Granite Foundation Models

IBM Research, Updated November 30th, 2023

***Abstract*—We introduce the Granite series of decoder-only foundation models for generative artificial intelligence (AI) tasks that are ready for enterprise use. We report on the architecture, capabilities, underlying data and data governance, training algorithms, compute infrastructure, energy and carbon footprint, testing and evaluation, socio-technical harms and mitigations, and usage policies.**

***Index Terms*—foundation model, large language model, generative AI, data governance, contrastive fine-tuning, energy consumption, evaluation, socio-technical harms, usage governance, transparent documentation**

I. Introduction

In this technical report, we present the Granite series of decoder-only foundation models for generative artificial intelligence (AI) tasks. The first in this series, granite.13b, is an English-only large language model (LLM). Using self-supervised learning, this base model has been trained on an IBM-curated pre-training dataset described in Section II. IBM relies on its internal end-to-end data and AI model lifecycle governance process and capabilities to develop enterprise-grade foundation models and is making similar capabilities available to customers of its watsonx platform.

The first versions (v1) of granite.13b models leveraged a base model trained on 1 trillion tokens. The second version of the granite.13b models leverages an updated base model trained on 2.5T trillion tokens. In both versions, the base model is the jumping-off point for two variants: granite.13b.instruct and granite.13b.chat. Granite.13b.instruct has undergone supervised fine-tuning to enable better instruction following [1] so that the model can be used to complete enterprise tasks via prompt engineering. Granite.13b.chat benefits from novel alignment methods to further improve the model's quality of generation, mitigate certain notions of harms, and encourage its outputs to follow certain social norms and have some notion of helpfulness [2]–[4]. We emphasize that these notions are not universal and discuss this point to a greater extent in Section VI on socio-technical harms and risks.

The granite.13b.instruct and granite.13b.chat models are made available by IBM through the watsonx platform [5]. IBM indemnifies customer use of these models on the watsonx platform, providing the same contractual intellectual property protections for IBM-developed AI models as it does for all of IBM's products according to IBM Standard Terms and Conditions.

A. Overview of Capabilities

The 13b in the name indicates the model has 13 billion parameters. Furthermore, the base granite.13b decoder-only model has multi-query attention with learned position embeddings, has been trained on tokens created with the GPT-NeoX 20B tokenizer [6], and has a context length of 8 thousand tokens. The first release of the granite.13b models (granite.13b.instruct.v1 and granite.13b.chat.v1) were trained using an early checkpoint of the base model that had been trained on 1 trillion tokens. The subsequent version of these models (granite.13b.instruct.v2 and granite.13b.chat.v2) were trained on a later checkpoint of granite.13b which saw an additional 1.5 trillion tokens of training, giving granite.13b.v2 models a final pre-training token count of 2.5 trillion tokens. As discussed in Section V, the Granite models are competitive in their 'weight class' on benchmark evaluations while being enterprise-ready in governance dimensions.

Some of the key enterprise tasks (common across sectors) for which the Granite models may be used are: retrieval-augmented generation, summarization, content generation, named entity recognition, insight extraction, and classification. The Granite models may be adapted to the specific tasks arising in particular enterprise applications through prompt engineering in the watsonx platform, which is illustrated in Fig. 1.

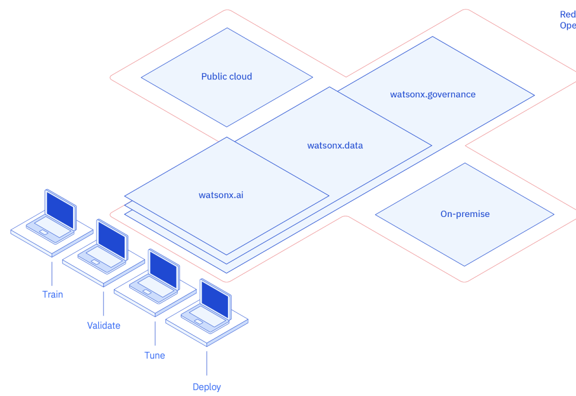


Fig. 1. A conceptual diagram of the watsonx platform.

B. Overview of the Granite Pre-Training Dataset

To support the training of large enterprise-grade foundation models, including granite.13b, IBM curated a massive dataset of relevant unstructured language data from sources across academia, the internet, enterprise (e.g., financial, legal), and code. In a rare move from a major provider of proprietary LLMs, IBM demonstrates its commitment to transparency and responsible AI by publishing descriptions of its training dataset in Section II.

The Granite pre-training dataset was created as a proprietary alternative to commonly used open-source data compilations for LLM training such as “The Pile” [7] or “C4” [8]. Some domains that are key for enterprise natural language processing are relatively under-represented in these compilations. Additionally these data compilations have been criticized for containing toxic, harmful, or pirated content [9]. By curating our own pre-training data corpus, IBM takes significant steps towards addressing these and other issues.

The IBM curated pre-training dataset is continually growing and evolving, with additional data reviewed and considered to be added to the corpus at regular intervals. In addition to increasing the size and scope of pre-training data, new versions of these datasets are regularly generated and maintained to reflect enhanced filtering capabilities (e.g., de-duplication and hate and profanity detection) and improved tooling.

C. Organization of Report

The remainder of this report is organized as follows. In Section II, we describe the data sources used in granite.13b’s pre-training. In Section III, we describe the data processing steps we undertake with a focus on the governance steps we follow. In Section IV, we provide further details about the pre-training and fine-tuning algorithms, the computation involved, and the energy consumption we estimate. Section V presents the testing and evaluation framework along with quantitative comparisons to other models. In Section VI, we discuss our approach to understanding and mitigating socio-technical harms from the Granite models. Section VII provides a brief discussion of the usage policies and the socio-technical documentation of Granite models. Finally in Section VIII, we conclude with areas of future work and discussion.

II. Data Sources

At kick-off for granite.13b’s initial phase of pre-training, IBM had curated 6.48 TB of data before pre-processing, 2.07 TB after pre-processing (detailed in Section III). All datasets were filtered English-text and code unstructured data files. There are no pre-defined labels or targets. All non-text artifacts (e.g., images, HTML tags, etc.) were removed.

Specifically, the first version of this base model, granite.13b.v1, was trained on 1 trillion tokens generated from a total of 14 datasets. The individual datasets used in the training are described below.

1) arXiv: Over 1.8 million scientific paper pre-prints posted to arXiv.

2) Common Crawl: Open repository of web crawl data.

3) DeepMind Mathematics: Mathematical question and answer pairs data.

4) Free Law: Public-domain legal opinions from US federal and state courts.

5) GitHub Clean: Code data from CodeParrot covering a variety of coding languages.

6) Hacker News: News on computer science and entrepreneurship, taken between 2007-2018.

7) OpenWeb Text: Open-source version of OpenAI’s Web Text corpus containing web pages through 2019.

8) Project Gutenberg (PG-19): A repository of free e-books with focus on older works for which U.S. copyright has expired.

9) Pubmed Central: Biomedical and life sciences papers.

10) SEC Filings: 10-K/Q filings from the US Securities and Exchange Commission (SEC) for the years 1934-2022.

11) Stack Exchange: Anonymized set of all user-contributed content on the Stack Exchange network, a popular collection of websites centered around user-contributed questions and answers.

12) USPTO: US patents granted from 1975 to May 2023, excluding design patents.

13) Webhose: Unstructured web content converted into machine-readable data feeds acquired by IBM.

14) Wikimedia: Eight English Wikimedia projects (enwiki, enwikibooks, enwikinews, enwikiquote, enwikisource, enwikiversity, enwikivoyage, enwiktionary). containing extracted plain text from pages and articles.

The second version of the base model, granite.13b.v2, continued pre-training of the granite.13b.v1 model on an additional 1.5T newly-curated tokens for a total of 2.5T tokens seen during pre-training. The datasets used in this second tranche of training tokens were a mixture of the same 14 datasets from granite.13b.v1 (with additional snapshots added from the Common Crawl) along with 6 new datasets described below; all new snapshots and datasets were processed according to the same procedure described in III.

15) Earnings Call Transcripts: Transcripts from the quarterly earnings calls that companies hold with investors. The dataset reports a collection of earnings call transcripts, the related stock prices, and the sector index.

16) EDGAR Filings: Annual reports from all the publicly traded companies in the US spanning a period of more than 25 years.

17) FDIC: The data is from the annual submissions of the FDIC.

18) Finance Text Books: A corpus from UMN’s Open Textbook Library, including a dump of all textbooks tagged as finance.

19) Financial Research Papers: Publicly available financial research paper corpus.

20) IBM Documentation: IBM redbooks and product documents.

III. Data Governance

As IBM is making Granite models available to customers to adapt to their own applications, we have invested heavily in a

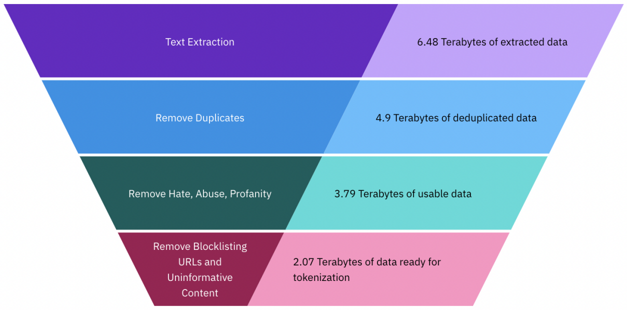


Fig. 2. Summary governance statistics on IBM's curated pre-training dataset at the time of granite.13b.v1's training.

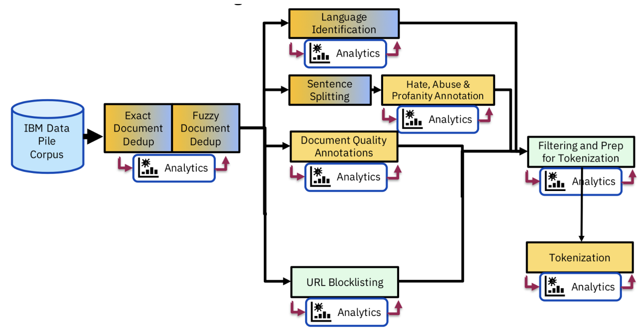


Fig. 3. IBM's Data pre-processing pipeline.

data governance process that evaluates datasets for governance, risk and compliance (GRC) criteria, including IBM's standard data clearance process, document quality checks, and other criteria. IBM has developed governance procedures for LLM pre-training datasets which are consistent with IBM AI Ethics principles and are guided by the IBM Corporate Legal Team. Best practices around LLM development is continually evolving with the ever-increasing understanding of AI models, their usage, and changing regulatory requirements, among other factors.

Addressing GRC criteria for data spans the lifecycle of training data, from data request to tokenization. An important objective for IBM is establishing an internal auditable link from a trained foundation model to the specific dataset version on which the model was trained, including information about each processing step performed prior to training. Summary statistics on IBM's curated pre-training dataset are provided in Fig. 2.

Data governance is organized into the following processes, corresponding to data lifecycle phases prior to model training:

A. Data clearance and acquisition;
B. Pre-processing; and
C. Tokenization.

Each process is composed of sub-processes focusing on specific governance aspects. The remainder of this section describes each phase in detail.

A. Data Clearance and Acquisition

The data clearance process assures that no datasets are used to train IBM foundation models, including the Granite series, without careful consideration. Before data is added to IBM's curated pre-training dataset, it is submitted to the data clearance process and subject to technical, business, and governance review. The clearance request captures comprehensive information about a dataset such as a thorough description, the data owner, the intended use, geographic location, data classification, licensing information (if available), usage restrictions and sensitivity (e.g., personal information). Additional information includes who will have access to the data, and how the data will be acquired.

Once a dataset completes the review process, it is tagged for potential inclusion, its metadata is moved into a catalog of approved datasets, and it is downloaded and prepared for the subsequent pre-processing stages.

Remark. *IBM's pre-training dataset currently addresses potentially infringing material through selective use of URL block-listing to exclude websites known to disseminate pirated information. Examples of datasets that are block-listed include the Books3 dataset, which is specifically excluded from use due to concerns around pirated information and its use in model training.*

B. Pre-Processing Pipeline

Once data has been cleared and downloaded, it is prepared for model training through a variety of steps collectively referred to as the *pre-processing pipeline*. An overview of the pre-processing pipeline for this release of Granite models is depicted in Fig. 3 and is composed of the following steps:

1) Text extraction
2) De-duplication
3) Language identification
4) Sentence splitting
5) Hate, abuse and profanity annotation
6) Document quality annotation
7) URL block-listing annotation
8) Filtering
9) Tokenization.

Some pre-processing steps follow an annotation/filtering pattern, where documents or sentences are annotated first and filtered later during the filtering task according to threshold definitions.

The completion of each pipeline step in the pipeline is logged. Logs are used to construct metadata reflecting the exact pre-processing steps performed on a dataset, laying the basis for end-to-end traceability of the model lifecycle.

We now describe each step of the pre-processing pipeline in greater detail.

1) Text Extraction: Text extraction is the first step in the pipeline, and is used to extract language from various documents into a standardized format for futher processing.

2) Data De-Duplication: Data de-duplication aims to identify and remove duplicate documents. De-duplication is performed on a per-dataset basis and is essential to ensuring the trained model does not learn artificial linguistic patterns due to repeated data in the dataset.

Two techniques are used: exact and fuzzy de-duplication, both of which use hash-based methods. As the name suggests, exact de-duplication removes exact duplicates among the documents in the dataset. Each document is hashed and documents with the same hash are fused to one. For example, if 50 documents in a dataset have the same hash, a single document will be used. Fuzzy de-duplication finds the Jaccard similarity between documents with locality sensitive hashing. If multiple updated snapshots of a dataset are downloaded, the exact de-duplication is performed across all snapshots.

3) Language Identification: Language identification is performed at a document level to detect the dominant language using the Watson Natural Language Processing (NLP) library [10].

The output of this task is an additional column in the parquet file containing a two letter ISO language code.

In the case of the Common Crawl dataset, language is already provided through folder names. The Watson NLP language identification algorithm is nevertheless run on Common Crawl documents, yielding two language classifications for these documents: Common Crawl and Watson NLP.

4) Sentence Splitting: Sentence splitting involves decomposing each document into its constituent sentences. Sentence splitting is key for hate, abuse, and profanity (HAP) annotation (to be discussed below) since HAP annotation is performed at a sentence level. As such, the sentence splitting stage must take place prior to the start of HAP annotation. Sentence splitting for the English language is performed using Watson NLP.

5) Hate, Abuse and Profanity Annotation: Data sources drawing from the open Internet, such as Common Crawl, inevitably contain abusive language. To reduce the possibility of Granite models producing profane content, each sentence in each document is assessed and scored as to its level of HAP content. The HAP detector is itself a language model trained by IBM and benchmarked against internal as well as public models such as OffensEval [11], AbusEval [12] and HatEval [13]. The IBM HAP detector performs comparably to HateBERT [14].

After a score is assigned to each sentence in the document, analytics are run over the sentences and scores to explore the distribution of annotations in each document with a HAP annotation. This serves both to determine the percentage of HAP sentences in a document as well as to determine threshold values used later during filtering.

6) Document Quality: Quality annotation aims to identify documents with low linguistic value using both heuristics and a classifier. The heuristics are derived from the Gopher Quality Filtering criteria [15]:

- total words: outside the range 50–100,000 words;
- average word length: outside the range 3–10 characters per word;
- symbol to word ratio: greater than 10%;
- bullet points ratio: greater than 90%;
- ellipsis line ratio: greater than 30%;
- alphabet words ratio: fewer than 80%;
- common English words: does not contain at least 2 from {the, be, to, of, and, that, have, with}.

The classifier assigns a perplexity score using the KenLM linear classifier pre-trained on Wikipedia documents [16], [17]. For any document, the model provides a score of the document's similarity to a training corpus (i.e., Wikipedia).

These heuristics and classifiers output columns with quality scores that are added to the parquet file. These annotations form the basis for quality filtering during the filtering step.

7) URL Block-Listing: Block-listing identifies documents to be blocked from being added to IBM's curated pre-training dataset. The block list is continuously maintained and includes URLs of known copyrighted material as well as block-listed sites such those contained in the 2022 Review of Notorious Markets for Counterfeiting and Piracy [18].

8) Filtering: Filtering occurs at the document level and is the last step before tokenization. It is here that annotations created in previous pre-processing steps are used to prevent documents from being used for tokenization. For example, documents are dropped which exceed HAP thresholds or do not meet a defined document quality. For the current English-only Granite models, the language identification annotations are used to filter out non-English documents.

C. Tokenization

Tokenization is the final pre-processing step prior to model training. For granite.13b, the cleaned and filtered text is converted from a sequence of characters to a vector of tokens using the GPT-NeoX 20B tokenizer [6].

IV. Training

In this section, we detail the training process for the decoder-only Granite models covering the algorithmic details of pre-training and fine-tuning, the computing involved, and an estimate of the carbon footprint.

A. Algorithmic Details

1) Granite.13b Pre-Training: We adopt most of the pre-training settings from [19]. Specifically, we use the standard decoder-only transformer architecture [20], Gaussian error linear unit (GELU) activation function [21], MultiQuery-Attention for inference efficiency [22], and learned absolute positional embeddings. We also adopt FlashAttention to speed up the training and reduce its memory footprint [23], allowing us to increase the context length to 8192 from the context length 2048 used by many existing LLMs.

The granite.13b.v1 base model is trained for 300K iterations, with a batch size of 4M tokens, for a total of 1.25 trillion tokens. The granite.13b.v2 base model continued pre-training on top of the granite.13b.v1 checkpoint for an additional 300K iterations and a total of 2.5 trillion tokens.

We train using the Adam optimizer [24], with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, and a weight decay of 0.1. We use a cosine learning rate schedule, with warmup of 2000 steps, and decay final learning rate down from 3×10^{-4} to 3×10^{-5}. We pre-train models with a 3D-parallel layout using both tensor and pipeline parallelism including sequence parallelism to enable training with 8K context length. Additionally, we used FlashAttention-2 [25] for training of granite.13b.v2 model, allowing much longer context length (e.g., 16K) for the same price as previously training a 8k context length model.

2) Granite.13b.instruct Alignment: Pre-training teaches the LLM to continue generating text based on the input. However in practice, users often expect the LLM to treat the input as instructions to follow. To enable instruction following, we perform supervised fine-tuning (SFT) with a mixture of datasets from different sources. Each sample consists of a prompt and an answer. We use a cosine learning rate schedule with an initial learning rate of 2×10^{-5}, a weight decay of 0.1, a batch size of 128, and a sequence length of 8192 tokens. We perform SFT for 3 epochs to obtain the granite.13b.instruct.v1 model.

The SFT data includes a subset of the Flan Collection [26], 15K samples from Dolly [2], Anthropic's human preference data about helpfulness and harmlessness [3], Instructv3 [27], and internal synthetic datasets specifically designed for summarization and dialogue tasks.

For granite.13b.instruct.v2, when the updated 2.5T token base model was available, the same data mixture was used on the new base checkpoint but with additional noise augmentations (to the input) to improve the model's robustness to white-spaces and special characters. Moreover, we adopt NEFTune [28], that adds noise to the embedding vectors during training (with no additional compute or data overhead) to improve the performance on conversational tasks.

3) Granite.13b.chat.v1 Alignment: Contrastive fine-tuning (CFT) is an instruction fine-tuning approach, used to align granite.13b.chat.v1, based on unlikelihood-based training [29], which penalizes the probability of data points from a negative data distribution while simultaneously increasing the probability of data points from a positive data distribution (see Fig. 4). In other words, we discourage an LLM from generating misaligned responses (e.g. responses that are harmful) while encouraging aligned responses (e.g. responses that are helpful) for each training prompt.

CFT requires both responses to be paired with the same prompt in order for the model to determine which response is worse. However, many publicly available human demonstration datasets lack paired aligned and misaligned responses for the same prompt. As such, one may wonder: *"How can one obtain both aligned and misaligned responses?"* A straightforward approach to obtain this negative data distribution is to have humans write misaligned responses for each prompt.

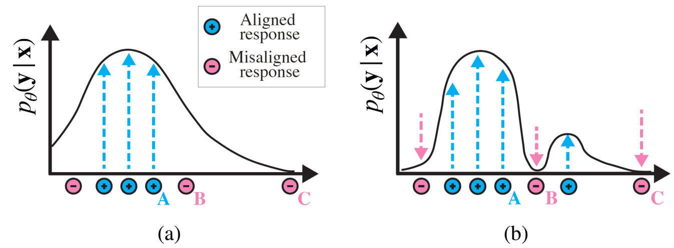


Fig. 4. Illustration for the resulting LLM distribution over responses y for a given prompt x. (a) SFT increases the likelihood of aligned responses but lacks control over misaligned responses. As a result, misaligned responses that closely resemble aligned ones can still have high likelihood. Examples A and B demonstrate this, where A is aligned and B is misaligned, yet they are similar. (b) Our approach, contrastive fine-tuning mitigates this by explicitly assigning low likelihood to misaligned responses.

However, such an approach can be cost-prohibitive.

Thus in our work we propose to use a separate LLM to serve as a 'negative persona' by mimicking individuals who tend to respond in a misaligned (e.g. harmful or untruthful) manner. We achieve this by fine-tuning an LLM on widely available misaligned human demonstration datasets. Consequently, for a given dataset of prompts and aligned human demonstrations, responses on the prompts from this negative persona LLM form the misaligned responses and the paired aligned responses are the demonstrations themselves.

For granite.13b.chat.v1, we use an early version of granite.13b.instruct.v1 as the separate LLM. The datasets for CFT are paired samples from Anthropic's human preference data about helpfulness and harmlessness that have been filtered using the OpenAssist reward model [3], samples from Dolly [2], and samples from ProsocialDialog [4].

As a part of the CFT step, the granite.13b.chat.v1 was also trained to work with the following system prompt [3] in order to support Human-Agent based dialogue:

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:

4) Granite.13b.chat.v2 Alignment: In the latest version of Granite.13b.chat.v2, we focus on enhancing the quality of generation, especially for longer-form generative tasks such as Retrieval-Augmented Generation (RAG) and Summarization.

Our strategy for improving generation quality involved addressing the primary limitation of fine-tuning on publicly available instruction-tuning datasets, such as the FLAN dataset. Despite FLAN's rich diversity, its instruction responses are characteristically brief, a feature empirically linked to suboptimal model performance [30], [31]. To mitigate this, we introduced a new synthetic dataset, generated through IBM's f-ORCA technique, into the instruction tuning of Granite.13b.chat.v2.

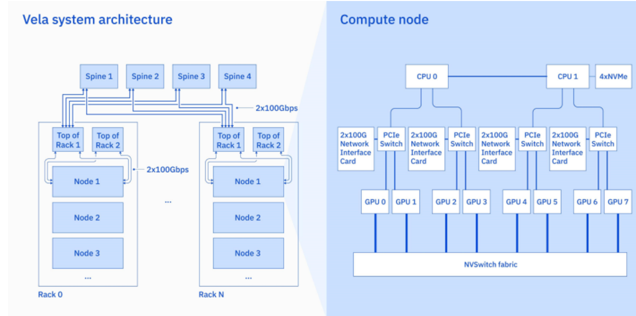
The f-ORCA methodology is a novel, exemplified-principles-based, in-context learning (ICL) technique, that is specifically designed to produce high-quality, varied responses that closely align with expert demonstrations for instruction tuning in large language models. While it shares similarities with the recently developed ORCA method—particularly in its capability to generate superior responses to the diverse prompts present in datasets such as FLAN—f-ORCA distinguishes itself in several key aspects. Primarily, it eschews the necessity for an aligned, instruction-tuned, and expensive black-box model, such as ChatGPT or GPT-4, which are foundational to ORCA's approach. Instead, f-ORCA only utilizes a pre-trained Falcon-180b model for the synthesis of high-quality responses. Furthermore, f-ORCA incorporates a rigorous quality control mechanism: an additional filtration step (f-UMPBACK) wherein the Falcon-180b model is repurposed as an auditor using a slight variation of the same exemplified-principles-based ICL technique. This allows for the model to self-assess and curate its outputs, ensuring adherence to high standards of accuracy, factual correctness, naturalness, and safety. This entire process results in a dataset of roughly 300K samples. We further incorporate a subset of datasets that proved helpful in the training of Granite.13b.chat.v1. The final dataset mixture is used to instruction tune to the Granite.13b.chat.v2 model.

In Granite.13b.chat.v2, to further improve model safety we applied SALMON method [32], an RLAIF algorithm recently proposed by IBM. This goal of SALMON is in line with the recent research on self-alignment [33], where the primary focus is to use AI models to improve themselves, e.g., with bootstrapping over the model-generated critiques [34] or self-refined outputs [35]. Central to SALMON is a principle-following reward model. Trained on synthetic preference data, this model can generate reward scores based on arbitrary human-defined principles. By merely adjusting these principles during the RL training phase, we gain full control over the preferences with the reward model, subsequently influencing the behavior of the RL-trained policies, and eliminating the reliance on the collection of online human preferences.

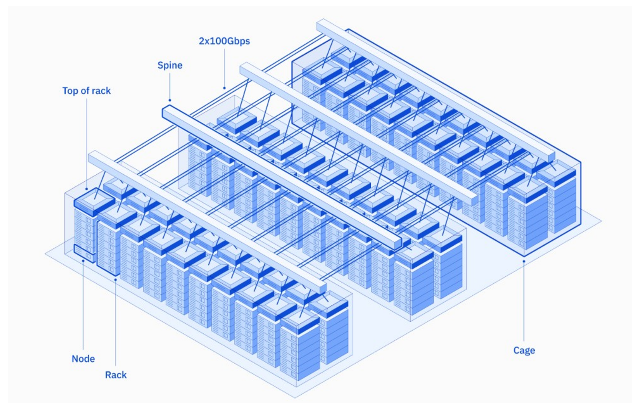
Following the method proposed in [32], we train the instruction-following reward model from a 40 billion parameter base model, using a human preference dataset and a synthetic principle-following preference dataset. Then, the reward is used in PPO training to align the granite.13b.chat.v2 with a set of principles defined by IBM researchers. The full list of principles can be found in Appendix B.

The datasets relied upon for both f-ORCA and SALMON steps in the granite.13b.chat.v2 model include IBM-generated synthetic data, HHRLHF [36], human-generated prompts taken from the sharegpt dataset [37] (no chatGPT responses were used), a filtered subset of the FLAN 2022 instruction tuning data collection, and OASST [38].

Unlike granite.13b.chat.v1, granite.13b.chat.v2 does not require a system prompt for inference. In future versions of this report, pending the full evaluation of the granite.13b.chat.v2 model, system prompts may be recommended for certain use cases.



(a)



(b)

Fig. 5. An (a) architectural and (b) infrastructure diagram of the AI supercomputer Vela.

B. Compute

IBM's primary computing infrastructure for training foundation models is the Vela AI supercomputer [39] (cf. diagram in Fig. 5). Vela uses a virtual machine-based approach for elasticity in resource allocation; with various optimizations, the 'virtual machine tax' is less than 5%. Each AI node has 8 Nvidia A100 GPU Cards, 96 vCPUs, 1.5 TB of DRAM and 4×3.2 TB NVMe drives. The nodes are interconnected via Ethernet. Each node has 2×100 Gbps Ethernet links. The Vela instance currently being used for model training is located in IBM Cloud's Washington D.C. Data Center. Future Granite models are planned to be trained using Vela, however, the granite.13b base model was trained on older infrastructure before the Vela instance was fully stood up. Granite.13b.v1 used 256 A100 GPUs for 1056 hours and 120 TFLOPs. Granite.13b.v2 was trained on the same infrastructure for an additional 1152 hours with 120 TFLOPS, bringing the total to 2208 hours.

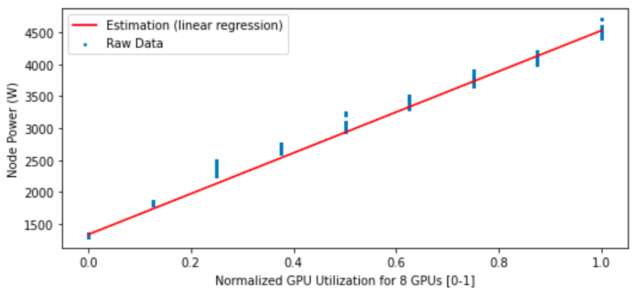


Fig. 6. Server (node) power vs. normalized GPU utilization.

C. Energy Consumption and Carbon Emissions

The methodology used to estimate the energy consumption and carbon emissions of the granite.13b base model is as follows. The carbon emissions $Carbon$ associated with a model M at a particular location L is given by:

$$Carbon(M, L) = E(M) \times PUE(L) \times CEF(L), \qquad (1)$$

where $E(M)$ is the electricity consumption of the model M, $PUE(L)$ is the power usage effectiveness at the location L, and $CEF(L)$ is the carbon emission factor applicable for the location L.

The information technology (IT) electricity consumption $E(M)$ is estimated using the average GPU utilization rate for all the GPUs. It is a proxy to estimate the power that is used to train the AI model M since the GPU utilization is typically highly correlated with the node power, as shown in Fig. 6. Then, the estimated node power is multiplied by the training time and the number of GPUs used to calculate the total compute energy consumption E.

Power usage effectiveness $PUE(L)$ is given by the ratio of the total electricity consumed by the data center (aggregate consumption by the IT and support overhead infrastructure) to that consumed by the IT infrastructure. We calculate the location-based carbon emission factor $CEF(L)$ following the GHG Protocol's Scope 2 Guidance [40].

Applying this estimation methodology to the granite.13b.v1 base model, we estimated 153074.3767 kWh energy consumption $E(M)$ and 0.12 kg/kWh carbon emission factor $CEF(L)$, yielding 22.2263995 tons of CO_2 equivalent $Carbon(M, L)$, which accounts for carbon dioxide and all other greenhouse gases, such as methane and nitrous oxide.

Energy consumption and carbon emissions for the granite.13b.v2 base model are still being estimated and will be published in a later release of this report.

A number of mitigation strategies may be used to reduce the energy and carbon footprint. For example, the amount of resources used in training may be adjusted as a function of the availability of renewable energy, or the resources usage may be capped to not exceed certain energy usage or emissions limits.

V. Testing and Evaluation

In this section, we describe the approach taken to test and evaluate the Granite models. We also provide empirical results along with comparisons to several other models that are of a similar capability level.

A. Foundation Model Evaluation Framework

We use a comprehensive foundation model evaluation framework (FM-eval) through the model's development lifecycle. FM-eval is running on RedHat OpenShift[1] cluster with GPU support, for efficient execution of evaluation benchmarks, in parallel and on multiple models. The automation framework can run any containerized evaluation framework or a wrapped external framework such as Eleuther AI's Language Model Evaluation Harness (lm-eval) [41] or Stanford's HELM (Holistic Evaluation Model) [42]. To allow easy addition of tasks, datasets and metrics to FM-eval, we developed Unitxt[2], an open-source Python library that provides a consistent interface and methodology for defining datasets, including the preprocessing required to convert raw datasets to the input required by LLMs, and the metrics used to evaluate the results.

Different types of tests are run during different phases of the lifecycle:

1) General knowledge benchmarks (during training)
2) HELM benchmarks (post-training)
3) Enterprise benchmarks (post-training)

These evaluations all leverage zero-shot and few-shot prompting. For clarity, zero-shot prompting uses a pre-existing LLM to generate text for a new task by only providing the instruction to execute the task in the prompt. In few-shot prompting, we provide multiple in-context examples, along with the task at hand, directly within the prompt. Both approaches allowed us to work with a single pre-trained model whose core parameters remained fixed.

The specific evaluations are detailed below.

1) General Knowledge Benchmarks During Training: The General Knowledge Benchmarks include a subset of existing benchmarks from lm-eval [41] and are used as light-weight tests run after every 100 billion tokens during training to validate model knowledge is advancing as training progresses.

Specifically, the following 12 datasets (organized by task) from lm-eval are:

- question answering for several domains (boolq, openbookqa, piqa, sciq);
- sentence completion (lambada)
- commonsense reasoning (arc_easy, arc_challenge, copa, hellaswag, winogrande);
- reading comprehension (race)
- multidisciplinary multiple-choice collection (mmlu);

[1] https://www.redhat.com/en/technologies/cloud-computing/openshift
[2] https://github.com/IBM/unitxt

In our evaluation framework these benchmarks are run in both the zero-shot and few-shot setting.

2) HELM: Part of the comprehensive assessment after pre-training relies on Stanford's Hollistic Evaluation of Language Models (HELM) Benchmark [42]. To evaluate our model, we use the 16 "core scenarios", consisting of a variety of tasks including question answering, information retrieval, summarization, sentiment analysis, and text classification [43]–[54] ,on which all HELM LLMs are evaluated and compared using their results and mean win rate (MWR).

3) Enterprise Evaluation Benchmarks: After training completes, we further evaluate our models on IBM-curated enterprise benchmarks to test our models performance in domains highly relevant to our customers. With this in mind, IBM curated 11 publicly available finance benchmarks for evaluating models in the financial domain, summarized in Table I.

The data source-provided train and test splits are used in the evaluation whenever possible. Model performance is reported based on test examples. If the test labels are not publicly available, model performance is reported on the validation set. If the train and test splits do not exist in the data source, 20% of the data is selected as test split and the rest is used as the train split.

All few-shot context examples are sampled from the training set. The number of few-shot examples provided to the model depends on the task, which is provided in Table I. For the current evaluation, all the models used the same parameters and the same standard prompt (see the techniques of few-shot-prompting and zero-shot-prompting and examples of prompts[3]), without task description, chain-of-thought prompting [55], or system prompts in place. For Financial Phrasesbank, News Headline and FiQA SA, the prompts were taken from BloombergGPT [56].

B. Granite Model Evaluation and Comparison

Full evaluation of granite.13b.v2 is still ongoing, results will be shared in an updated version of this report as soon as they are available. Evaluation results for granite.13b.v1 can be found below, but it should be noted that this initial release of granite.13b only saw 1T tokens during training, making all of these evaluations preliminary.

1) General Knowledge Benchmarks During Training: In this section, we leverage the lighter-weight General Knowledge Benchmarks to assess a series of snapshots of the granite.13b.v1 base model taken every 100B tokens during training along with the fine-tuned granite.13b.instruct.v1 and granite.13b.chat.v1 variants. As visualized in Fig. 7 and further detailed in Table II, progressively training on each 100B tokens steadily improved General Knowledge as expected with further boosts in performance achieved in both fine-tuned variants of granite.13b.v1. Note system prompts were not used for this evaluation of granite.13b.chat.v1.

[3]https://www.promptingguide.ai/techniques/fewshot

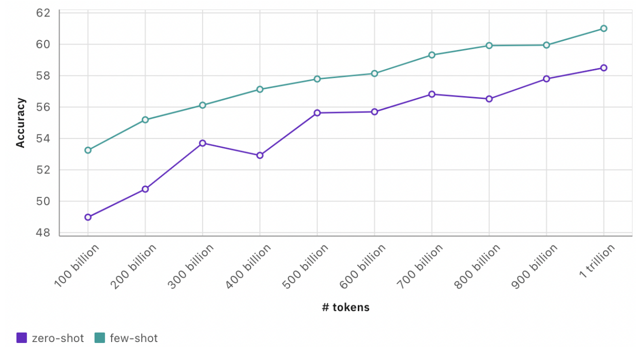


Fig. 7. Granite.13b General Knowledge Performance during Training.

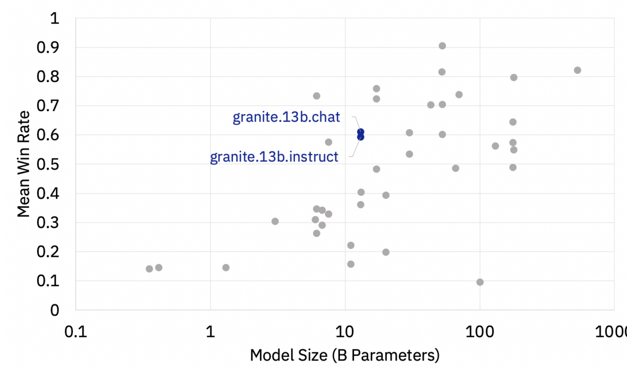


Fig. 8. Granite.13b.instruct.v1 and Granite.13b.chat.v1 Performance on HELM Tasks vs Model Size.

2) HELM Benchmarks: In this study, we comprehensively assess our models across HELM's 16 core scenarios, comparing them to all other models in the v0.2.3 release[5].

Our evaluation follows the default two-tiered process as suggested by HELM: first, we evaluate each model on individual evaluation datasets, then we aggregate these results into scenarios (as detailed in Table III). To facilitate a fair comparison of LLMs, we employ HELM's Mean Win Rate (MWR) metric across scenarios for model ranking.

Figure 8 illustrates the positioning of granite.13b.instruct.v1, granite.13b.chat.v1, and all v0.2.3 models along the model size MWR axes. Note, models where exact model size is not published by the LLM-provider are excluded from this visualization.

The figure highlights that the granite models strike a desirable balance between model size and HELM performance. Granite.13b.chat.v1 and granite.13b.instruct.v1 are respectively ranked 15 and 18, out of all models evaluated. Further, the granite.13b.chat.v1 and granite.13b.instruct.v1 models were respectively the top-2 and top-3 models evaluated under 17b parameters in size. Only Cohere Command beta (6.1b) exceeded their performance in this size category. These results also hold for the other aspects evaluated by HELM, such as robustness, and fairness. In calibration granite.13b.chat.v1 and granite.13b.instruct.v1 are ranked 9 and 28, respectively.

[5]https://crfm.stanford.edu/helm/latest/?group=core_scenarios

TABLE I
FINANCE BENCHMARKS OVERVIEW

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

TABLE II
GRANITE.13B GENERAL KNOWLEDGE PERFORMANCE DURING TRAINING

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**

3) Enterprise Benchmarks: This evaluation is conducted by augmenting HELM's framework to encompass 11 publicly available task datasets from the financial services domain. Baseline models are selected based on model size, type of training data, accessibility, and model tuning. To be specific, granite models are compared with GPT-NeoX-20B [6] and FLAN-UL2 [67], some of the best performing open-sourced models under 50 billion parameters, along with state-of-the-art available LLaMA2 models [68] between 7 billion to 13 billion parameters.

Table IV presents the detailed performance scores of the models on the 11 financial tasks. An asterisk is given next to the LlaMA2 models, as these models have seen 2T tokens of pre-training data, imparting significant advantage to the models. All other models evaluated, including granite, have seen 1T tokens of training data. Despite having been trained on half the amount of data as the LlaMA2 models, the Granite.v1 models are competitive across each of the tasks, often meeting or outperforming LlaMA2. This bodes well for future planned versions of granite models, which will be trained on 2T+ tokens of pre-training data.

VI. SOCIO-TECHNICAL HARMS AND RISKS

Numerous potential socio-technical harms and risks of LLMs have been identified in recent years, including misinformation, hallucination, lack of faithfulness or factuality, leakage of private information, plagiarism or inclusion of copyrighted

TABLE III
HELM RESULTS PER SUB SCENARIO IN THE CORE-SCENARIOS.

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

TABLE IV
FINANCE BENCHMARK EVALUATION RESULTS PER TASK.

	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.v1	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

content, hate speech, toxicity, human-computer interaction harms such as bullying and gaslighting, malicious uses, and adversarial attacks [69], [70].

In Table V, we present the catalogue of risks compiled by the IBM AI Ethics Board, a central, cross-disciplinary body that defines the AI ethics vision and strategy with the objective of supporting a culture of ethical, responsible, and trustworthy AI throughout the IBM Corporation [71], [72]. The table is organized across several dimensions [73]:

- Whether the risk is from the data or other inputs to the foundation model, from the generated output of the foundation model, or from other concerns.
- Whether the risk arises in the training/tuning of the model, during inference, or in broader considerations such as governance, legal compliance, or societal impact.
- What higher-level grouping the risk falls under, e.g. fairness, robustness, intellectual property, and misuse.
- Whether the risk is new or amplified. 'Traditional' risks are present in earlier forms of AI models and continue to be present in foundation models. 'Amplified' risks are known from earlier forms of AI models but are intensified by foundation models due to their generative capabilities. 'New' risks are emerging risks, intrinsic to foundation models due to their generative capabilities.

As part of creating and releasing the granite.13b.instruct and granite.13b.chat models, we have addressed some of the risks as follows. The data governance processes of the IBM's pre-training dataset, including the block-listing and filtering of hate, abuse and profanity have mitigated some intellectual property and misuse risks. Toward fairness, an additional component of the data pre-processing pipeline not described in Section III is annotating documents by religion, gender, race, stigma, age, and political ideology. We have created keyword lists for these dimensions and use keyword matching to annotate sentences. The annotations may be used to identify under-represented and over-represented groups. We have not been overly aggressive in HAP filtering and have not filtered with respect to groups because it would prevent us from having training data that reclaims slurs and positively describes marginalized identities, and might skew the pre-training dataset in other unintended ways [74].

Through model alignment, we have encouraged prosocial and less harmful model behavior with the aim to mitigate certain aspects of misuse and value alignment risks. However, one of the biggest socio-technical risks would be our own hubris to believe that the datasets we used for fine-tuning (or other existing datasets we could use in the future) are aligned to the needs, wants, and desires of the peoples and organizations that will be deploying the Granite models to meet their own goals. Every enterprise has its own regulations to conform to, whether they come from laws, social norms, industry standards, market demands, or architectural requirements [75]; we believe that enterprises should be empowered to personalize their models according to their own values (within bounds) [76], e.g. using tools in the watsonx platform.

In addition, through FM-eval, we have tested the Granite models on benchmark datasets that cover several risk dimensions. However, evaluating on benchmarks is a limited approach for revealing socio-technical harms [77]. After enterprises have further aligned the Granite models to their own values, they should enlist a red team with members of varying socio-cultural and lived experience to find additional harms and undesirable LLM behaviors within the context of a precise use case [78].

TABLE V
SOCIO-TECHNICAL HARMS AND RISKS

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

VII. Usage Policies and Documentation

A. Machine-Generated Content

IBM's licensing terms and conditions govern downstream applications and services that use IBM models.

In addition, IBM is setting up an Acceptable Use Provision (AUP) that states guidelines and practices that the users of IBM models are required to follow as they develop and deliver downstream applications and services.

The AUP provides acceptable use of AI Models and confers to IBM the right to terminate the license to these models if necessary.

B. European Union-Specific Controls

The licensing terms and conditions to IBM Models are augmented with an Acceptable Use Policy (AUP) that states guidelines and practices that are specific to deployments of downstream applications and services in specific Countries.

C. Downstream Documentation

For downstream usage of its pre-trained models, IBM makes available the following documentation:

- Terms and Conditions
- Product documentation
- Technical reports, such as this report

Together, this information is designed so that not only IBM complies with legal and ethical requirements, but also the users of these models can comply with their own obligations.

1) Terms and Conditions: The latest Terms and Conditions for the watsonx platform can be found at https://www.ibm.com/support/customer/csol/terms/?id=i126-6883 .

2) Product documentation: The IBM Granite models are currently available through IBM's watsonx platform. As part of watsonx, each Granite model is accompanied by a model card that details key facts and provenance of the model.

VIII. Conclusion

In this technical report, we have presented IBM's Granite family of foundation models designed for enterprise generative AI applications. IBM's ethical and governance frameworks provide the context within which these models are created and made available. Aligned with IBM's commitment to transparent and responsible AI, we have presented descriptions of exact datasets, pre-processing steps, training infrastructure, energy consumption, and testing/evaluation methodologies used throughout the model development lifecycle.

We are continuing to develop the Granite series in several directions. Whereas this initial Granite release only supports English, future models will be trained on multiple natural languages. Alongside, HAP annotation is being refined and expanded for additional languages. Furthermore, Granite models for other modalities such as code as well as industry-specific content are being developed. On the model safety evaluation front, we are developing a comprehensive red-teaming framework. The adversarial prompts will test the models across a variety of domains, including (but not limited to) HAP, bias and stigma, factual correctness and harmful topics.

We are continuing to develop additional data annotations for IBM's curated pre-training dataset, such as scoring documents for their inclusion of personally-identifiable information and for their conversationality [79], [80]. We are working toward instrumenting our compute infrastructure to obtain precise rather than estimated measurement of energy and carbon footprints [81]. Finally, we are exploring the application of various methods for mitigating unwanted biases [82]–[84].

References

[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-open-commercially-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-watsonx-ai, 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. Mitrović, 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. Ré, "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.

[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. Köpf, Y. Kilcher, D. von Rütte, 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.

[39] 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.

[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. Ré, 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. Kočiský, 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. Rozière, 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. Röttger, 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.
[84] 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
Release Notes

September 15th, 2023

- Initial report released.

November 7th, 2023

- Table IV, updated with new values for FiQA - Opinion and Insurance QA metrics. New values were calculated after correcting a bug found in HELM's ranking metric protocol. Oasst-sft-pythia-12b was additionally temporarily removed from the analysis as a benchmark, as it was not immediately available to the evaluation team to rerun after the HELM ranking metric fix was implemented.
- Several minor typo and grammar corrections updated throughout.

November 30th, 2023

- Updated entire report with new documentation on the granite.13b.v2 models. Evaluation results were still pending at the time of this report's release and will be shared in an updated version of this report at a later date.
- Updated language of the remark on copyrighted materials for clarity.

Appendix B
Principles used in SALMON process

1) Honest and Accurate: The AI must furnish reliable and factual information, and candidly disclose its limitations and the extent of its knowledge.
2) Ethical: The AI should produce content that is free from offensive, discriminatory, or harmful material, and should not participate in or endorse risky activities.
3) Educational and Engaging: The AI's responses should be enriched with accurate, relevant, and current information, serving to educate while keeping the user engaged.
4) Creative: The AI should be adept at generating original content, such as poems, stories, code, essays, songs, parodies, summaries, translations, and more.
5) Multilingual: The AI should be capable of conversing in the language used by the user, for instance, replying in Chinese if the query is in Chinese."
6) Comprehensive: For information-seeking tasks, the AI should offer extensive and relevant details to ensure a thorough and in-depth response. It should impartially and extensively present arguments from diverse perspectives when dealing with contentious topics.
7) Natural Language: The AI should respond with diverse and natural language, avoiding repetition and awkward phrasing.
8) Consistent Reasoning: The AI should deliver responses that are clear and logically sound, ensuring they do not contain self-contradictions.
9) Numerical Sensitive: The AI should ensure that any numerical specifications given in the instruction are carefully adhered to, avoiding any errors in numerical computations.
10) Analytical Structure: For information analysis tasks, the AI should articulate its response in a manner that begins with a summary, followed by numerous key points, each underscored by a thorough analysis.
11) Vivid: The AI should employ vibrant, energetic language, enhancing user engagement by making all interactions lively and dynamic.
12) Privacy Protection: The AI should avoid generating any personal identifiable information (PII) or external URLs in its responses.
13) Candor: The AI should avoid sharing false information. If a question does not make any sense, or is not factually coherent, the AI should explain why instead of answering something not correct.
14) Stand-Alone: The AI must function as a stand-alone text-based system, avoiding interactions with any external sources, including URLs, images, or videos.