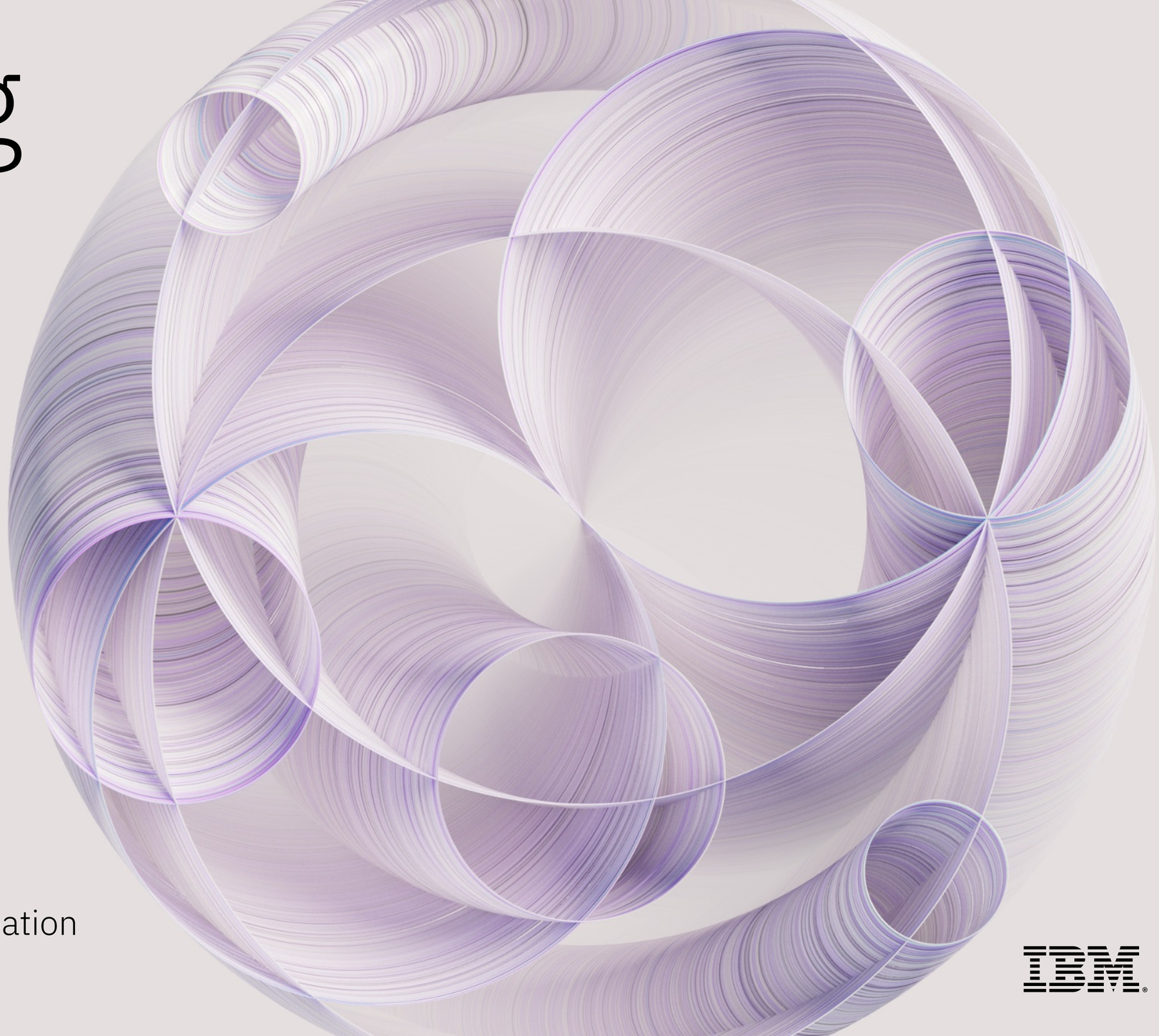
Introducing Granite



Maryam Ashoori
Director of Product Management
watsonx Foundation Models

Kate Soule
Sr Manager, Business Strategy and Incubation
AI Models, IBM Research

Legal disclaimer

This presentation marked as **"IBM and Business Partner Internal Use Only"** is for IBM and Business Partner use and should not be shared with clients or anyone else outside of IBM or the Business Partners' company.

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth, or other results.

All client examples described are presented as illustrations of how those clients have used IBM products and the results, they may have achieved. Actual environmental costs and performance characteristics may vary by client.

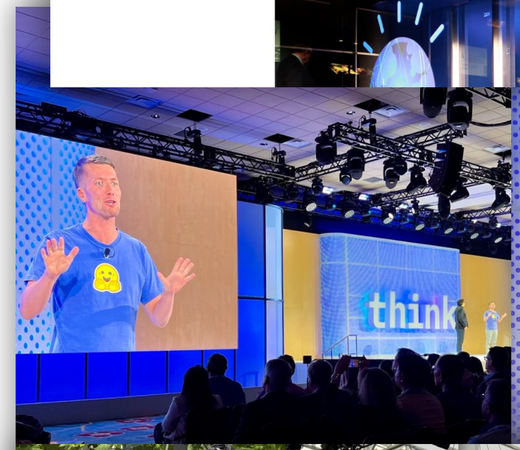
IBM and Business Partner
Internal Use Only

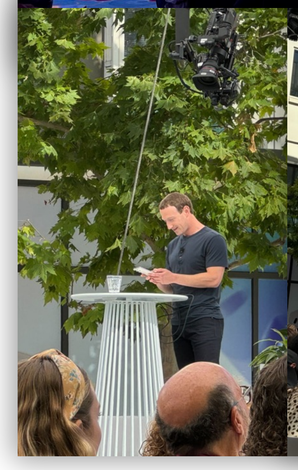
The New York Times

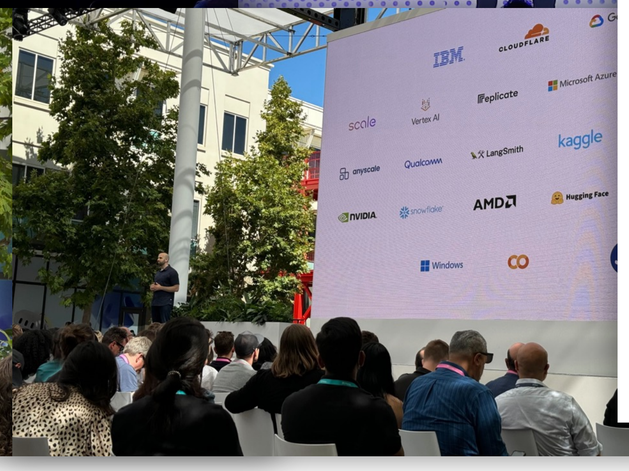
IBM Tries to Ease Customers' Qualms About Using Generative A.I.

The company will assume the legal risk of businesses that use its A.I. systems and will publish the technology's underlying data.

Share full article







Bloomberg the Company & Its Products | Bloomberg Anywhere Remote Login | Bloomberg Terminal Demo Request

News | Podcasts | Research Tools | Bloomberg Law | Log In | Sign Up For Newsletter

IP Law

IBM logo.
Chris J. Ratcliffe/Bloomberg

Sept. 28, 2023, 5:00 PM EDT

IBM Joins Microsoft, Adobe in Protecting AI Customers From Suits

Isaiah Poritz
Legal Reporter

Yann LeCun
@ylecun

Big Tech is not the problem.
Closed and proprietary is the problem.
Meta, IBM are Big Tech and open.
Google, Apple are Big Tech and closed.
OpenAI, Anthropic are Small Tech and closed.
Hugging Face, Mistral are Small Tech and open.

2:02 PM · 9/23/23 · **12.1K** Views

REUTERS | World | Business | Markets | Sustainability | Legal | Breakingviews | More

Disrupted

IBM to launch Meta's Llama 2 on watsonx AI platform for businesses

Reuters
August 9, 2023 8:10 AM EDT · Updated 2 months ago

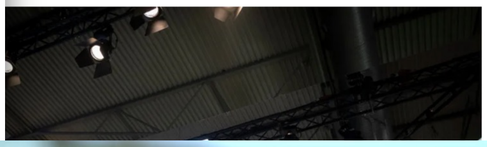
TC

Startups

IBM rolls out new generative AI features ...odels

...kyle_l_wiggers / 8:00 AM EDT • September 7, 2023

Comment



FORBES › INNOVATION › CLOUD

IBM Enables Safe Enterprise AI with Granite Foundation Models

Steve McDowell Contributor
Principal analyst and founding partner at NAND Research.

Follow

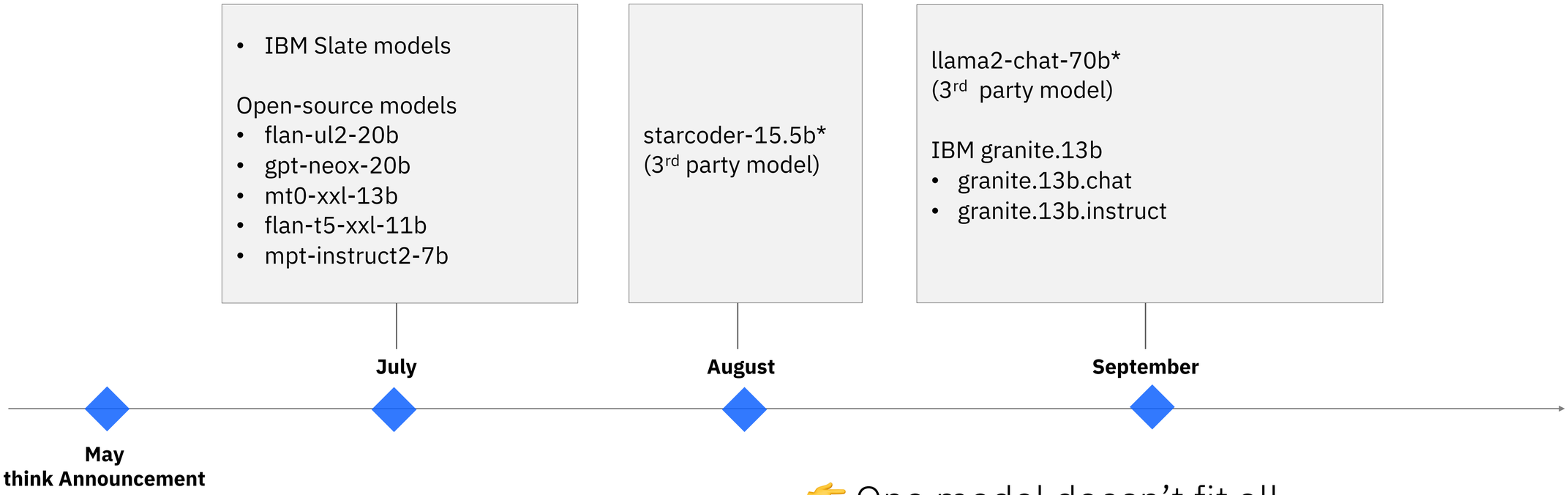
0

Oct 3, 2023, 08:47pm EDT

Listen to article 7 minutes

IBM watsonx NURPHOTO VIA GETTY IMAGES

SaaS Foundation Models available today



Models will be revised quarterly and may be removed from the SaaS collection based on performance and market demand. The removed models will remain in the watsonx.ai Foundation Model Library for BYOM use cases.

* Llama 2 and StarCoder have non-standard open-source terms with additional Acceptable Use Policies.

👉 One model doesn't fit all
👉 Bigger is not always better
👉 Specialized models can produce better results than larger, general-purpose models, and do so with lower infrastructure requirements to achieve improved price-performance

Easy win with Granite Series

Client Protection -- IBM stands behind Granite models and offers a peace of mind to clients by
- not requiring them to indemnify IBM for their use of its models, and
- not capping its IP indemnification liability.

Targeted for Business -- superior performance for finance tasks like credit risk assessment, insurance QA, Conversational Finance QA, and summarization (details)

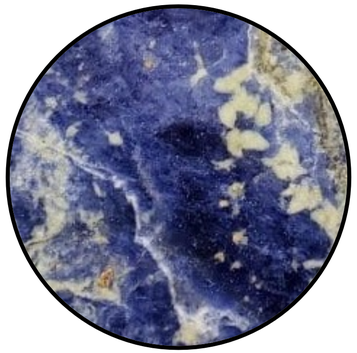
Built for Trust -- IBM has published details of training data sets to demonstrate its commitment to transparency and responsible AI.

	Indemnification	Company's Obligation to Indemnify	No Cap for Company's Indemnity	Customer is not obliged to Indemnify Company	Company's Obligation to Indemnify for Output
IBM		✓	✓	✓	x
Adobe		✓	x	◐	✓
Oracle		✓	x	x	x
Google		✓	✓	x	x
OpenAI		x		x	x
Microsoft		✓	✓	x	✓
AWS		✓	✓	x	✓
Salesforce		✓	x	x	x
Cohere		✓	x	x	x
Meta (LLaMa)		x		x	x
Anthropic		x		x	x
Google (PaLM)		x		x	x

Granite’s Differentiation

Trusted	IBM’s AI is responsible and governed
Targeted	IBM’s AI is designed for enterprise and targeted at business domains
Open	IBM’s AI is transparent, publishing key details such as training dataset names

Introducing Granite

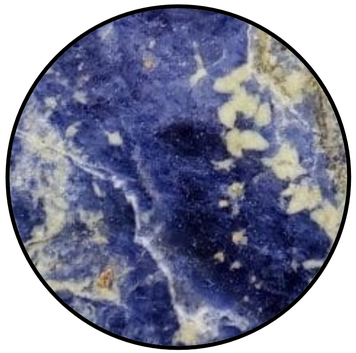


granite.13b
(.instruct, .chat)

Model Size:
13B Parameters

Intended Use:
English language tasks including generation, summarization, classification, extraction, and question answering

Available today in watsonx.ai



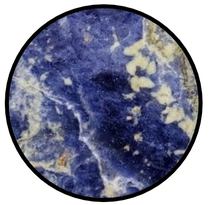
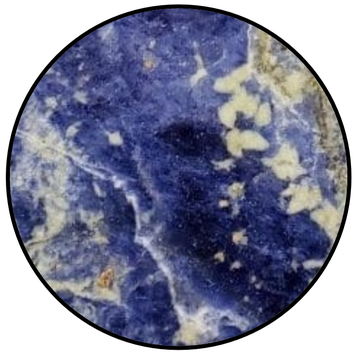
granite.20b.code
(.cobol, .ansible)

Model Size:
20B Parameters

Intended Use:
Code generation, summarization, and translation

Available in Q4 in watsonx Code Assistant

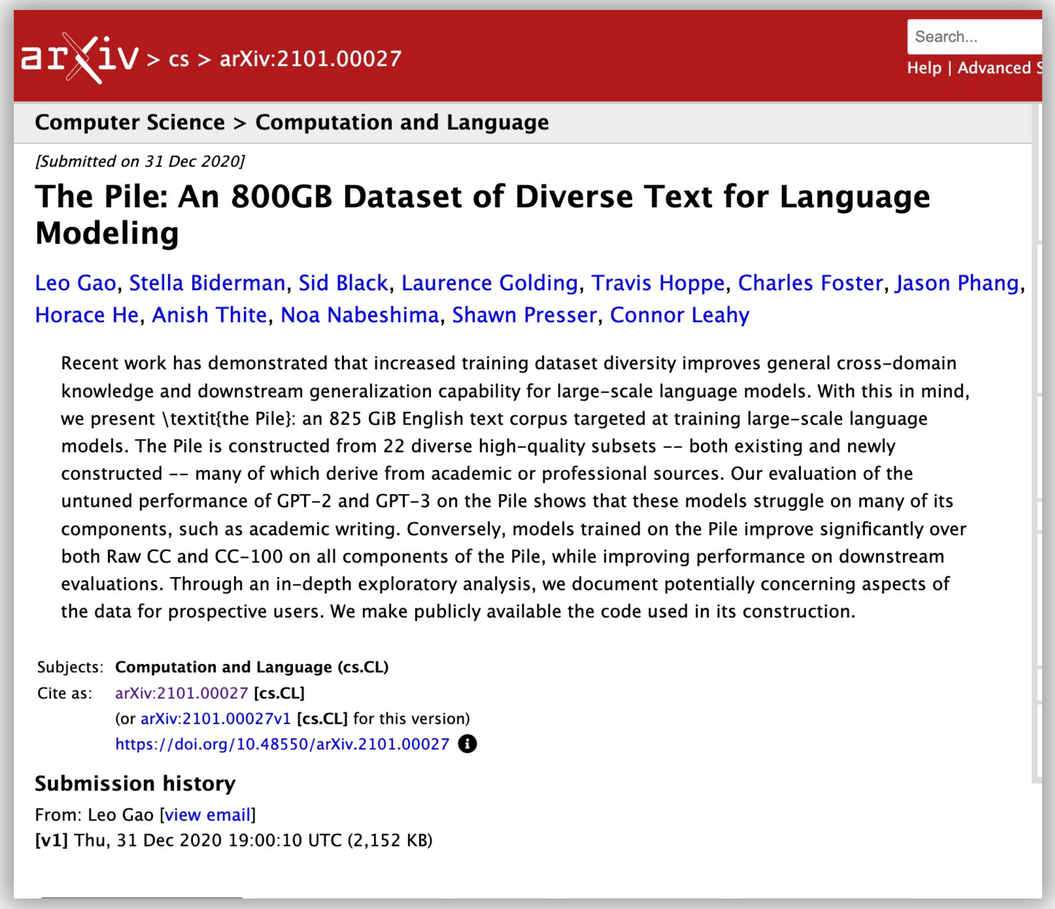
LLM =
Architecture + Data + Compute + Alignment

Model	Architecture	Data	Compute	Alignment
 granite.13b	• 13b parameters • 8k context length • Decoder-only	1T tokens of English language	• 256 A100s on CCC (original compute architecture) • 1056 GPU Hours	.instruct Designed to follow short instructions and returns concise response .chat Designed for human / agent conversations and question answering
 granite.20b.code	• 20b parameters • 8k – 32k context length • Decoder-only	1.6T tokens of 100+ different coding languages, including • Cobol, • Ansible, • Python, • Java, • and many more	• 768 A100s on Vela (new super-computer from IBM Research)	.cobol Designed to translate cobol to java .ansible Designed to generate ansible playbooks for IT Ops automation

Granite’s Differentiation

Trusted	IBM’s AI is responsible and governed
Targeted	IBM’s AI is designed for enterprise and targeted at business domains
Open	IBM’s AI is transparent, publishing key details such as training dataset names

Public Data – “The Pile”



arXiv > cs > arXiv:2101.00027

Search... Help | Advanced S

Computer Science > Computation and Language

[Submitted on 31 Dec 2020]

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, Connor Leahy

Recent work has demonstrated that increased training dataset diversity improves general cross-domain knowledge and downstream generalization capability for large-scale language models. With this in mind, we present \textit{the Pile}: an 825 GiB English text corpus targeted at training large-scale language models. The Pile is constructed from 22 diverse high-quality subsets -- both existing and newly constructed -- many of which derive from academic or professional sources. Our evaluation of the untuned performance of GPT-2 and GPT-3 on the Pile shows that these models struggle on many of its components, such as academic writing. Conversely, models trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, while improving performance on downstream evaluations. Through an in-depth exploratory analysis, we document potentially concerning aspects of the data for prospective users. We make publicly available the code used in its construction.

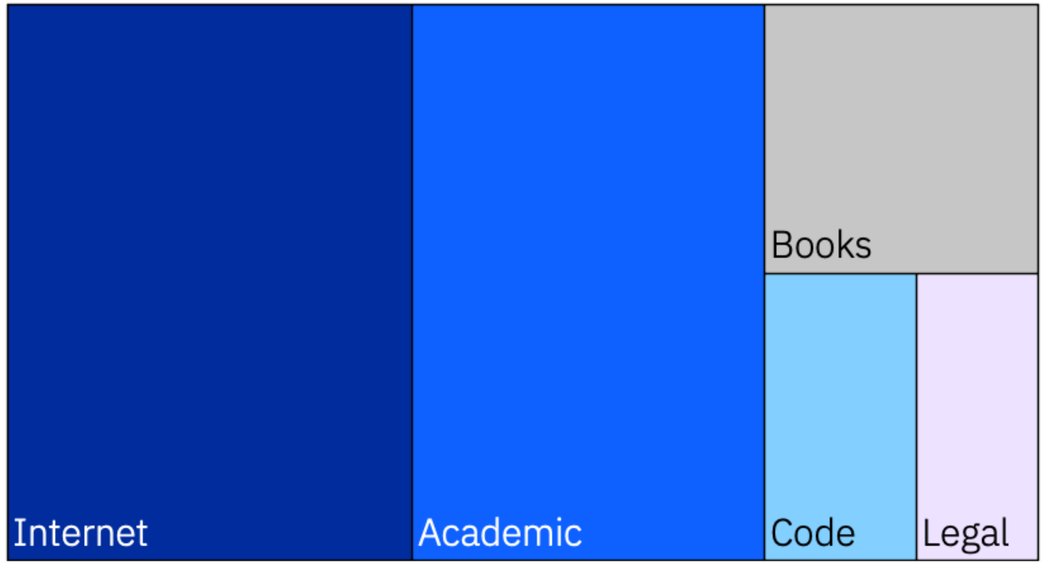
Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:2101.00027 **[cs.CL]**
(or arXiv:2101.00027v1 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2101.00027

Submission history

From: Leo Gao [view email]
[v1] Thu, 31 Dec 2020 19:00:10 UTC (2,152 KB)

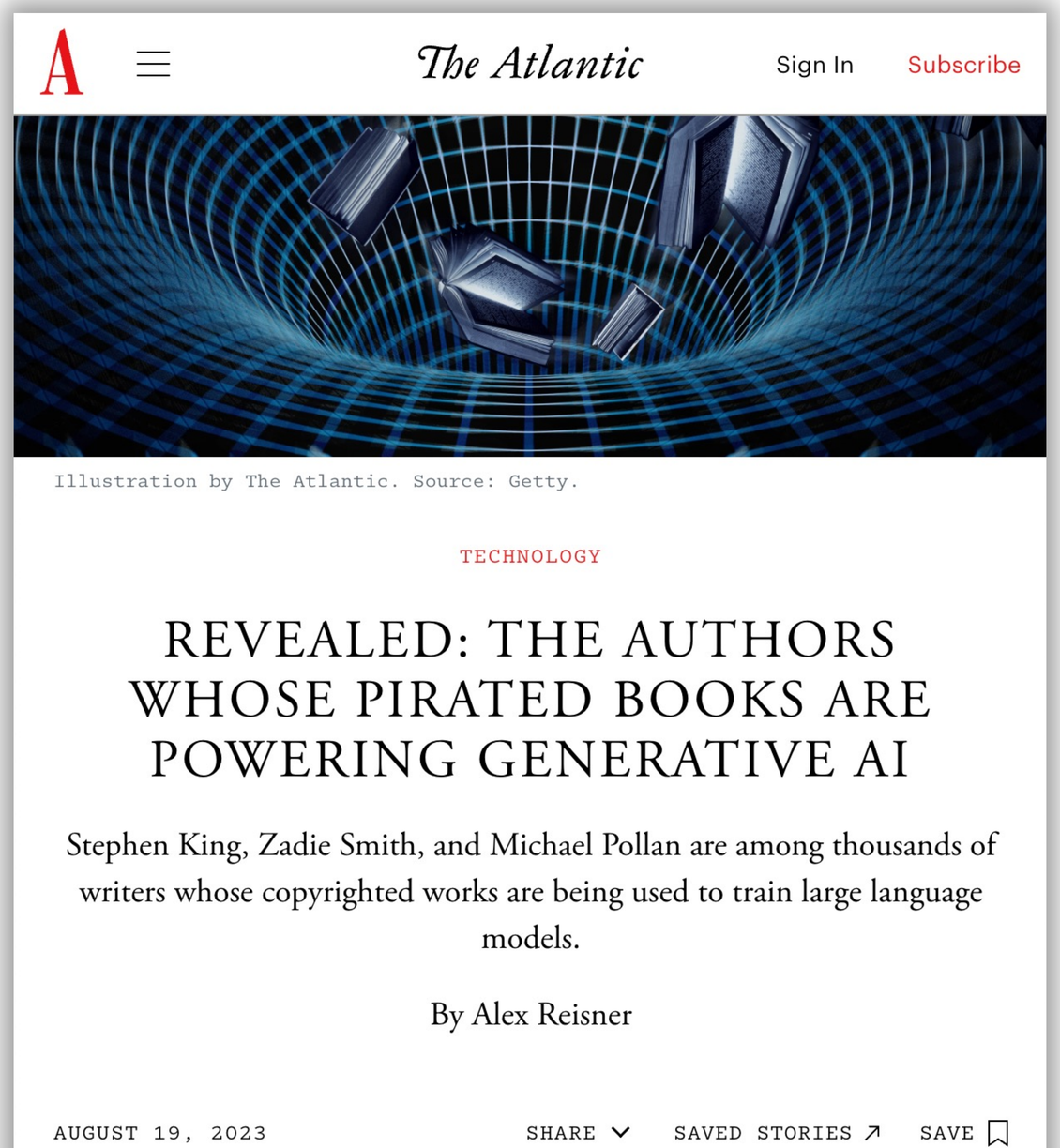
Source: https://arxiv.org/pdf/2101.00027.pdf

“The Pile” (800GB)



Effective size of datasets by domain type

Public Data – “The Pile”



The Atlantic Sign In Subscribe

Illustration by The Atlantic. Source: Getty.

TECHNOLOGY

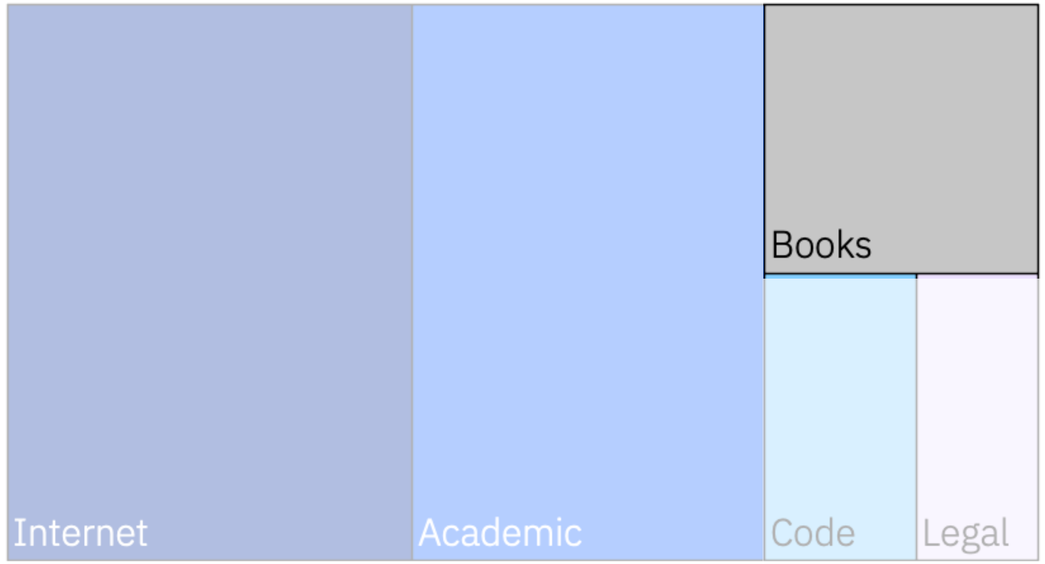
REVEALED: THE AUTHORS WHOSE PIRATED BOOKS ARE POWERING GENERATIVE AI

Stephen King, Zadie Smith, and Michael Pollan are among thousands of writers whose copyrighted works are being used to train large language models.

By Alex Reisner

AUGUST 19, 2023 SHARE SAVED STORIES SAVE

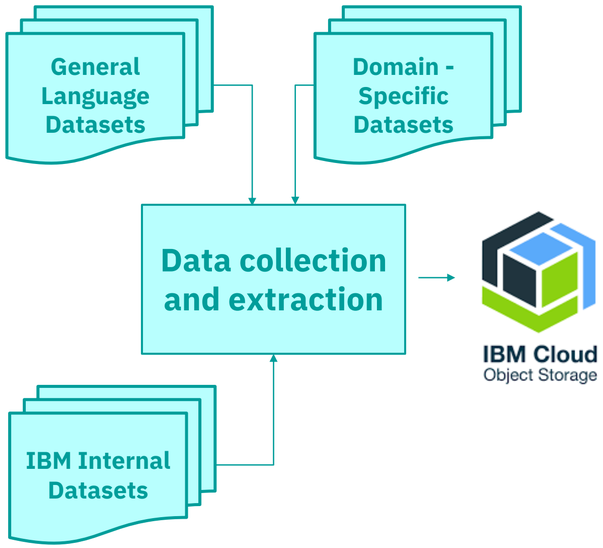
“The Pile” (800GB)



Effective size of datasets by domain type

IBM’s trusted and transparent data curation process for Generative AI development

Dataset Acquisition



Key Data Acquisition Stakeholders:
- Domain Experts
- IBM Legal
- Chief Privacy Office
- Responsible Business Alignment (IBM Procurement)
- IBM’s AI Ethics Board

Dataset Preprocessing (model agnostic)

Data Deduplication

- Document id generation
- Deduplication
- Sentence splitting

Data Annotation

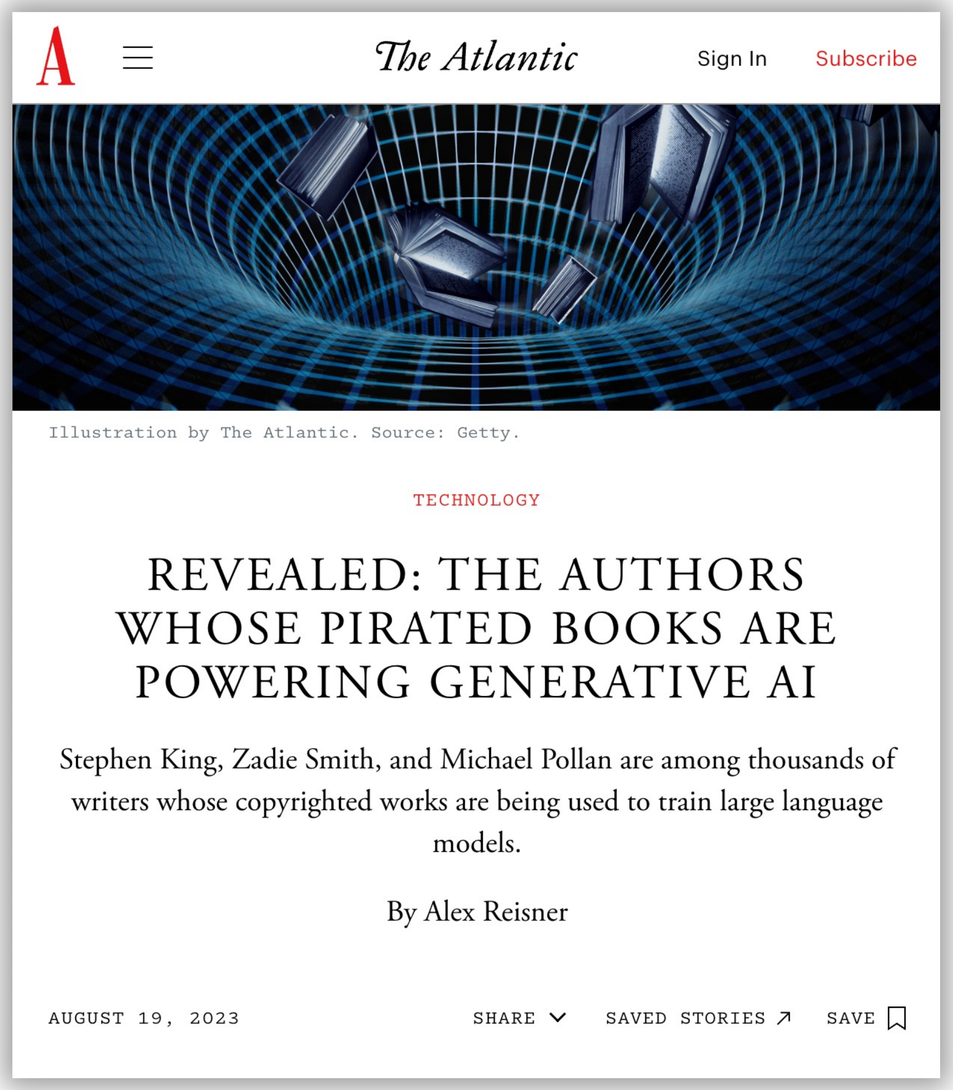
- Language detection
- Hate, abuse, profanity
- Document quality
- URL Blocklisting
- License detection (code)

Data Preprocessing (model specific)

Data Filtering

Tokenization

An Example of Datasets Excluded from Data Pile v0.3



The Atlantic Sign In Subscribe

Illustration by The Atlantic. Source: Getty.

TECHNOLOGY

REVEALED: THE AUTHORS WHOSE PIRATED BOOKS ARE POWERING GENERATIVE AI

Stephen King, Zadie Smith, and Michael Pollan are among thousands of writers whose copyrighted works are being used to train large language models.

By Alex Reisner

AUGUST 19, 2023 SHARE SAVED STORIES SAVE

Books3

~200,000 copyrighted books

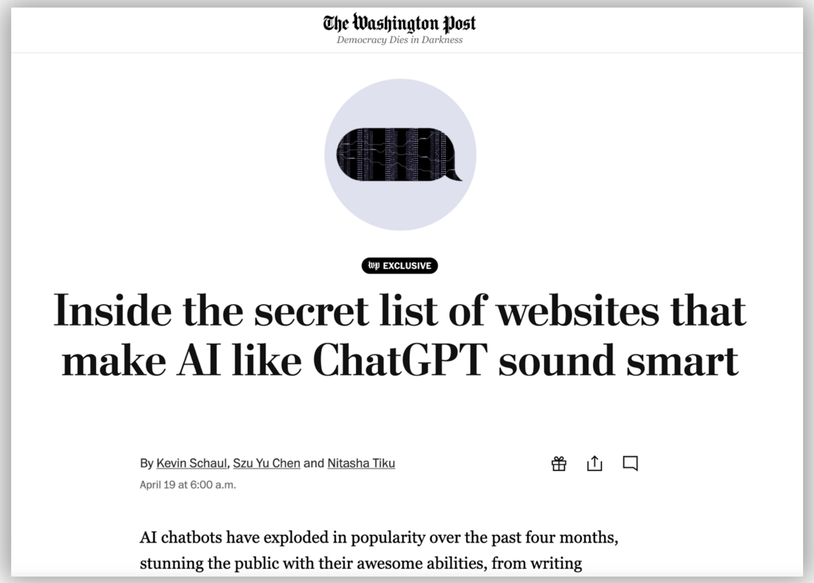
{'title': '07 LEGO Ninjago - The Search For Zane (Scholastic) - Kate Howard (retail)'
'text': '\n\nTITLE PAGE\n\nFROM THE JOURNAL OF SENSEI GARMADON\n\nCHAPTER 1\n\nCHAPTER 2\n\nCHAPTER 3\n\nCHAPTER 4\n\nCHAPTER 5\n\nCHAPTER 6\n\nCHAPTER 7\n\nCHAPTER 8\n\nCHAPTER 9\n\nCOPYRIGHT\n\nThroughout Ninjago", five ninja are well-known for their speed, strength, and of course the elemental powers that help them protect our world from evil. But there are others who possess some of the same powers as the ninja. Others who may not always use their powers for good.\n\nBefore now, the ninja believed they were special. They di.....'}

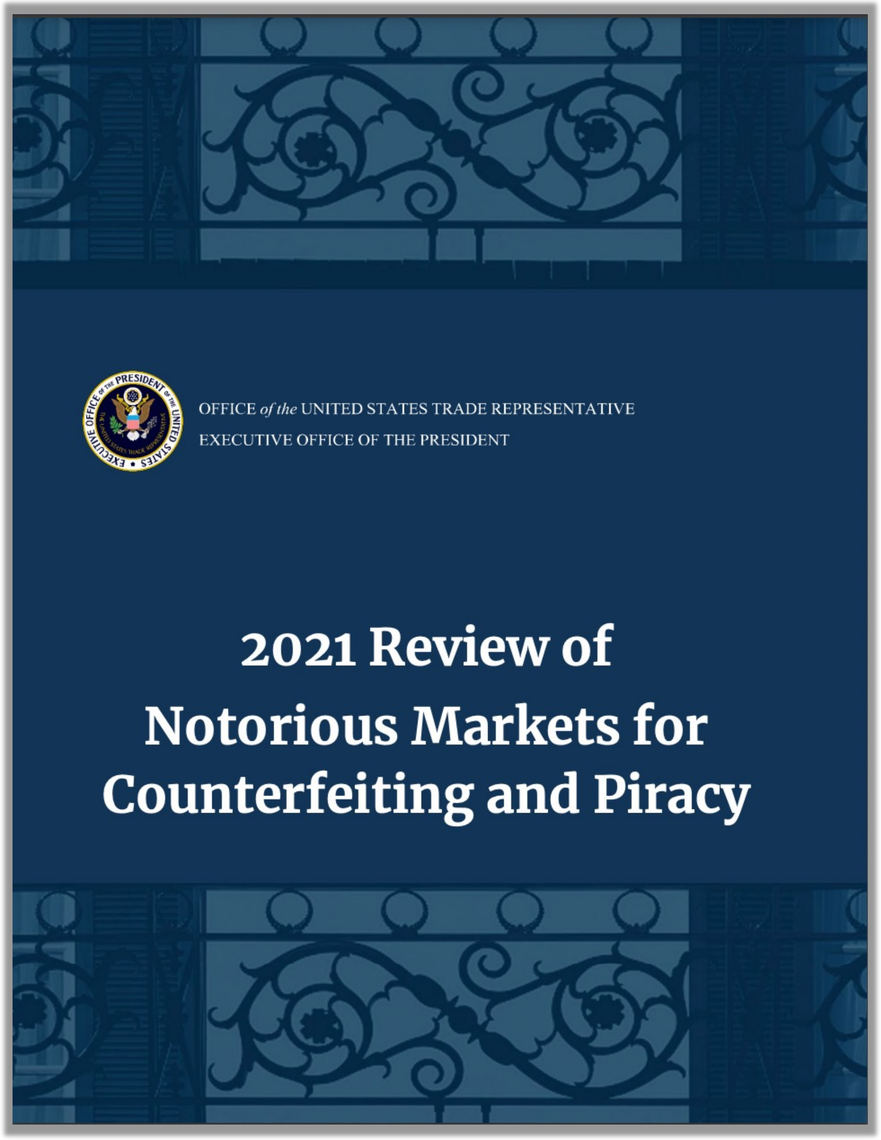
https://huggingface.co/datasets/the_pile_books3

An Example of Datasets Excluded from Data Pile v0.3

US Federal Report on Websites Known to Perpetuate Pirated Information

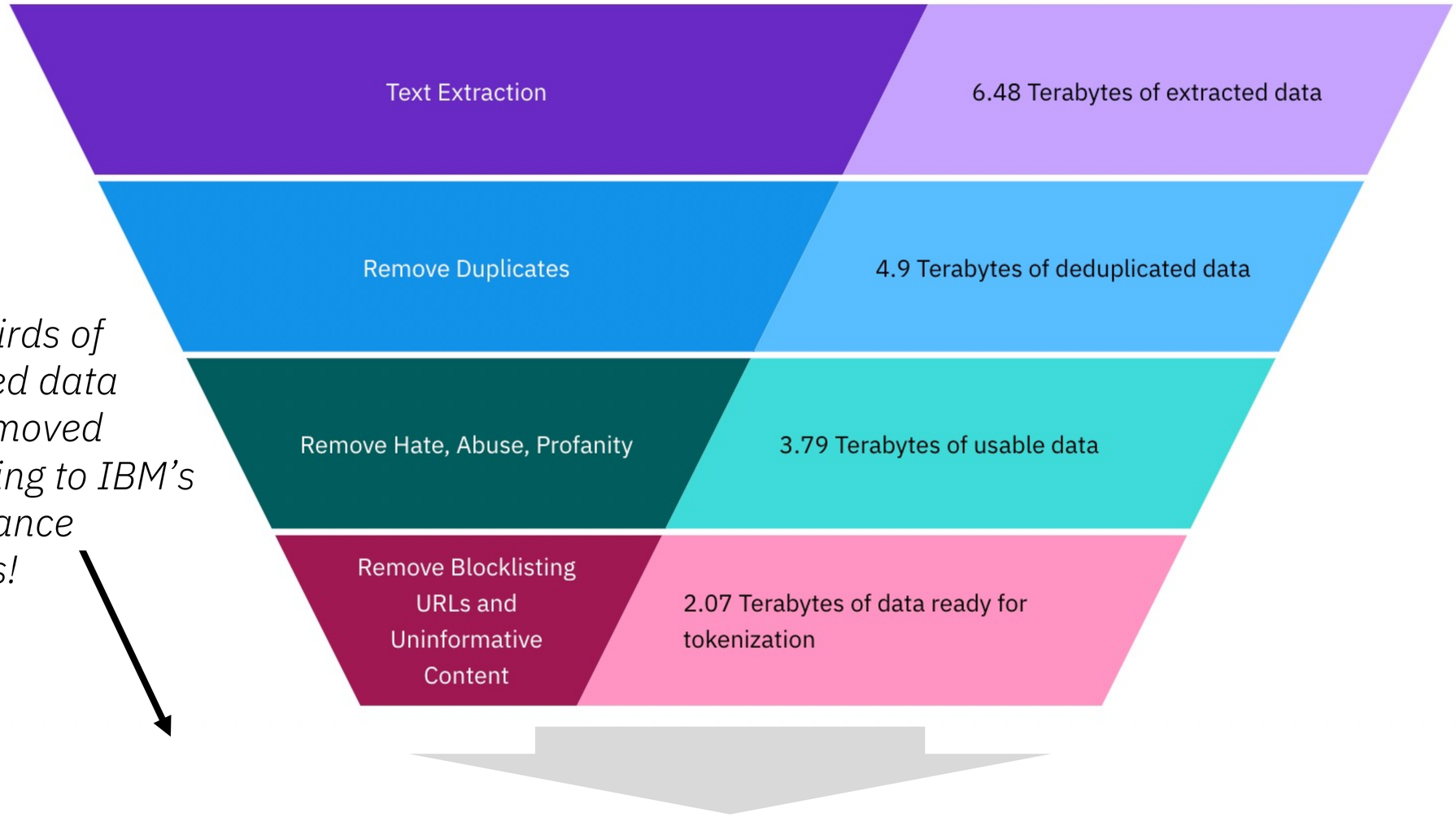
E.g. ThePirateBay:
The most frequently visited bit torrent index site in the world





Granite.13b:
Training data governance funnel

Two-thirds of collected data was removed according to IBM's governance process!

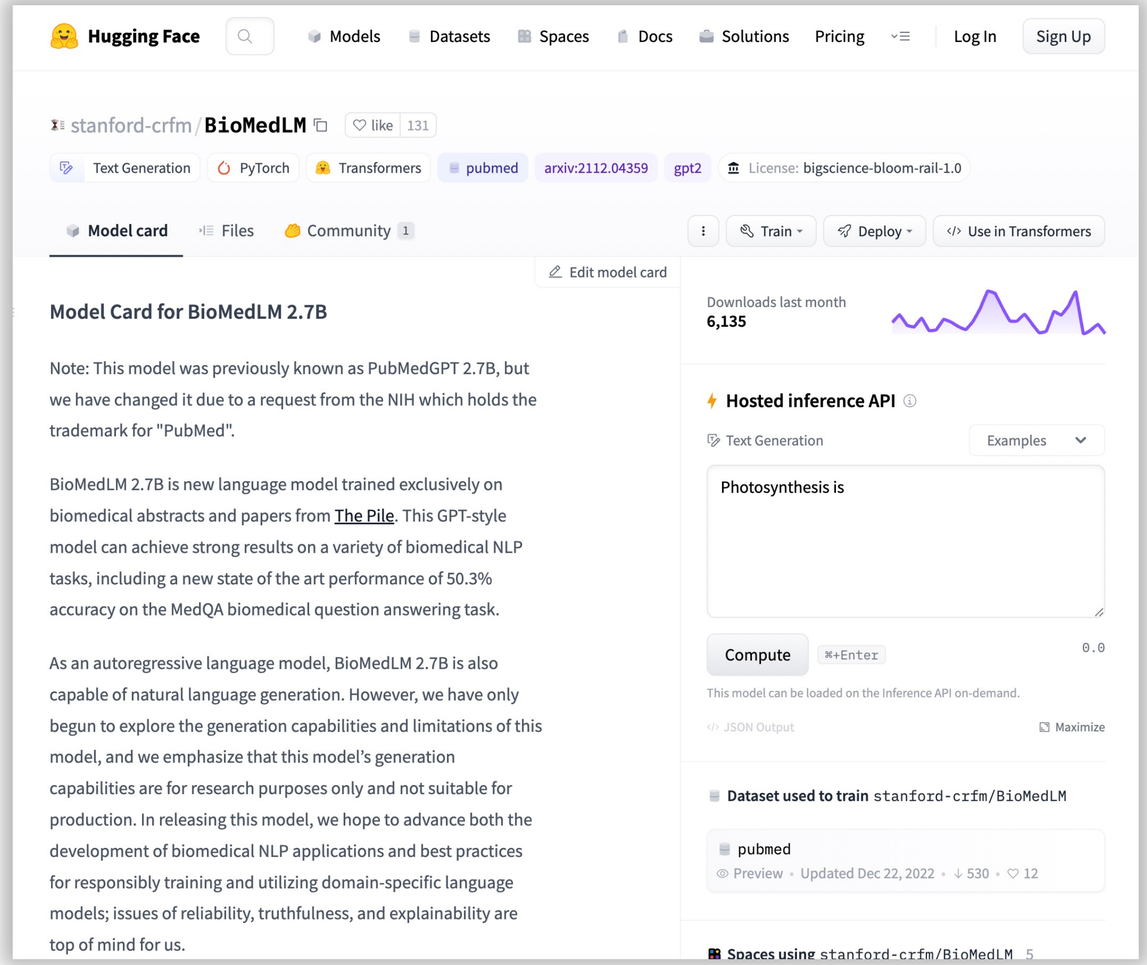


1T Tokens of data for training

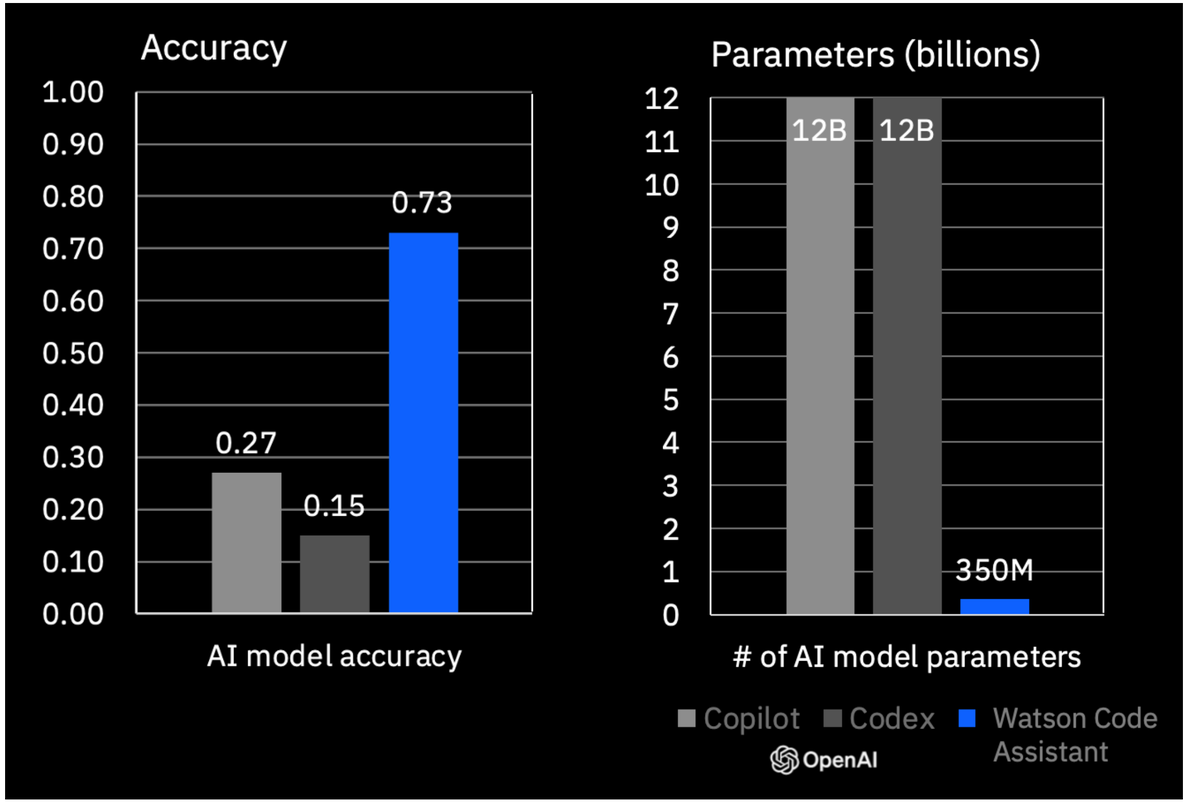
Granite’s Differentiation

Trusted	IBM’s AI is responsible and governed
Targeted	IBM’s AI is designed for enterprise and targeted at business domains
Open	IBM’s AI is transparent, publishing key details such as training dataset names

Small, specialized models can outperform large, generalist models



BioMedLM (Stanford)

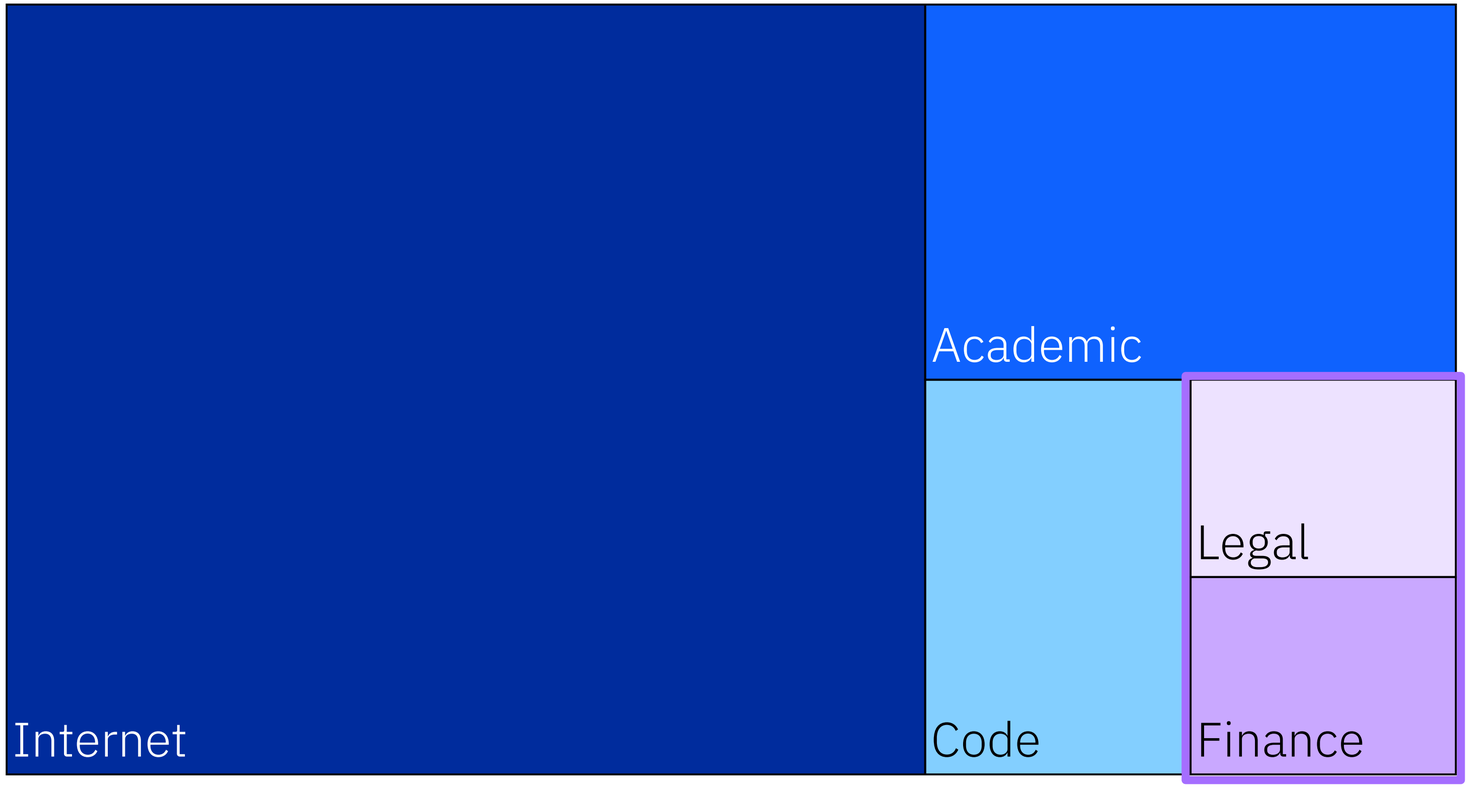


Watson Code Assistant

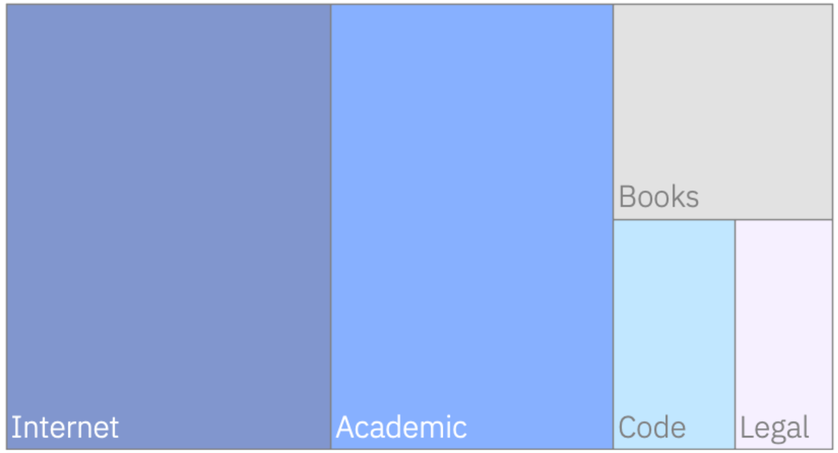
Granite.13b

Specializing LLM training on enterprise-relevant datasets

Datasets used in training granite.13b (2.4 TB)



"The Pile" (800GB)



10% of data comes from specialized enterprise domains

Research Evaluation:
BloombergGPT Finance Panel (3 tasks)

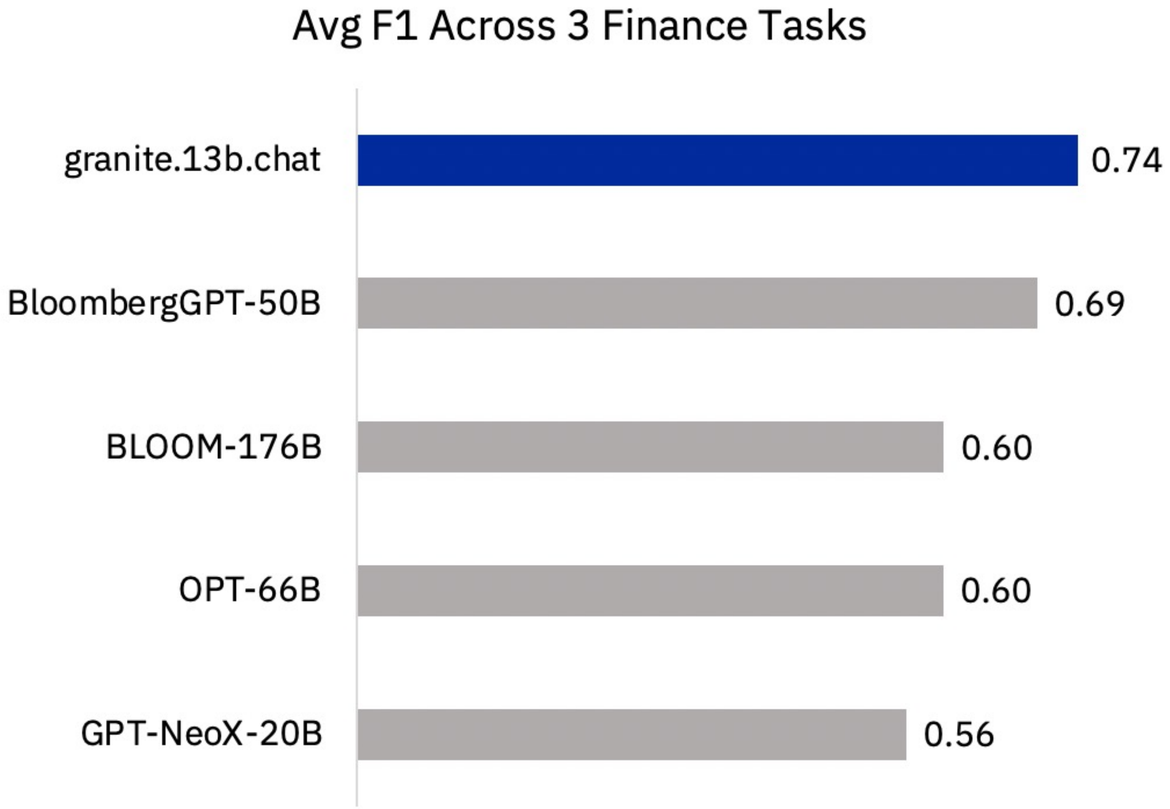
Goal: Compare performance to the leading finance-specialized model.

BloombergGPT Finance Panel

Bloomberg shared the exact test data for 3 out of 5 of their external finance benchmarks:

Sentiment analysis
- Financial Phrase Bank
- Financial Question Answering

Classification
- News Headline Classification



Evaluation Caveats:

The BloombergGPT paper excluded many key competitor models from their analysis.

BloombergGPT analyzed performance on many other metrics that we did not have data for.

Source: BloombergGPT: A Large Language Model for Finance (2023)

Granite.13b

Finance Evaluation (11 tasks)

Goal: Compare financial performance to the leading general-purpose models

IBM Research Finance Panel

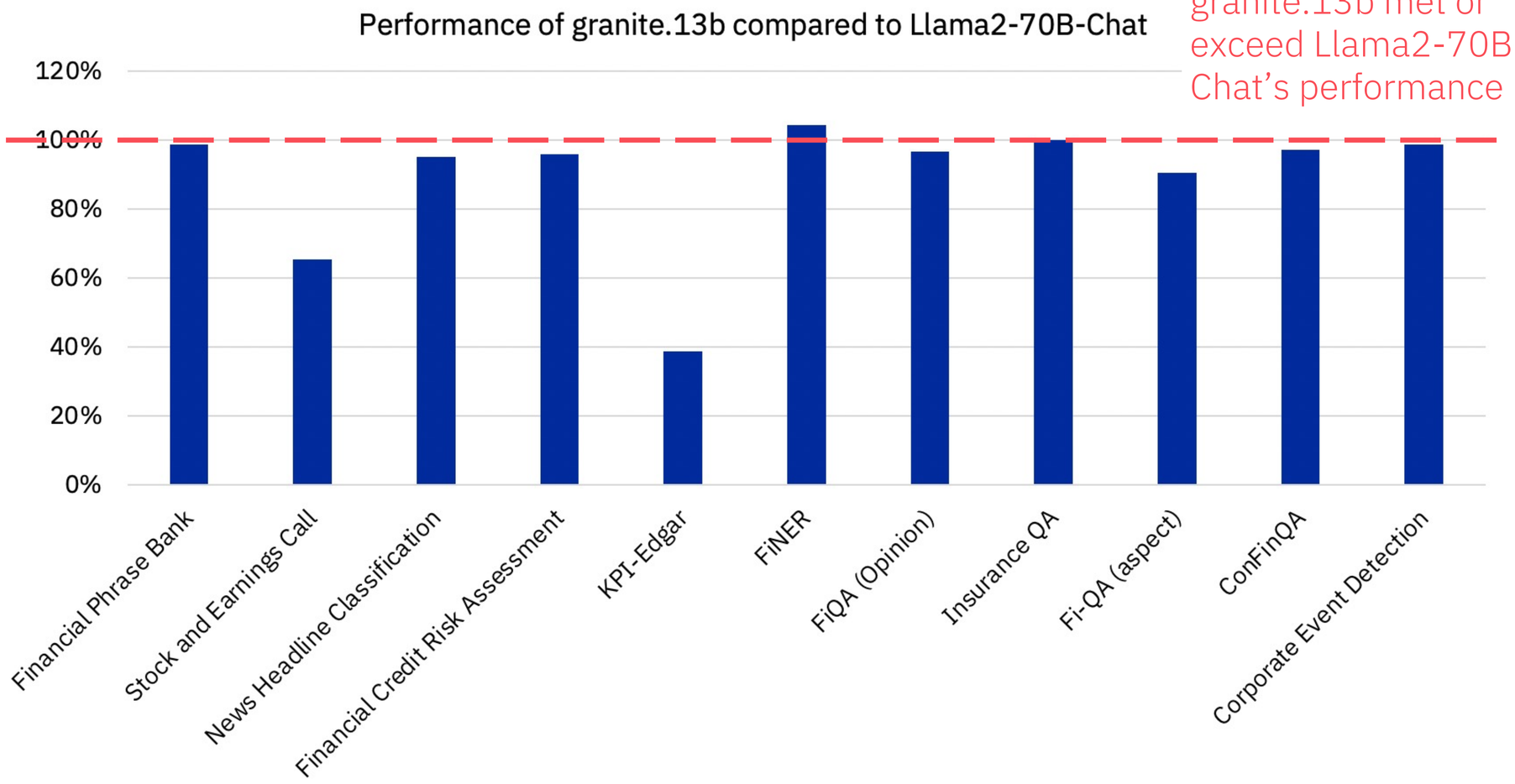
Sentiment analysis
- Financial Phrase Bank
- Stock and Earnings Call Transcripts
- Financial Question Answering

Classification
- News Headline Classification

Entity Extraction
- Edgar SEC Filings (x2)
- Financial Credit Risk Assessment

Question and Answering
- Financial Question Answering (x2)
- Insurance

Summarization
- Corporate Event Detection



On 9 out 11 tasks granite.13b met or exceed Llama2-70B-Chat's performance

Granite’s Differentiation

Trusted	IBM’s AI is responsible and governed
Targeted	IBM’s AI is designed for enterprise and targeted at business domains
Open	IBM’s AI is transparent, publishing key details such as training dataset names

https://www.ibm.com/downloads/cas/X9W4O6BM

Granite Foundation Models

IBM Research

***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 base model is the jumping-off point for two variants: granite.13b.instruct and granite.13b.chat. The first variant, 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. The second variant, granite.13b.chat, has undergone a novel contrastive fine-tuning after supervised fine-tuning to further improve the model's instruction following, mitigate certain notions of harms, and encourage its outputs to follow certain social norms and have some notion of helpfulness [2]–[4]. We

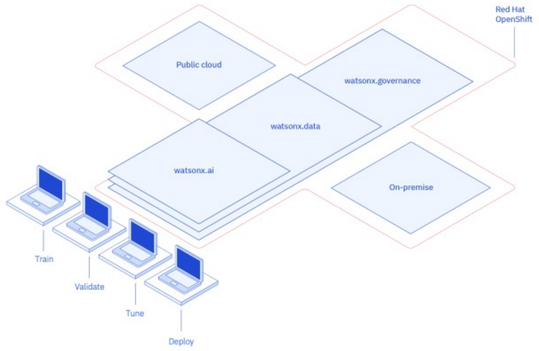


Fig. 1. A conceptual diagram of the watsonx platform.

has been trained on 1 trillion tokens created with the GPT-NeoX 20B tokenizer [6], and has a context length of 8 thousand 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. Other series of models that IBM is developing are Sandstone: encoder-decoder models designed to be tuned for specific tasks and Obsidian: modular universal transformer models suitable for high inference efficiency.

II. Data Sources

At the time of granite.13b's 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, for the purposes of training granite.13b, 1 trillion tokens were 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.

IBM’s AI Principles

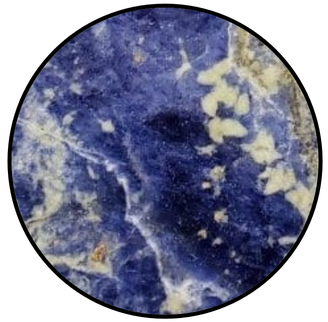
Trusted	IBM’s AI is responsible and governed
Targeted	IBM’s AI is designed for enterprise and targeted at business domains
Open	IBM’s AI is transparent, publishing key details such as training dataset names

Granite Roadmap

IBM’s enterprise-grade, decoder-only family of large language models

Research Projections Only, See Roadmap on Seismic for Formal Committed Product Deliveries

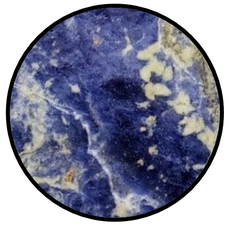
Early Access Targeted EOY, See Seismic Roadmap for Committed Deliveries



granite.13b.v2

2T tokens of English language

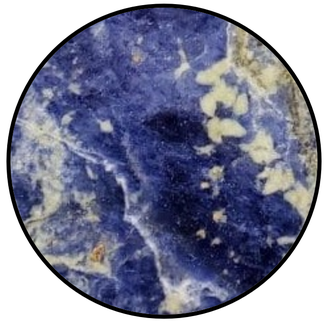
RLAIF-based alignment for .chat variant



granite.8b .japanese

1T tokens of English language

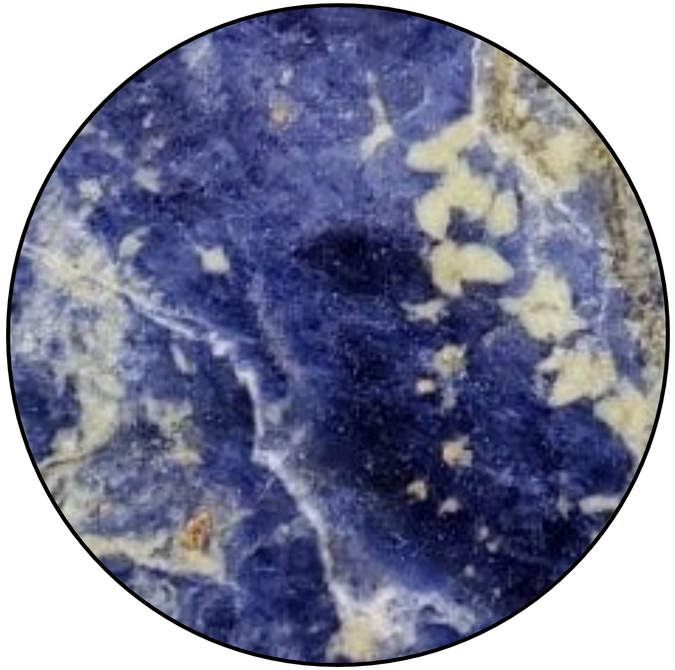
1T tokens total of Japanese



granite.13b .multilingual

1T tokens of English language

1T tokens total of Portuguese, Spanish, French, and German



granite.20b – granite.200b

2T-5T+ tokens of English, Code, and Multilingual Data

Picking the right LLM
- price/performance
- GPU compute
- Trust (regulations)

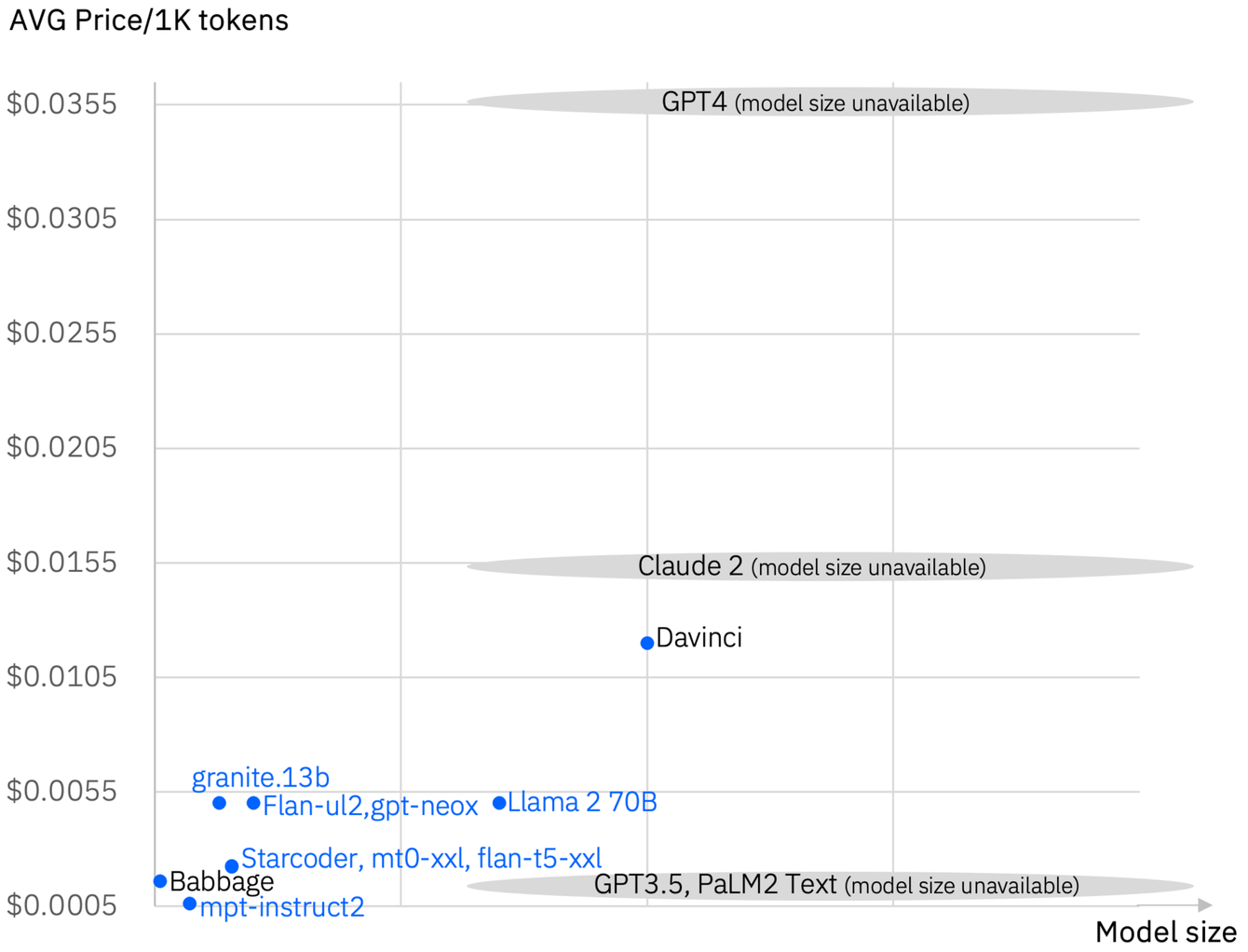
Picking the right platform
- Hybrid Cloud
- Governance
- Security, privacy, scalability

Model Name	Provider	Use cases	Context length	Price (USD)[1]
granite-13b-chat Featured model	IBM	Supports Q&A and Generate tasks.	8192	0.005
granite-13b-instruct Featured model	IBM	Supports Extract, Summarize and Classify tasks.	8192	0.005
llama-2-70b-chat	Meta	Supports Q&A, Generate, Extract, Summarize, Classify tasks. Optimized for dialogue use cases.	4096	0.005
flan-t5-xxl-11b	Google	Supports Q&A, Generate, Summarize, Classify tasks.	4096	0.0018
flan-ul2-20b	Google	Supports Q&A, Generate, Extract, Summarize, Classify tasks.	4096	0.005
gpt-neox-20b	EleutherAI	Supports Q&A and Generate tasks. Works well with special characters, which can be useful for generating structured output.	8192	0.005
mpt-7b-instruct2	Mosaic, fine-tuned by IBM	Supports Q&A and Generate tasks.	2048	0.0006
mt0-xxl-13b	BigScience	Supports Q&A, Generate, Extract, Summarize, Classify tasks.	4096	0.0018
starcoder-15.5b	BigCode	Task specific model for code by generating and translating code from a natural language prompt.	8192	0.0018

			Base FM - Inference
	Model Offerings	Estimated # of Parameters	Average Price/1K tokens
watsonx	Granite.13B	13B	$0.0050
	Llama 2 70B	70B	$0.0050
	StarCoder	15.5B	$0.0018
	flan-ul2-20b	20B	$0.0050
	gpt-neox-20b	20B	$0.0050
	mt0-xxl-13b	13B	$0.0018
	mpt-instruct2-7b	7B	$0.0006
	flan-t5-xxl-11b	11B	$0.0018
Microsoft Azure/Open AI	Text-Babbage	1.3B	$0.0016
	Text-Davinci	175B	$0.0120
	gpt-3.5-turbo		$0.0016[1]
	gpt-4 (8k context)		$0.0360[2]
Anthropic	Claude 2		$0.0153[3]
Google	PaLM 2 Text		$0.002[4]

1. Translated from original price: Prompt $0.0015/k tokens, Completion: $0.002/k tokens; assuming 4x length of prompt vs. completion to average.
2. Translated from original price: Prompt $0.03/k tokens, Completion: $0.06/k tokens; assuming 5x length of prompt vs. completion to average.
3. Translated from original price: Prompt $0.01102/k tokens, Completion: $0.03268/k tokens; assuming 5x length of prompt vs. completion to average.
4. Translated from original price: $0.002/1K tokens ($0.0005/1K characters), characters are counted by UTF-8 code points; assuming 1 token = 4 characters)

Pricing data as of Oct 5,2023.



The cost of summarizing a 30 min meeting with OpenAI GPT4 is 6.6 times higher than Llama 2 70B model on watsonx.ai [simple math]

watsonx: Tailored to the enterprise

Open

→ Based on the best open technologies available.

→ Access to the innovation of the open community and multiple models.

Trusted

→ Offering security and data protection.

→ Governance, transparency, and ethics that support increasing regulatory compliance demands.

Targeted

→ Designed for targeted for business use cases, that unlock new value.

→ Models that can be tuned to your proprietary data.

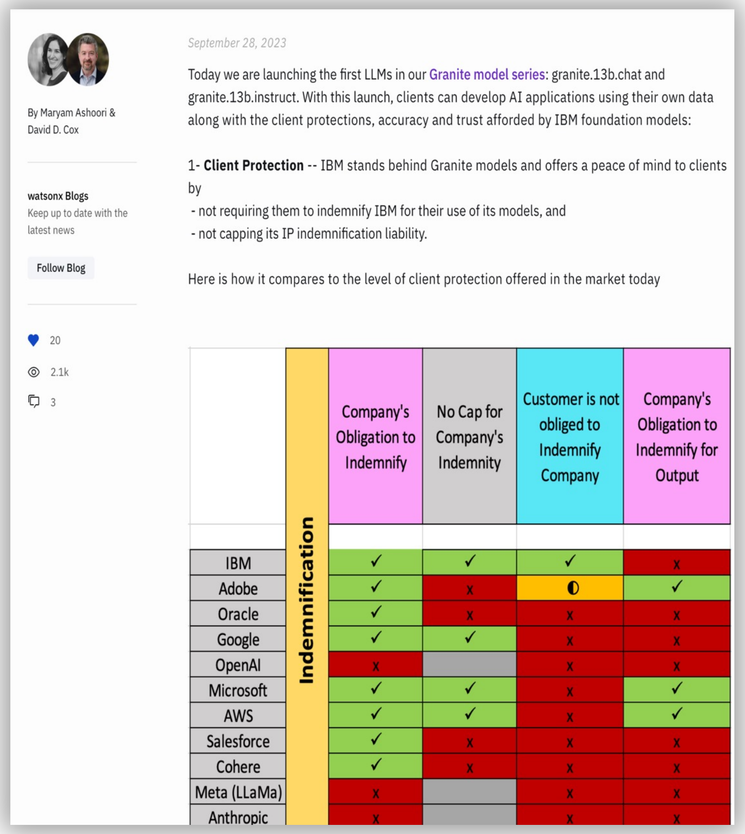
Empowered

→ A platform to bring your own data and AI models that you tune, train, deploy, and govern.

→ Running anywhere, designed for scale and widespread adoption.

Where to start

Granite’s internal blog



By Maryam Ashoori & David D. Cox

watsonx Blogs
Keep up to date with the latest news

Follow Blog

20
2.1k
3

September 28, 2023

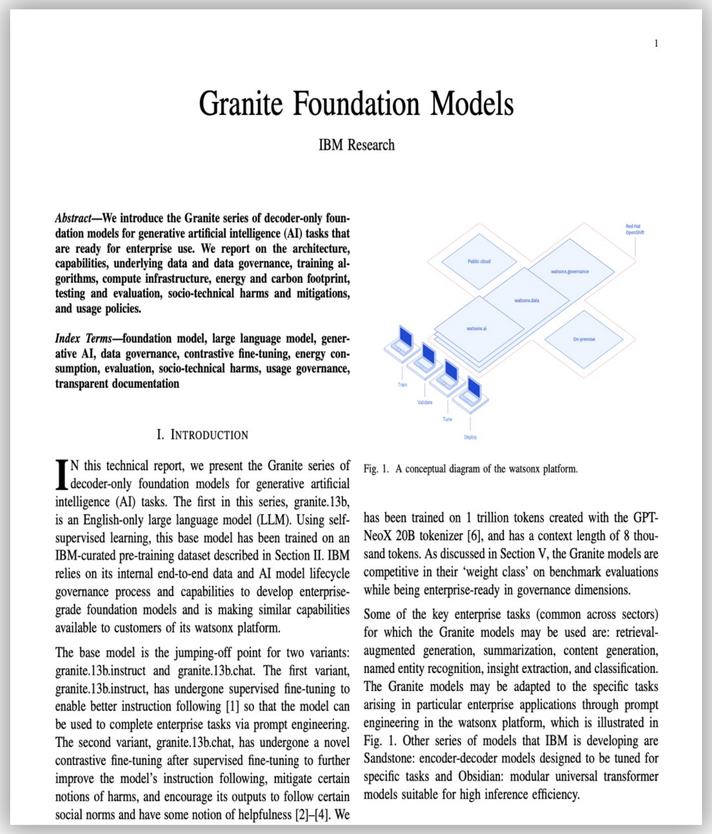
Today we are launching the first LLMs in our Granite model series: granite.13b.chat and granite.13b.instruct. With this launch, clients can develop AI applications using their own data along with the client protections, accuracy and trust afforded by IBM foundation models:

1- **Client Protection** -- IBM stands behind Granite models and offers a peace of mind to clients by
- not requiring them to indemnify IBM for their use of its models, and
- not capping its IP indemnification liability.

Here is how it compares to the level of client protection offered in the market today

	Indemnification	Company's Obligation to Indemnify	No Cap for Company's Indemnity	Customer is not obliged to Indemnify Company	Company's Obligation to Indemnify for Output
IBM		✓	✓	✓	x
Adobe		✓	x	◐	✓
Oracle		✓	x	x	x
Google		✓	✓	x	x
OpenAI		x		x	x
Microsoft		✓	✓	x	✓
AWS		✓	✓	x	✓
Salesforce		✓	x	x	x
Cohere		✓	x	x	x
Meta (LLaMa)		x		x	x
Anthropic		x		x	x

Granite’s technical paper



1

Granite Foundation Models

IBM Research

***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.

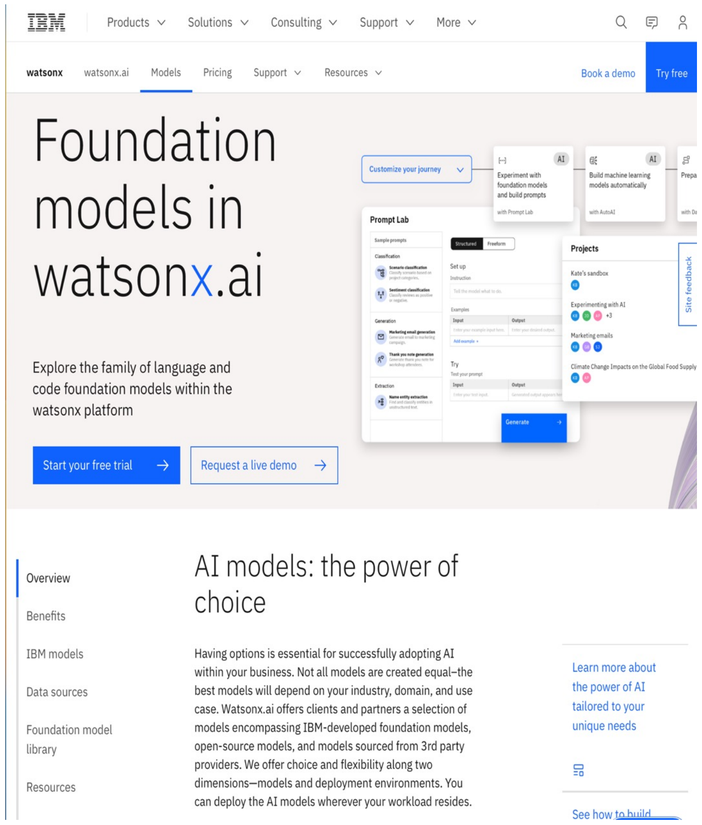
The base model is the jumping-off point for two variants: granite.13b.instruct and granite.13b.chat. The first variant, granite.13b.instruct, has undergone supervised fine-tuning to enable better instruction following [1] so that the model can be used to complete enterprise tasks via prompt engineering. The second variant, granite.13b.chat, has undergone a novel contrastive fine-tuning after supervised fine-tuning to further improve the model’s instruction following, mitigate certain notions of harms, and encourage its outputs to follow certain social norms and have some notion of helpfulness [2]–[4]. We

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

has been trained on 1 trillion tokens created with the GPT-NeoX 20B tokenizer [6], and has a context length of 8 thousand 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. Other series of models that IBM is developing are Sandstone: encoder-decoder models designed to be tuned for specific tasks and Obsidian: modular universal transformer models suitable for high inference efficiency.

Product page



IBM Products Solutions Consulting Support More

watsonx watsonx.ai Models Pricing Support Resources Book a demo Try free

Foundation models in watsonx.ai

Explore the family of language and code foundation models within the watsonx platform

Start your free trial → Request a live demo →

Overview
Benefits
IBM models
Data sources
Foundation model library
Resources

AI models: the power of choice

Having options is essential for successfully adopting AI within your business. Not all models are created equal–the best models will depend on your industry, domain, and use case. Watsonx.ai offers clients and partners a selection of models encompassing IBM-developed foundation models, open-source models, and models sourced from 3rd party providers. We offer choice and flexibility along two dimensions—models and deployment environments. You can deploy the AI models wherever your workload resides.

Learn more about the power of AI tailored to your unique needs

See how to build

- Reason to call
- Watsonx foundation model roadmap

Thank you!