Eleuther AI の Language Model Evaluation Harness (lm-eval) [41]や Stanford の HELM (Holistic Evaluation Model) [42]のようなラップされた外部フレームワークやコンテナ化された評価フレームワークを実行することができる。FM-eval にタスク、データセット、メトリクスを簡単に追加できるようにするため、我々はオープンソースの Python ライブラリである Unitxt（https://github.com/IBM/unitxt）を開発した。Unitxt は、データセットを定義するための一貫したインターフェースと手法を提供し、これには生のデータセットを LLM が必要とする入力に変換するために必要な前処理、および結果を評価するために使用されるメトリクスが含まれる。

ライフサイクルのさまざまな段階で、さまざまな種類のテストが実行される：

1) 一般知識のベンチマーク（General Knowledge Benchmarks、学習中）
2) HELM ベンチマーク（学習後）
3) 企業ベンチマーク（学習後）

これらの評価はすべて、zero-shot プロンプトと few-shot プロンプトを活用したものである。注釈として述べておくが、zero-shot プロンプトは、既存の LLM を使用し、プロンプトとしてタスクの実行命令のみを提供することで、新しいタスクのテキストを生成する。few-shot プロンプトでは、プロンプトの中でタスクの記述とともに複数の例示を提供する。両アプローチとも、一つの事前学習済みモデルを、コア・パラメーターを固定したまま使うことを可能にする。

具体的な評価は以下の通り。

1) 学習中のための一般知識ベンチマーク：General Knowledge Benchmarks は lm-eval [41]の既存のベンチマークのサブセットを含み、学習中に 1,000 億トークンごとに実行される軽量テストとして使用され、学習の進行に伴ってモデルの知識が進歩していることを検証する。具体的には、lm-eval の以下の 12 個のデータセット（タスクごとに整理されている）である：

- いくつかのドメインの質問応答（boolq、openbookqa、piqa、sciq）；
- 文章補完
- 常識推論（arc easy, arc challenge, copa, hellaswag, winogrande）；
- 機械読解
- 分野横断的な多肢選択コレクション（mmlu）；

私たちの評価フレームワークでは、これらのベンチマークは zero-shot と few-shot の両方の設定で実行される。

2) HELM: 事前学習が完了した後の包括的な評価の一部は Stanford の Holistic Evaluation of Language Models (HELM) Benchmark [42]に依存する。私たちのモデルを評価するために、質問応答、情報検索、要約、感情分析、テキスト分類を含む様々なタスクから構成される 16 の「コアシナリオ」[43]〜[54]を使用する。

3) エンタープライズ評価ベンチマーク：学習が完了したら、さらに IBM が作成したエンタープライズ・ベンチマークでモデルを評価し、お客様に関連性の高いドメインにおけるモデルのパフォーマンスをテストする。そのような意図で、IBM は金融ドメインでモデルを評価するために、11 個の公開された金融ベンチマークをキュレーションし、表 I にまとめた。データ提供元から提供された学習セットとテストセットの分割を、可能な限り評価に使用する。モデルの性能はテストセットに基づいて報告される。テストラベルが公開されていない場合、モデルの性能は検証セットに基づいて報告される。学習セットとテストセットの分割がデータ提供元から提供されていない場合は、データの 20%をテスト用として選択し、残りを学習用として使用する。

few-shot プロンプトで与えるコンテキスト例はすべて学習セットからサンプリングする。モデルに提供される few-shot の例示の数はタスクによって異なり、表 I に示す。今回の評価では、すべてのモデルが同じパラメータと同じ標準的なプロンプトを使用し、タスク記述、chain-of-thought プロンプト[55]、システム・プロンプトは使用しなかった（few-shot プロンプトと zero-shot プロンプトのテクニックとプロンプトの例（https://www.promptingguide.ai/techniques/fewshot）を参照されたい）。Financial Phrasesbank、News Headline、FiQA SA では、プロンプトは BloombergGPT [56]から引用した。

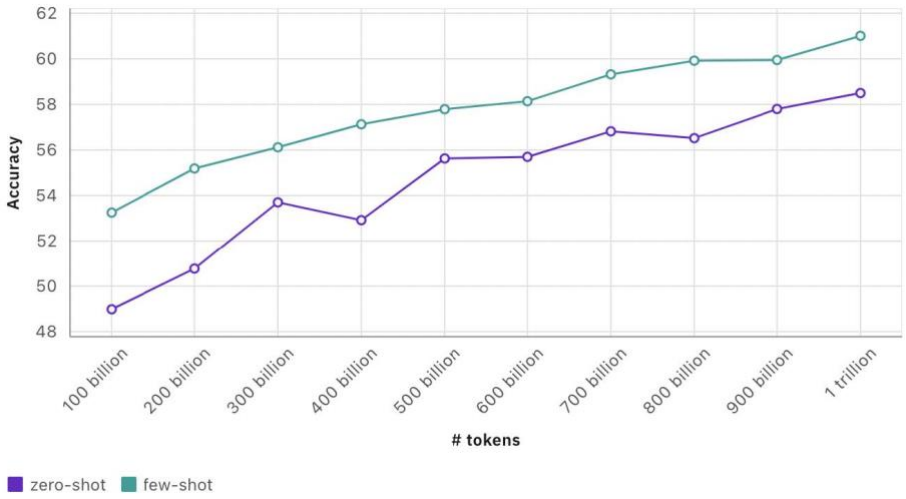


図 7：Granite.13b 学習中の一般知識パフォーマンス

B. Granite モデルの評価と比較

granite.13b.v2 の完全な評価はまだ進行中であり、結果は得られ次第このレポートの更新版で公表される。granite.13b.v1 の評価結果を以下に報告するが、この初期リリースでは学習に 1 兆個のトークンを用いただけであり、これらの評価はすべて予備的なものであることに注意すべきである。

学習中の一般知識ベンチマーク：このセクションでは、軽量な一般知識ベンチマークを活用して、学習中に 1,000 億トークンごとに取得した granite.13b.v1 のベース・モデルのスナップショットと、ファイン・チューニングした granite.13b.instruct.v1 と granite.13b.chat.v1 のバリエーションを評価する。図 7 に可視化され、さらに表 II に詳述されているのは、予想された通り、学習を進めることで 1,000 億トークンごとに granite.13b.v1 の一般知識精度が着実に向上していること、そしてファイン・チューニングされたバリエーション・モデルではさらにパフォーマンスが向上していることである。ただし、granite.13b.chat.v1 のこの評価では、システム・プロンプトは使用されていない。

HELM ベンチマーク：この調査では、HELM の 16 のコア・シナリオで私たちのモデルを網羅的に評価し、v0.2.3 リリース（https://crfm.stanford.edu/helm/latest/?group=core_scenarios）の他のすべてのモデルと比較した。私たちの評価は HELM によって推奨されている通り 2 段階のプロセスで進めた。すなわち、まず各モデルを個別の評価データセットで評価し、次にこれらの結果をシナリオに集約する（表III に詳述）。LLM の公正な比較を容易にするため、モデルのランク付けに HELM の Mean Win Rate（MWR）指標を、シナリオを通して採用する。

図 8 は、granite.13b.instruct.v1、granite.13b.chat.v1、およびすべての v0.2.3 モデルの、モデルサイズと MWR 軸に対する位置関係を示している。ただし、正確なモデルサイズが LLM プロバイダーによって公表されていないモデルは、この可視化から除外されている。

この図は、granite モデルがモデルサイズと HELM 性能の間で望ましいバランスをとっていることを表している。granite.13b.chat.v1 と granite.13b.instruct.v1 は、評価されたすべてのモデルの中で、それぞれ 15 位と 18 位である。　さらに、granite.13b.chat.v1 と granite.13b.instruct.v1 は、それぞれサイズが 170 億以下のパラメータで評価されたモデルの中で、トップ 2 とトップ 3 だった。Cohere Command beta (61 億)だけが、このサイズのカテゴリーでその性能を上回った。これらの結果は、HELM によって評価された他の側面、例えば頑健性や公平性についても同様である。キャリブレーションにおいて、granite.13b.chat.v1 と granite.13b.instruct.v1 は、それぞれ 9 位と

28 位である。

エンタープライズ・ベンチマーク：この評価は、HELM のフレームワークを拡張し、金融サービス領域から公開されている 11 のタスク・データセットを包含することによって実施される。比較対象のベースライン・モデルは、モデル・サイズ、学習データの種別、アクセシビリティー、モデル・チューニングに基づいて選択される。具体的に言えば、Granite モデルは、500 億パラメータ以下のオープンソースモデルの中で最もパフォーマンスが良い GPT-NeoX-20B [6]、FLAN-UL2 [67]、また 70 億から 130 億のパラメータを持つ、公開されている最先端のモデルである LLaMA2 [68]と比較される。

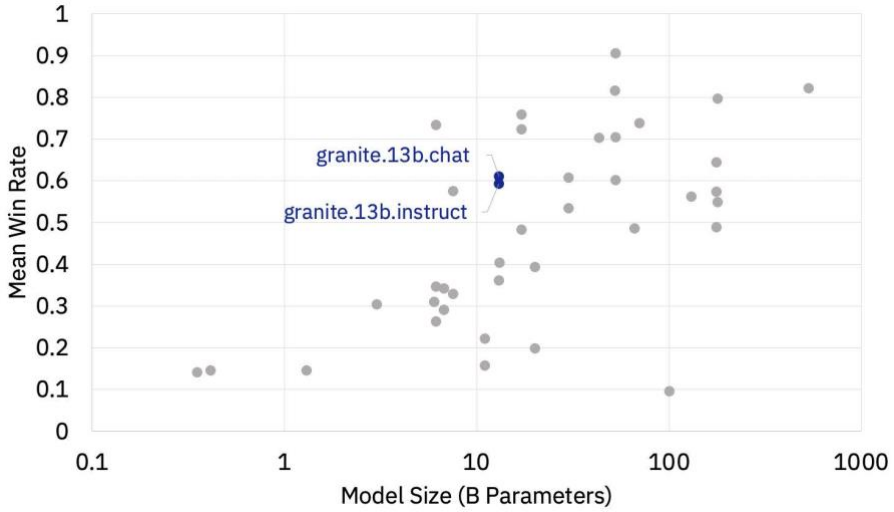


図 8：サイズに対する HELM タスクでのパフォーマンス

表 IV は、11 の金融タスクに関する各モデルの詳細な性能スコアを示している。LLaMA2 モデルは 2T トークンの事前学習データを利用しており、モデルに大きなアドバンテージを与えているため、LLaMA2 モデルの横にアスタリスクを付けている。granite を含む他の評価モデルはすべて 1T トークンの学習データを使っている。Llama2 モデルの半分のデータ量で学習したにもかかわらず、Granite.v1 モデルは各タスクで競争力があり、しばしば Llama2 を上回る。このことは、2T トークンを超える事前学習データで学習される予定の、将来計画されている granite モデルのバージョンにとって良い兆候である。

VI. 社会技術的弊害とリスク

近年、LLM の潜在的な社会技術的弊害やリスクは数多く指摘されている。それには、 誤情報、幻覚、忠実性や事実性の欠如、個人情報の漏洩、著作権で保護されたコンテンツの包含や剽窃、ヘイトスピーチ、有害性、いじめやガスライティングのような人間とコンピュータの相互作用の害、悪意のある利用、敵対的攻撃などが含まれる[69]、[70]。

表 I：ファイナンス・ベンチマークの概要

Task	Task Description	Dataset	Dataset Description	N-shot Prompt	Metric
Sentiment Classification	3 classes	Financial Phrasebank [57]	Financial news categorised by sentiment	5-shot	Weighted F1
	2 classes	Earnings Call Transcripts [58]	Earnings call transcripts, the related stock prices and the sector index in terms of volume	5-shot	Weighted F1
Classification	9 classes	News Headline [59]	The gold commodity news annotated into various dimensions	5-shot	Weighted F1
Named Entity Recognition	4 numerical entities	Credit Risk Assessment (NER) [60]	Eight financial agreements (totalling 54,256 words) from SEC filings were manually annotated for entity types: location, organization person and miscellaneous	20-shot	Entity F-1
	4522 numerical entities	KPI-Edgar [61]	A dataset for Joint Named Entity Recognition and Relation Extraction building on financial reports uploaded to the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, where the main objective is to extract Key Performance Indicators (KPIs) from financial documents and link them to their numerical values and other attributes	20-shot	Modified Adjusted F1
	139 numerical entities	FiNER-139 [62]	1.1M sentences annotated with extensive Business Reporting Language (XBRL) tags extracted from annual and quarterly reports of publicly-traded companies in the US, focusing on numeric tokens, with the correct tag depending mostly on context, not the token itself.	10-shot	Entity F1
Question Answering	Document relevance ranking	Opinion-based QA (FiQA) [63]	Text documents from different financial data sources (microblogs, reports, news) for ranking document relevance based on opinionated questions, targeting mined opinions and their respective entities, aspects, sentiment polarity and opinion holder.	5-shot	RR@10
	3 classes	Sentiment Analysis (FiQA SA) [63]	Text instances in the financial domain (microblog message, news statement or headline) for detecting the target aspects which are mentioned in the text (from a pre-defined list of aspect classes) and predict the sentiment score for each of the mentioned targets.	5-shot	Weighted F1
	Ranking	Insurance QA [64]	Questions from real world users and answers with high quality composed by professionals with deep domain knowledge collected from the website Insurance Library [4]	5-shot	RR@5
	Exact value match	Chain of Numeric Reasoning (ConvFinQA) [65]	Multi-turn conversational finance question answering data for exploring the chain of numerical reasoning	1-shot	Accuracy
Summarization	Long documents	Financial text summarization (EDT) [66]	303893 news articles range from March 2020 to May 2021 for abstractive text summarization	5-shot	Rouge-L

表 II：Granite.13b 学習中の一般知識パフォーマンス

Model	Tokens (B)	Avg Accuracy (Zero-Shot)	Avg Accuracy (Few-Shot)
granite.13b (base)	100	49.0	53.3
granite.13b (base)	200	50.8	55.2
granite.13b (base)	300	53.7	56.1
granite.13b (base)	400	52.9	57.1
granite.13b (base)	500	55.6	57.8
granite.13b (base)	600	55.7	58.1
granite.13b (base)	700	56.8	59.3
granite.13b (base)	800	56.5	59.9
granite.13b (base)	900	57.8	60.0
granite.13b (base)	**1000**	**58.5**	**61.0**
granite.13b.instruct.v1	**1000**	**59.3**	**61.5**
granite.13b.chat.v1	**1000**	**61.2**	**62.6**

表 III：コア・シナリオ中のサブ・シナリオごとの HELM 結果

Model (Metric)	MMLU (EM)	BoolQ (EM)	NarrativeQA (F1)	NaturalQuestions closed-book (F1)	NaturalQuestions open-book (F1)	QuAC (F1)	HellaSwag (EM)	OpenbookQA (EM)	TruthfulQA (EM)	MS MARCO (RR@10)	MS MARCO (TREC) (NDCG@10)	CNN/DailyMail (ROUGE-2)	XSUM (ROUGE-2)	IMDB (EM)	CivilComments (EM)	RAFT (EM)
granite.13b.instruct.v1	0.377	0.809	0.668	0.188	0.659	0.373	0.338	0.296	0.203	0.431	0.638	0.135	0.11	0.953	0.637	0.693
granite.13b.chat.v1	0.378	0.776	0.698	0.212	0.684	0.391	0.305	0.276	0.208	0.396	0.634	0.14	0.115	0.948	0.6	0.709

表 IV：タスクごとのファイナンス・ベンチマーク評価結果

	Financial Phrase-bank	Earnings Call Tran-scripts	News Head-line	Credit Risk As-sessment	KPI-Edgar	FiNER-139	FiQA - Opin-ion	Insurance QA	FiQA SA	ConFinQA	Summarization
Metrics	Weighted F1	Weighted F1	Weighted F1	Entity F1	Adj F1	Entity F1	RR@10	RR@5	Weighted F1	Accuracy	Rouge-L
granite.13b.v1 (base)	0.306	0.443	0.811	**0.477**	0.344	0.699	0.400	0.169	0.780	**0.365**	0.173
granite.13b.instruct.v	0.590	0.443	0.764	0.407	0.281	0.699	0.658	0.605	0.590	0.346	0.323
granite.13b.chat.v1	0.714	0.443	0.779	0.361	0.290	0.746	0.624	0.422	0.758	0.334	**0.376**
llama2.7b*	0.244	0.486	0.752	0.408	0.419	0.660	0.617	0.255	0.744	0.233	0.195
llama2.7b.chat*	**0.758**	**0.677**	**0.829**	0.458	0.450	0.626	0.644	0.443	0.693	0.254	0.345
llama2.13b*	0.378	0.410	0.584	0.467	0.463	0.689	0.560	0.539	0.800	0.226	0.252
llama2.13b.chat*	0.608	0.572	0.744	0.445	**0.538**	0.671	0.625	0.227	**0.849**	0.261	0.269
gpt-neox-20b	0.561	0.318	0.630	0.469	0.308	**0.774**	0.496	0.163	0.771	0.266	0.205
flan-ul2	0.240	0.318	**0.829**	0.394	0.011	0.446	**0.793**	**0.747**	0.811	0.254	0.310

表 V では、IBM AI 倫理委員会がまとめたリスクのカタログを示す。この委員会は、IBM コーポレーション全体を通じて、倫理的で、責任感があり、信頼できる AI の文化を支援することを目的として、AI 倫理のビジョンと戦略を定義する主要な分野横断的組織である [71], [72]。この表はいくつかの軸にわたって構成されている [73]：

- そのリスクは、基盤モデルのためのデータまたはその他の入力に起因するのか、基盤モデルの生成出力に起因するのか、あるいはその他の懸念事項に起因するものなのか。
- そのリスクは、モデルの学習やファイン・チューニング中、もしくは推論の際に生じるか、あるいはガバナンス、法令遵守、社会的影響など、より広範な考慮事項において生じるのか。
- そのリスクは、例えば、公平性、堅牢性、知的財産、悪用など、ハイレベルのグループのどれに属するのか。
- そのリスクは新しいものか、増幅されたものなのか。「Traditional（従来リスク）」は、以前の AI モデルにも存在し、基盤モデルにも引き続き存在するリスクである。「Amplified（増幅されたリスク）」は、以前の AI モデルから知られていたものであるが、生成能力によって基盤モデルに関して悪化したものである。「New（新しいリスク）」は、その生成能力によって基盤モデル固有に生まれた新たなリスクである。

granite.13b.instruct および granite.13b.chat モデルの作成とリリースにあたって、私たちはリスクのいくつかに次に述べるような方法で対処した。ヘイトスピーチ、暴力的な表現、下品な表現のブロックリストとフィルタリングを含む、IBM の事前学習データセットのデータ・ガバナンス・プロセスは、いくつかの知的財産と誤用のリスクを軽減した。公平性の観点については、セクション III で説明したデータ前処理パイプラインへの追加要素として、宗教、性別、人種、社会的烙印、年齢、政治的イデオロギーによる文書のアノテーションがある。私たちはこれらの観点でキーワードリストを作成し、キーワードマッチングを使って文章にアノテーションを付けている。このアノテーションは、過小評価グループ（見逃されているマイノリティー）や過大評価グループ（その逆）を特定するために使われる。HAP フィルタリングを過度に積極的に行わず、グループに関するフィルタリングも行っていないのは、中傷を撤回したり、疎外されたアイデンティティーを肯定的に表現したりする学習データが得られなくなることや、他の意図しない方法で事前訓練データセットを歪める可能性があるからである[74]。

ファイン・チューニングを通じて、私たちは、誤用や価値観との不一致によるリスクの一部の側面を軽減することを目的として、向社会的で有害性の少ないモデルの挙動を実現しようとしてきた。しかし、最大の社会技術的リスクのひとつは、ファイン・チューニングのために使用したデータセット（あるいは将来使用する可能性のある他の既存のデータセット）が、個々の目的のために Granite モデルをデプロイする人々や組織のニーズ、欲求、願望に合致していると信じる私たち自身の思い上がりだろう。すべての企業には、法律、社会規範、業界標準、市場要求、またはアーキテクチャー要件 [75]のいずれに由来するものであれ、準拠すべき独自の規制がある。私たちは企業に、例えば watsonx プラットフォームのツールを使うことで、独自の価値観に従ってモデルをパーソナライズする権限が（範囲内で）与えられるべきだと考えている [76]。

さらに、FM-eval を通じて、いくつかのリスク次元をカバーするベンチマーク・データセットを用いて、Granite モデルをテストした。しかし、ベンチマークでの評価は、社会技術的弊害を明らかにするためのアプローチとして限定的である [77]。企業は、Granite モデルを自社の価値観にさらに整合させた後、社会文化的・生活的経験の異なるメンバーからなるレッド・チームを結成し、正確なユースケースの文脈の中で、ほかの弊害と望ましくない LLM の行動を見つけるべきである [78]。

VII. 利用ポリシーと文書化

A. 機械生成コンテンツ

IBM のライセンス条件は、IBM モデルを使用する下流アプリケーションとサービスを管理する。それに加えて、IBM は、IBM モデルのユーザーが下流アプリケーションやサービスを開発・提供する際に従うことが求められるガイドラインやプラクティスを記載した AUP（Acceptable Use Provision）を設定している。AUP は、AI モデルの許容される使用を規定し、必要に応じてこれらのモデルのライセンスを終了する権利を IBM に付与する。

B. 欧州連合特有の規制

IBM モデルのライセンス条件には、特定の国での下流アプリケーションおよびサービスの展開に固有のガイドラインと慣行が記載された利用規定（AUP）が追加されている。

C. 下流のドキュメンテーション

IBM は、事前学習済みモデルを下流で使用するために、以下の文書を提供している：

- 利用条件
- 製品ドキュメント
- 本レポートのようなテクニカル・レポート

これらの情報は、IBM が法的および倫理的な要件を遵守するだけでなく、これらのモデルの使用者が自ら設定した義務を遵守できるように設計されている。

1) 利用規約：watsonx プラットフォームの最新の利用規約は、以下の URL でご覧いただける。
https://www.ibm.com/support/customer/csol/terms/?id=i126-6883
2) 製品ドキュメント：IBM Granite モデルは現在、IBM の watsonx プラットフォームを通じて利用可能である。watsonx の一部として、各 Granite モデルには、モデルの主要な事実と出所を詳述したモデル・カードが付属している。

VIII. 結論

このテクニカルレポートでは、企業の生成 AI アプリケーション向けに設計された IBM の Granite 基盤モデル・シリーズを紹介した。IBM の倫理とガバナンスのフレームワークは、これらのモデルが作成され、利用可能であるコンテキストを提供する。透明で責任ある AI に対する IBM のコミットメントに沿い、正確なデータセット、前処理ステップ、学習インフラ、エネルギー消費、およびモデル開発のライフサイクルを通じて使用されるテスト/評価手法に関する説明を示した。

表 V：社会技術的弊害とリスク

Source	Phase	Group	Risk	Indicator
Input	Training and Tuning	Fairness	Bias	Amplified
Input	Training and Tuning	Robustness	False samples	Traditional
Input	Training and Tuning	Value Alignment	Undesirable output for retraining purposes	New
Input	Training and Tuning	Data Laws	Legal restrictions on moving or using data	Traditional
Input	Training and Tuning	Intellectual Property	Copyright and other IP issues with content	Amplified
Input	Training and Tuning	Transparency	Disclose data collected, who has access, how stored, how it will be used	Amplified
Input	Training and Tuning	Privacy	Inclusion or presence of SPI or PII	Traditional
Input	Training and Tuning	Privacy	Provide data subject rights (e.g., opt-out)	Amplified
Input	Inference	Privacy	Disclose PII or SPI as part of prompt to model	New
Input	Inference	Intellectual Property	Disclose copyright or other IP information as part of prompt to model	New
Input	Inference	Robustness	Vulnerabilities to adversarial attacks like evasion (create incorrect model output by modifying data sent to train model)	Amplified
Input	Inference	Robustness	Vulnerabilities to adversarial attacks like prompt injection (force different output), prompt leaking (disclose system prompt),or jailbreaking (avoid guardrails)	New
Output	Inference	Fairness	Bias in generated content	New
Output	Inference	Fairness	Performance disparity across individuals or groups	Traditional
Output	Inference	Intellectual property	Copyright infringement, compliance with open source license agreements	New
Output	Inference	Value alignment	Hallucination (generation of false content)	New
Output	Inference	Value alignment	Toxic, hateful, abusive, and aggressive output	New
Output	Inference	Misuse	Spread disinformation (deliberate creation of misleading information	Amplified
Output	Inference	Misuse	Generate toxic, hateful, abusive, and aggressive content	New
Output	Inference	Misuse	Nonconsual use of people's likeness (deepfakes)	Amplified
Output	Inference	Misuse	Dangerous use (e.g., creating plans to develop weapons or malware)	New
Output	Inference	Misuse	Deceptive use of generated content (e.g., intentional nondisclosure of AI generated content)	New
Output	Inference	Harmful code generation	Execution of harmful generated code	New
Output	Inference	Privacy	Expose PI or SPI in generated content	New
Output	Inference	Explainability	Challenges in explaining the generated output	New
Output	Inference	Traceability	Challenges in identifying source and facts for generated output	New
Other	Governance	Transparency	Document data and model details, purpose, potential use and harms	Traditional
Other	Governance	Accountability	Identify responsibility for misaligned output along AI lifecycle and value chain	Amplified
Other	Legal compliance	Intellectual property	Determine creator of downstream models	New
Other	Legal compliance	Intellectual property	Determine creator of open source foundation models	New
Other	Legal compliance	Intellectual property	Determine owner of AI-generated content	New
Other	Legal compliance	Intellectual property	Uncertainty about IP rights related to generated content	New
Other	Legal compliance	Legal uncertainty	Determine downstream obligations	Amplified
Other	Societal impact	Impact on jobs	Human displacement (AI induced job loss)	Amplified
Other	Societal Impact	Human dignity	Human exploitation (ghost work in training), poor working conditions, lack of healthcare, unfair compensation	Amplified
Other	Societal Impact	Environment	Increased carbon emission (high energy requirements for training and operation)	Amplified
Other	Societal Impact	Diversity and inclusion	Homogenizing culture and thoughts	New
Other	Societal Impact	Human agency	Misinformation and disinformation generated by foundation models	Amplified
Other	Societal Impact	Impact on education	Bypass learning process, plagiarism	New

私たちは Granite シリーズの開発を複数の方向へ続けている。この Granite の初期リリースでは英語しかサポートしていないが、将来的には複数の自然言語でモデルを学習させる予定である。これと並行して、HAP アノテーションも改良し、追加言語向けに拡張する。さらに、コードや業界固有のコンテンツなど、他のデータ種別用の Granite モデルも開発中である。モデルの安全性評価の面では、包括的なレッド・チーミング・フレームワークを開発中である。そこでは敵対的なプロンプトを用いて、HAP、バイアスと社会的烙印、事実の正しさ、有害なトピックやそれだけに終わらない、さまざまな領域にわたってモデルをテストする。

私たちは、個人を特定できる情報が含まれているかどうかや、会話性があるかどうかで文書をスコアリングするなど、IBM のキュレーションした事前学習データセットのための追加データ・アノテーションを開発し続けている[79]、[80]。私たちは、エネルギーとカーボン・フットプリントの推定測定ではなく、正確な測定を得られるように計算インフラストラクチャーの計装に取り組んでいる[81]。最後に、望ましくないバイアスを軽減するためのさまざまな手法の適用を検討している [82]-[84]。

参考文献

[1] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” 2021.

[2] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin, “Free Dolly: Introducing the world’s first truly open instruction-tuned LLM,” https://www.databricks.com/blog/2023/04/12/dolly-first-opencommercially-viable-instruction-tuned-llm , Apr. 2023.

[3] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan et al., “Training a helpful and harmless assistant with reinforcement learning from human feedback,” arXiv preprint arXiv:2204.05862, 2022.

[4] H. Kim, Y. Yu, L. Jiang, X. Lu, D. Khashabi, G. Kim, Y. Choi, and M. Sap, “ProsocialDialog: A prosocial backbone for conversational agents,” in Proc. Conf. Empir. Meth. Nat. Lang. Proc., Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 4005–4029.

[5] D. D. Cox, “Introducing the technology behind watsonx. ai, IBM’s AI and data platform for enterprise,” https://www.ibm.com/blog/introducing-the-technology-behind-watsonxai , May 2023.

[6] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, “Gpt-neox-20b: An open-source autoregressive language model,” 2022.

[7] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, “The Pile: An 800gb dataset of diverse text for language modeling,” arXiv preprint arXiv:2101.00027, 2020.

[8] https://huggingface.co/datasets/c4 .

[9] K. Schaul, S. Y. Chen, and N. Tiku, “Inside the secret list of websites that make AI like ChatGPT sound smart,” Washington Post, Apr. 2023.

[10] IBM Corporation. Watson Natural Language Processing library. [Online]. Available:

https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/watson-nlp.html?context=cpdaas
[11] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," CoRR, vol. abs/1903.08983, 2019. [Online]. Available: http://arxiv.org/abs/1903.08983
[12] T. Caselli, V. Basile, J. Mitrovi´c, I. Kartoziya, and M. Granitzer, "I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language," in Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, May 2020, pp. 6193–6202. [Online]. Available: https://aclanthology.org/2020.lrec-1.760
[13] A. Capozzi, M. Lai, V. Basile, C. Musto, M. Polignano, F. Poletto, M. Sanguinetti, C. Bosco, V. Patti, G. Ruffo, G. Semeraro, and M. Stranisci, "Computational linguistics against hate: Hate speech detection and visualization on social media in the "contro l'odio" project," 11 2019.
[14] T. Caselli, V. Basile, J. Mitrovic, and M. Granitzer, "Hatebert: Retraining BERT for abusive language detection in english," CoRR, vol. abs/2010.12472, 2020. [Online]. Available: https://arxiv.org/abs/2010.12472
[15] J. W. Rae et al., "Scaling Language Models: Methods, Analysis & Insights from Training Gopher," 2022. [Online]. Available: https://arxiv.org/abs/2112.11446
[16] Kenneth Heafield. (2011) KenLM: Faster and smaller language model queries. [Online]. Available: https://kheafield.com/papers/avenue/kenlm.pdf
[17] kenml GitHub source code repository. [Online]. Available: https://github.com/kpu/kenlm
[18] Office of the United States Trade Representative (USTR). (2022) 2022 Review of Notorious Markets for Counterfeiting and Piracy. [Online]. Available: https://ustr.gov/sites/default/files/2023-01/2022%20Notorious%20Markets%20List%20(final).pdf
[19] R. Li, L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. Marone, C. Akiki, J. Li, J. Chim et al., "Starcoder: may the source be with you!" arXiv preprint arXiv:2305.06161, 2023.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
[21] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," arXiv preprint arXiv:1606.08415, 2016.
[22] N. Shazeer, "Fast transformer decoding: One write-head is all you need," arXiv preprint arXiv:1911.02150, 2019.
[23] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. R´e, "Flashattention: Fast and memory-efficient exact attention with io-awareness," Advances in Neural Information Processing Systems, vol. 35, pp. 16 344–16 359, 2022.
[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[25] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," arXiv preprint arXiv:2307.08691, 2023.
[26] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei et al., "The

flan collection: Designing data and methods for effective instruction tuning," arXiv preprint arXiv:2301.13688, 2023.
[27] (2023) Mpt-30b: Raising the bar for open-source foundation models. [Online]. Available: https://www.mosaicml.com/blog/mpt-30b
[28] N. Jain, P.-y. Chiang, Y. Wen, J. Kirchenbauer, H.-M. Chu, G. Somepalli, B. R. Bartoldson, B. Kailkhura, A. Schwarzschild, A. Saha et al., "Neftune: Noisy embeddings improve instruction finetuning," arXiv preprint arXiv:2310.05914, 2023.
[29] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," arXiv preprint arXiv:1908.04319, 2019.
[29] T. Gershon, S. Seelam, J. Jubran, E. Gampel, and D. Thorstensen, "Why we built an AI supercomputer in the cloud," https://research.ibm.com/blog/AI-supercomputer-Vela-GPU-cluster , Feb. 2023.
[30] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah, "Orca: Progressive learning from complex explanation traces of gpt-4," arXiv preprint arXiv:2306.02707, 2023.
[31] Y. Wang, H. Ivison, P. Dasigi, J. Hessel, T. Khot, K. R. Chandu, D. Wadden, K. MacMillan, N. A. Smith, I. Beltagy et al., "How far can camels go? exploring the state of instruction tuning on open resources," arXiv preprint arXiv:2306.04751, 2023.
[32] Z. Sun, Y. Shen, H. Zhang, Q. Zhou, Z. Chen, D. Cox, Y. Yang, and C. Gan, "Salmon: Self-alignment with principle-following reward models," arXiv preprint arXiv:2310.05910, 2023.
[33] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon et al., "Constitutional ai: Harmlessness from ai feedback," arXiv preprint arXiv:2212.08073, 2022.
[34] Y. Fu, H. Peng, T. Khot, and M. Lapata, "Improving language model negotiation with self-play and in-context learning from ai feedback," arXiv preprint arXiv:2305.10142, 2023.
[35] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language model with self generated instructions," arXiv preprint arXiv:2212.10560, 2022.
[36] D. La Barbera, K. Roitero, and S. Mizzaro, "Combining human intelligence and machine learning for fact-checking: Towards a hybrid human-in-the-loop framework," Intelligenza Artificiale, no. Preprint, pp. 1–10.
[37] R. AI, "Ryokoai/sharegpt52k · datasets at hugging face." [Online]. Available: https://huggingface.co/datasets/RyokoAI/ShareGPT52K
[38] A. K¨opf, Y. Kilcher, D. von R¨utte, S. Anagnostidis, Z.-R. Tam, K. Stevens, A. Barhoum, N. M. Duc, O. Stanley, R. Nagyfi et al., "Openassistant conversations–democratizing large language model alignment," arXiv preprint arXiv:2304.07327, 2023.
[40] Mary Sotos. (2015) GHG Protocol Scope 2 Guidance. [Online]. Available: https://ghgprotocol.org/sites/default/files/ghgp/standards/Scope%202%20Guidance_Final_0.pdf
[41] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, "A framework for few-shot language model evaluation," Sep. 2021. [Online]. Available:

https://doi.org/10.5281/zenodo.5371628
[42] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. R´e, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda, "Holistic evaluation of language models," 2022.
[43] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. X. Song, and J. Steinhardt, "Measuring massive multitask language understanding," ArXiv, vol. abs/2009.03300, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221516475
[44] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 2924–2936.
[45] T. Ko˘cisk´y, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette, "The NarrativeQA reading comprehension challenge," Transactions of the Association for Computational Linguistics, vol. 6, pp. 317–328, 2018. [Online]. Available: https://aclanthology.org/Q18-1023
[46] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "HellaSwag: Can a machine really finish your sentence?" in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4791–4800. [Online]. Available: https://aclanthology.org/P19-1472
[47] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2381–2391. [Online]. Available: https://aclanthology.org/D18-1260
[48] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," Transactions of the Association for Computational Linguistics, vol. 7, pp. 452–466, 2019. [Online]. Available: https://aclanthology.org/Q19-1026
[49] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015
[50] D. F. Campos, T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, and B. Mitra, "Ms marco: A human generated machine reading comprehension dataset," ArXiv, vol. abs/1611.09268, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:1289517
[51] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware

convolutional neural networks for extreme summarization," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1797–1807. [Online]. Available: https://aclanthology.org/D18-1206

[52] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. [Online]. Available: https://aclanthology.org/2022. acl-long.229

[53] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083. [Online]. Available: https://aclanthology.org/P17-1099

[54] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer, "QuAC: Question answering in context," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2174–2184. [Online]. Available: https://aclanthology.org/D18-1241

[55] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[56] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," 2023.

[57] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," Journal of the Association for Information Science and Technology, vol. 65, 2014.

[58] D. Roozen and F. Lelli, "Stock values and earnings call transcripts: a sentiment analysis," Preprints 2021, 2021020424, 2021. [Online]. Available: https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/TJE0D0

[59] A. Sinha and T. Khandait, "Impact of news on the commodity market: Dataset and results," 2020.

[60] J. C. Salinas Alvarado, K. Verspoor, and T. Baldwin, "Domain adaption of named entity recognition to support credit risk assessment," in Proceedings of the Australasian Language Technology Association Workshop 2015, Parramatta, Australia, Dec. 2015, pp. 84–90. [Online]. Available: https://aclanthology.org/U15-1010

[61] T. Deußer, S. M. Ali, L. Hillebrand, D. Nurchalifah, B. Jacob, C. Bauckhage, and R. Sifa, "KPI-EDGAR: A novel dataset and accompanying metric for relation extraction from financial documents," in 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, dec 2022. [Online]. Available: https://doi.org/10.1109%2Ficmla55696.2022.00254

[62] L. Loukas, M. Fergadiotis, I. Chalkidis, E. Spyropoulou, P. Malakasiotis, I. Androutsopoulos, and P. George, "Finer: Financial numeric entity recognition for xbrl tagging," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022). Association for Computational Linguistics, 2022. [Online]. Available: https://arxiv.org/abs/2203.06482

[63] Https://sites.google.com/view/fiqa/home .
[64] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou, “Applying deep learning to answer selection: A study and an open task,” 2015.
[65] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah, and W. Y. Wang, “Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering,” 2022.
[66] Z. Zhou, L. Ma, and H. Liu, “Trade the event: Corporate events detection for news-based event-driven trading,” 2021.
[67] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, H. S. Zheng, D. Zhou, N. Houlsby, and D. Metzler, “Ul2: Unifying language learning paradigms,” 2023.
[68] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozi`ere, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “LLaMA: Open and efficient foundation language models,” arXiv:2302.13971, 2023.
[69] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel, “Taxonomy of risks posed by language models,” in Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 214–229.
[70] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla, J. Gallegos, A. Smart, E. Garcia, and G. Virk, “Sociotechnical harms: Scoping a taxonomy for harm reduction,” arXiv preprint arXiv:2210.05791, 2022.
[71] IBM Corporation. IBM AI Ethics. [Online]. Available: https://www.ibm.com/impact/ai-ethics
[72] B. Green, D. Heider, K. Firth-Butterfield, and D. Lim, “Responsible use of technology: The IBM case study,” World Economic Forum, White Paper, Sep. 2021.
[73] “Foundation models: Opportunities, risks and mitigations,” IBM AI Ethics Board, Tech. Rep., Jul. 2023.
[74] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.
[75] L. Lessig, “The new chicago school,” The Journal of Legal Studies, vol. 27, no. S2, pp. 661–691, 1998.
[76] H. R. Kirk, B. Vidgen, P. R¨ottger, and S. A. Hale, “Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback,” arXiv preprint arXiv:2303.05453, 2023.
[77] I. D. Raji, E. Denton, E. M. Bender, A. Hanna, and A. Paullada, “AI and the everything in the whole wide world benchmark,” in Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.
[78] S. Fazelpour and M. De-Arteaga, “Diversity in sociotechnical machine learning systems,” Big Data & Society, vol. 9, no. 1, p. 20539517221082027, 2022.
[79] IBM Corporation. IBM Natural Conversation Framework. [Online]. Available:

https://ibm.biz/natconv
[80] R. J. Moore, S. An, and G.-J. Ren, “The IBM natural conversation framework: a new paradigm for conversational UX design,” Human Computer Interaction, vol. 38, no. 3-4, pp. 168–193, 2023. [Online]. Available: https://doi.org/10.1080/07370024.2022.2081571
[81] M. Amaral, H. Chen, T. Chiba, R. Nakazawa, S. Choochotkaew, E. K. Lee, and T. Eilam, “Kepler: A framework to calculate the energy consumption of containerized applications,” in IEEE International Conference on Cloud Computing, 2023.
[82] P. Sattigeri, S. Ghosh, I. Padhi, P. Dognin, and K. R. Varshney, “Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting,” Advances in Neural Information Processing Systems, vol. 35, pp. 35 894–35 906, 2022.
[83] G. Zhang, Y. Zhang, Y. Zhang, W. Fan, Q. Li, S. Liu, and S. Chang, “Fairness reprogramming,” Advances in Neural Information Processing Systems, vol. 35, pp. 34 347–34 362, 2022.
[74] S. Basu, P. Sattigeri, K. N. Ramamurthy, V. Chenthamarakshan, K. R. Varshney, L. R. Varshney, and P. Das, “Equi-tuning: Group equivariant fine-tuning of pretrained models,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 6, 2023, pp. 6788–6796.

APPENDIX A
英語版リリースノート

2023 年 9 月 15 日
- 本レポートの最初のバージョンの公表。

2023 年 11 月 7 日
- FiQA （Opinion and Insurance QA）尺度の新しい値によって表 IV を更新。HELM のランキング尺度の手順に見つかったバグの修正後、新しい数値が計算された。さらに、HELM ランキング尺度の修正が実装された後、Oasst-sft-pythia-12b について再計算する方法が評価チームにはなかったため、ベンチマークから一時的に削除。
- 全体にわたって幾つかのタイポと文法の修正によって更新。

2023 年 11 月 30 日
- granite.13b.v2 についての新しい情報についてレポート全体を更新。評価結果についてはまだ進行中であり、後日、本レポートの更新版によって公開される予定。
- 著作権のあるコンテンツについての注意書きの文言を明確性のために更新。

APPENDIX B
SALMON の処理で使われるガイド（の和訳）

1) 正直で正確であること: AI は、信頼でき事実に基づいた情報を提供し、知識の範囲や限界についてはつまびらかにしなければならない。
2) 倫理的であること: AI は、攻撃的、差別的、また害のあるコンテンツを生成してはならず、リスクのある活動に関わったりそれを支持したりすべきではない。

3) 教育的で魅力的であること： AI の回答は正確で、関連性があり、最新の情報によって裏付けられ、ユーザーを惹きつけながら教育的であるべきである。
4) 創造的であること： AI は、詩、物語、コード、エッセイ、歌、パロディ、翻訳などのオリジナルなコンテンツを生成することに長けていなければならない。
5) 多言語：AI は、たとえばクエリが中国語の場合、中国語で返答するなど、ユーザーが使用する言語で会話すべきである。
6) 包括的であること： 情報検索タスクの場合、AI は広範に関連性の高い詳細情報を提供し、完全で細部に及んだ応答を提供すべきである。また、物議を醸す話題を扱う場合は、多様な視点から公平かつ幅広い論点を提示するべきである。
7) 自然言語： AI は、繰り返しやぎこちない表現を避け、多様で自然な言葉で応答するべきである。
8）一貫した推論：AI は、自己矛盾を含まないように、明確で論理的な回答を提供するべきである。
9) 数値に敏感であること： AI は
指示された数値仕様が慎重に遵守されていることを確認すること。
数値計算に誤りがないようにすること。
を避ける。
10) 分析的構造： 情報分析タスクの場合、AI は応答文を要約で始めた後、それぞれ徹底的な分析で強調したキーポイントを多数続けるべきである。
11) 生き生きとしていること： AI は生き生きとしたエネルギッシュな言葉を使い、すべてのインタラクションを生き生きとダイナミックにすることでユーザーを惹きつけるべきである。
12) プライバシー保護： AI は個人を特定できる情報（PII）や外部 URL を応答文中に生成しないこと。
13) 誠実さ： AI は虚偽の情報を共有することを避けるべきである。
質問が意味をなさない場合、あるいは事実にそぐわない場合、AI は正しくない回答をする代わりに理由を説明すべきである。
14) スタンドアロン： AI は、URL、画像、動画などの外部ソースとの相互作用を避け、テキストベースの独立したシステムとして機能しなければならない。