

WHITEPAPER

# Fine-Tuning Granite Model Using InstructLab Methodology for Dodd-Frank Act (Section 1033 Compliance)



# Table of contents

<b>Abstract</b>	<b>03</b>
<b>Introduction</b>	<b>04</b>
<b>Model Training Process</b> (Granite + InstructLab)	<b>05</b>
<b>Key Findings and Insights</b>	<b>13</b>
<b>Technical Details and Performance Metrics</b>	<b>14</b>
<b>Case Studies:</b> Applying the Trained Granite Model in Banking Compliance	<b>16</b>
<b>Conclusion and Future Directions</b>	<b>18</b>
<b>References</b>	<b>19</b>
<b>Authors' bio</b>	<b>20</b>

# Abstract

---

The increasing adoption of open banking regulations has created a demand for AI tools that can interpret complex financial legislation with precision. In the U.S., the Consumer Financial Protection Bureau's (CFPB's) Personal Financial Data Rights rule (Section 1033 of the Dodd–Frank Act) will reshape how banks share consumer data, with phased deadlines extending through 2030.

For compliance leaders, the challenge goes beyond meeting technical API requirements. They must ensure that decision-makers, frontline employees, and customer-facing teams interpret the rule accurately and consistently across all operations.

This white paper presents a case study on customizing IBM's Granite-3-8B large language model (LLM) to interpret Section 1033 using IBM's InstructLab methodology on watsonx.ai. Beginning with 15 high-quality, subject-matter-expert (SME)-authored question-and-answer (Q&A) pairs and the complete statutory text of Section 1033, we leveraged the InstructLab pipeline to generate a grounded synthetic training dataset. This dataset was then used to fine-tune the model for accurate, context-aware compliance responses.

The aligned model delivered substantial accuracy gains over the baseline Granite-3-8B, providing plain-language, regulation-grounded answers tailored for compliance officers, legal teams, and customer-facing staff. This work demonstrates how targeted alignment can rapidly adapt a foundation model to a highly specialized regulatory domain. It also enhances retrieval-augmented generation (RAG) performance by sharpening query interpretation and producing concise, structured prompts.

The approach is repeatable, resource-efficient, and scalable to other high-stakes, domain-specific use cases.

# Introduction

Financial regulations, such as the Dodd-Frank Act, present significant interpretation challenges that AI can help address. Section 1033 of the Act establishes rules around consumer financial data access, disclosure, and transparency, which are critical to U.S. banking compliance.

For many institutions, translating these legal requirements into operational guidance is both resource-intensive and error-prone. The rule outlines requirements for data sharing, consent management, API standards, and security protocols. As enforcement deadlines approach, financial institutions are preparing to comply, yet the regulation's complexity continues to challenge stakeholders in both understanding and implementation.

Using IBM's InstructLab methodology, LTIMindtree and IBM customized the Granite-3-8B model to support compliance teams with accurate, plain-language responses grounded in statutory text.

The project had two primary goals:

Demonstrate the value of generative AI in turning dense regulatory text into precise, compliance-ready insights.

Show how a base model can be rapidly adapted and deployed in the watsonx.ai ecosystem for specialized, domain-specific tasks.

This approach enables enterprises to train AI on niche domains such as banking compliance, saving time, improving understanding, and boosting productivity across legal, risk, and technical functions.

# Model Training Process (Granite + InstructLab)

We began with the Granite-3-8B base model and applied **InstructLab**—IBM’s LAB methodology for alignment tuning through the watsonx.ai platform. InstructLab provides an end-to-end pipeline that includes:

## 01

### Taxonomy-driven data curation

Creating a dedicated Section 1033 node under Banking Regulations and populating it with high-quality, grounded Q&A seed examples.

## 02

### Large-scale synthetic data generation (SDG)

Expanding seed examples into thousands of grounded pairs using a teacher model, with automated filtering for factual accuracy.

## 03

### Multi-phase instruction tuning

Sequentially training for domain knowledge and reasoning skills, with replay to prevent forgetting.

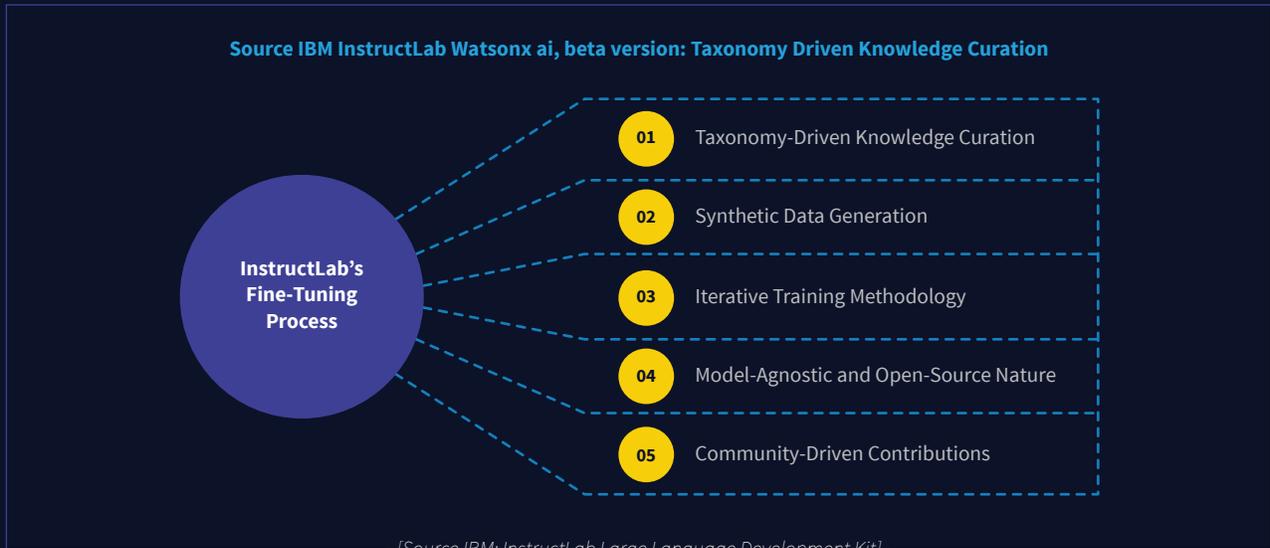
## 04

### Testing and Validation

After training model will be subjected to the domain specific as well as standard benchmarks.

### RAG compatibility

Alignment tuning sharpened the model’s domain understanding, enabling more precise query intent and shorter prompts in downstream RAG workflows. This reduced retrieval noise and token costs.



## 1. Taxonomy-Driven Knowledge Curation

At the core of the InstructLab methodology is the taxonomy—a hierarchical structure that organizes knowledge and skills into clearly defined branches and tasks.

For this project, we extended the existing Knowledge Taxonomy branch by adding a dedicated Section 1033 – Dodd–Frank Act leaf node (Figure 2: Dodd\_Act\_v2).

Rather than taking an unstructured “blender” approach of dumping raw Section 1033 text, we anchored the content in a clear hierarchy aligned with the regulation’s scope and the practical questions a bank might ask.

Working closely with subject-matter experts (SMEs), we:

Selected the full statutory text of Section 1033 as the primary knowledge source.

Authored 15 high-quality instruction–response seed pairs covering key areas, such as:

- Consumer rights (e.g., timelines for providing account data)
- Data format requirements
- Exceptions and limitations

Example questions included:

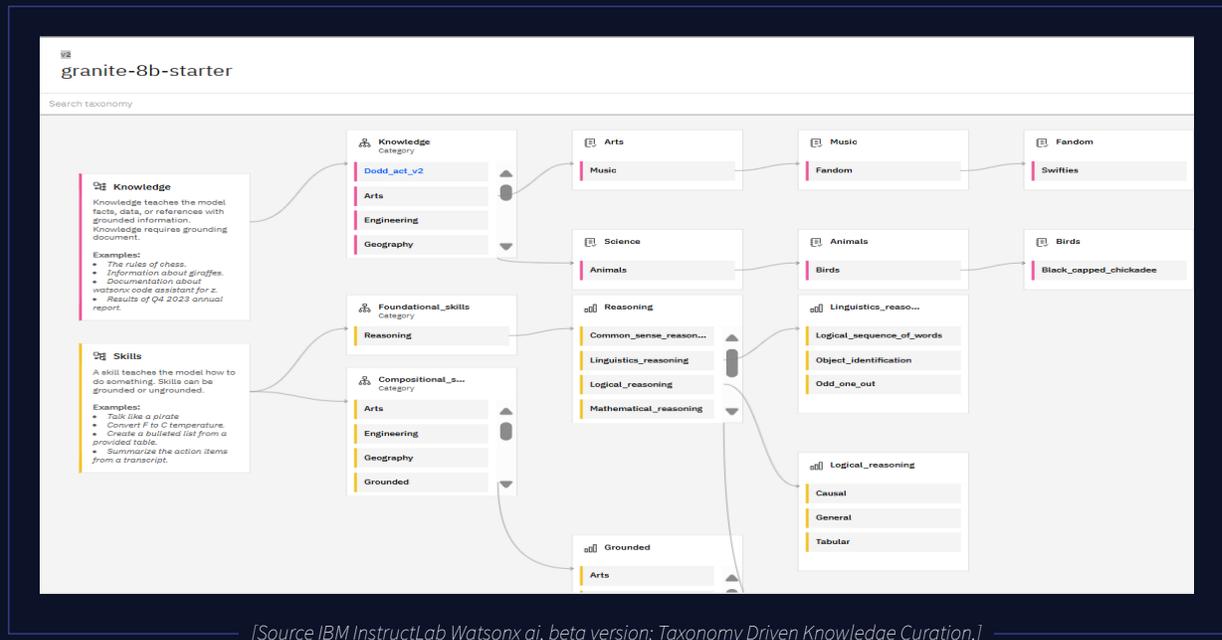
- “What data must be shared, and in what format?”
- “What are the consumer consent requirements under Section 1033?”
- “How does the CFPB interpret its authority under CFPB Section 1033?”
- “What’s the compliance timeline for large versus small banks?”
- “A depository institution holds \$2 billion in total assets. What is its compliance date under § 1033?”

By embedding these seed examples inside the taxonomy, we ensured:

- **Structure:** Clear, navigable knowledge hierarchy
- **Coverage:** Comprehensive representation of Section 1033 provisions
- **Intent:** Targeted application to real compliance scenarios

Where reasoning was required, intermediate steps were shown to reinforce transparency and explainability.

This structured, SME-validated approach laid the foundation for the next LAB workflow phase: large-scale synthetic data generation. As a result, our fine-tuned model was primed to deliver responses with precision, context, and regulatory alignment.



## 2. Large Scale Synthetic Training Data Generation

With the taxonomy and seed data in place, InstructLab moved to synthetic data generation—a scalable, automated process powered by a teacher model.

This process does not merely replicate the teacher model’s knowledge. Instead, it generates diverse, structured training data based on carefully designed prompts.

The SME-crafted examples served as the foundation for a large-scale SDG process using the mixtral-8x7B teacher model. This model interpreted Section 1033 and produced 6,614 additional instruction–response pairs in the same style and tone as the SME-authored seeds, eliminating the need for labor-intensive manual annotation.

The synthetic output underwent strict filtering to remove low-quality or irrelevant samples. The final dataset was:

### Extensive:

Scaled from a handful of SME-crafted seeds to thousands of high-fidelity examples.

### Consistent:

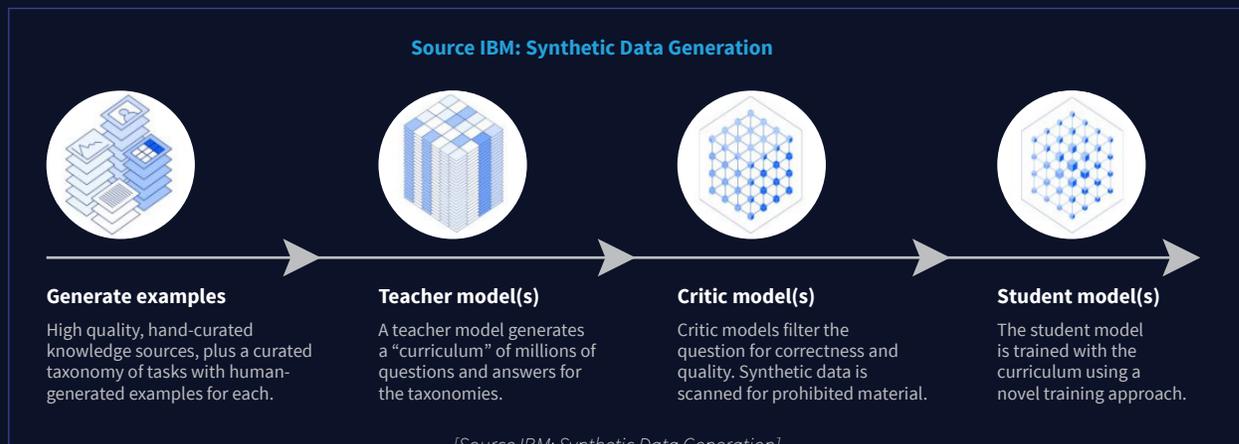
Maintained the same reasoning style, terminology, and compliance framing as the human-authored seeds.

### Efficient:

Required far less manual effort than traditional instruction tuning.

For example, when prompted about acceptable formats for data sharing, the generated answer correctly stated that financial institutions should use secure, standardized formats such as APIs with JSON or XML, accompanied by encryption, authentication, and compliance with the **GLBA safeguards rule**.

Using **watsonx.ai’s InstructLab Tuning Studio**, synthetic data could be reviewed, validated, and regenerated as needed. The **volume of generated data** depended on the length and complexity of the knowledge source (in this case, Section 1033 of the Dodd–Frank Act).



### 3. Multi-Phase Instruction Tuning

We fine-tuned the Granite-3-8B model using InstructLab’s two-phase instruction tuning process, designed to improve model performance incrementally while preserving training stability.



#### Knowledge tuning

In the first phase, the model internalized factual content from Section 1033. Training began with short, targeted answers to build a foundational understanding, followed by longer, context-rich responses.

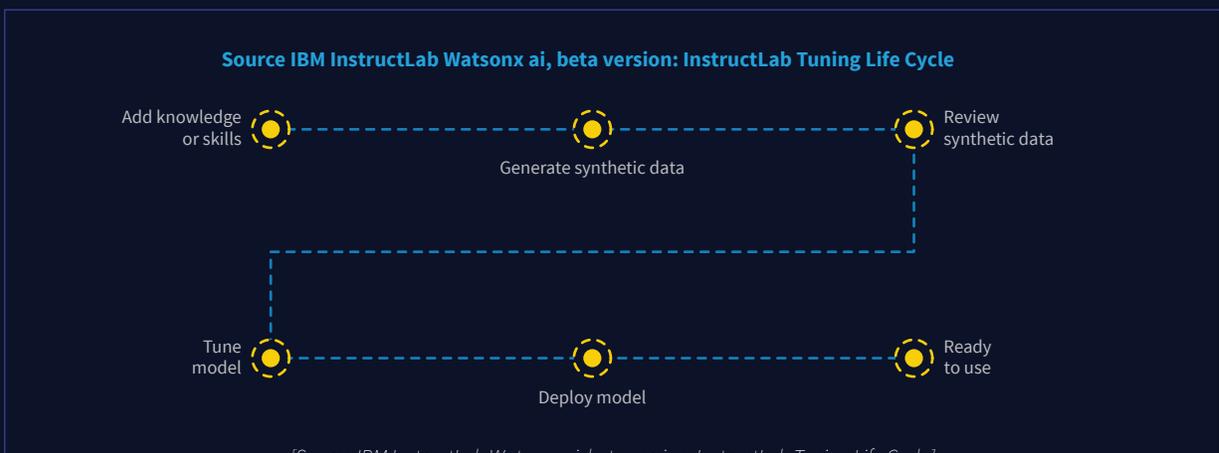


#### Skills tuning

The second stage enhanced the model’s ability to reason, synthesize, and apply knowledge to complex tasks. This stage focused on reasoning, problem-solving, synthesis, and decision-making, mirroring real-world compliance demands.

Training occurred over 16 total epochs. Throughout the process, research benchmarks were used to identify optimal checkpoints, and a replay buffer prevented catastrophic forgetting, ensuring previously learned information was retained.

Despite starting with a relatively small seed dataset, the process was effective. Granite successfully internalized Section 1033 content and produced accurate, regulation-aligned responses.



## 4. Testing and Validation

After training, the fine-tuned model was deployed in **watsonx.ai** using the “Bring Your Own Model” (BYOM) feature. We assessed the model using both domain-specific tests and standard benchmarks.



### Domain-specific test

The model was evaluated on prepared questions and slight variations to gauge its robustness. Interactive testing confirmed that it could answer complex compliance questions directly from Section 1033 in plain language, with citations to the statutory text.

#### Results

- The results were very positive. For our set of approximately ten core questions, the model’s answers were accurate and specific, often mirroring language from the regulation. We compared outputs with those from the original base model.
- Example: On the question “What does Section 1033 require banks to do before sharing customer data with a third party?”, the base (untrained) model gave a vague response about privacy in general. Whereas the Trained Granite Model answered with specifics: Banks must provide the customer with a clear authorization disclosure that identifies the third party and the data to be shared and obtain the customer’s explicit informed consent before any data sharing occurs.

This comparison illustrates the trained model’s ability to deliver precise, actionable answers.



### Standard benchmarks

- **Massive multitask language understanding (MMLU) benchmark**

The MMLU benchmark evaluates both a model’s breadth of knowledge and its ability to apply reasoning across diverse domains. It covers 57 subjects across STEM, humanities, social sciences, and other disciplines. Each test consists of multiple-choice questions with four answer options, assessing factual recall and problem-solving skills.

Scores are based on the percentage of correct answers, with the average across all subjects representing the final benchmark score. Since answers are scored against a predefined key, no judge model is required.

### In this work

- We evaluated MMLU performance of the tuned model against the base model to measure the effect of training on additional domain-specific knowledge.
- Multiple-choice questions were generated during the SDG phase from the Section 1033 corpus by a teacher model to measure factual recall and comprehension.

### Purpose of MMLU

Assess factual understanding gained during knowledge tuning.

### Results

- **Knowledge\_dodd-act-phase-2:** Improved from **0.58 → 0.72** (+0.15)
- **Final MMLU Score: 0.56 / 1.0**



### Multi-turn benchmark (MT-Bench)

The MT-bench benchmark evaluates a model's ability to handle **multi-turn conversations** and execute skills defined in InstructLab's taxonomy. It measures how well the model manages **specialized, domain-specific tasks** requiring contextual understanding across turns.

Each evaluation consists of **80 multi-turn questions**, where the model's first answer is followed by a **related follow-up question** to assess continuity, reasoning, and adaptability:

<Question 1> → <Model Answer 1> → <Follow-up Question> → <Model Answer 2>

Responses are scored by a **strong LLM judge** (prometheus-8x7b-v2-0). The judge rates each answer on a **scale of 1 to 10** based on correctness, coherence, and reasoning quality.

### In this work

- MT-Bench was used to evaluate the tuned model's skill performance on custom Section 1033 compliance tasks.
- Open-ended prompts reflected domain skills curated via InstructLab's taxonomy.

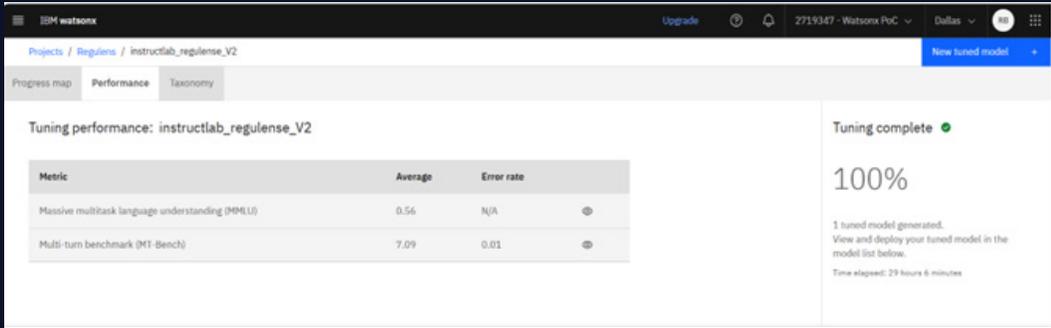
**Purpose of MMLU**

Measure how well the tuned model applies knowledge in realistic, multi-turn interactions requiring context retention and reasoning.

**Results**

- **MT-Bench Score: 7.9/10**

Source IBM InstructLab Watsonx ai, beta version: Final MMLU, MT Bench Score



Metric	Average	Error rate
Massive multitask language understanding (MMLU)	0.56	N/A
Multi-turn benchmark (MT-Bench)	7.09	0.01

Tuning complete 100%

1 tuned model generated. View and deploy your tuned model in the model list below.

Time elapsed: 29 hours 6 minutes

[Source IBM InstructLab Watsonx ai, beta version: Final MMLU, MT Bench Score]

# Key Findings and Insights

Fine-tuning Granite-3-8b on Section 1033 yielded several important outcomes and lessons:

**01**

## Accuracy improvement

The fine-tuned model correctly answered 80–90% of Section 1033 queries, a significant improvement from the base model’s approximate 20% accuracy. Even minimal fine tuning enabled precise identification of regulatory obligations.

**02**

## Limitations and trust

Minor omissions and loosely related concepts appeared occasionally, likely from residual pretraining data. We mitigated this by cross-checking answers against Section 1033 and plan further fine-tuning or retrieval augmentation. Strong early accuracy built internal stakeholder trust, but production deployment will require transparent citations of regulation text in every answer.

**03**

## Consistency and focus

The model stayed focused on Section 1033, using accurate terminology and avoiding irrelevant content. This highlights the value of domain-specific training in compliance-critical environments.

**04**

## Depth of understanding

The model demonstrated the ability to synthesize and structure information beyond repeating legal text. Responses were logically organized, often in step-by-step formats.

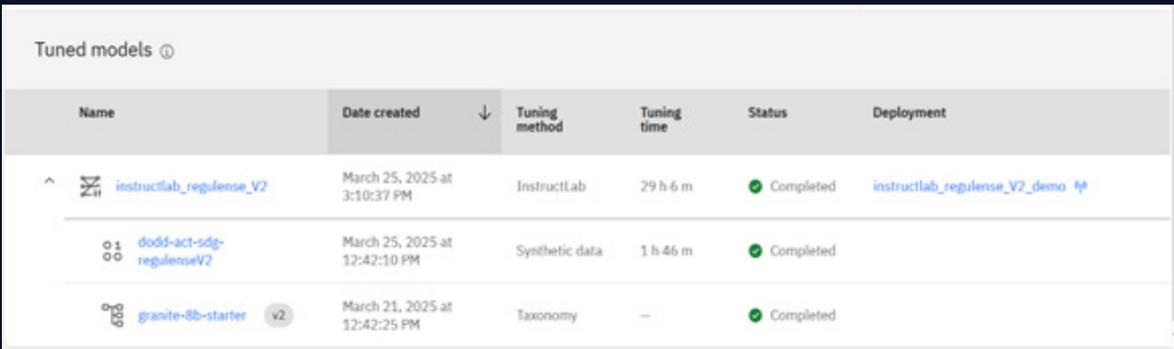
For example, when asked about consumer rights under Section 1033, the model listed access, consent, revocation, and breach notification clearly. This indicates a developed internal representation of the regulation, enabling comprehensive, well-structured answers that reflect multiple facets of the rule.

**Summary:** Fine-tuning the Granite LLM on a targeted domain (Section 1033) resulted in a highly accurate, domain-specific conversational agent.

# Technical Details and Performance Metrics

## Performance metrics

- Accuracy: 80% on in-domain test queries vs. ~20% for the base model
- Training time: ~30 hours for alignment tuning + ~2 hours for synthetic data generation
- Model size: ~16 GB size of the model (Granite 3–8B model, FP16)



Name	Date created	Tuning method	Tuning time	Status	Deployment
instructlab_regulense_V2	March 25, 2025 at 3:10:37 PM	InstructLab	29 h 6 m	Completed	<a href="#">instructlab_regulense_V2_demo</a>
dodd-act-sdg-regulenseV2	March 25, 2025 at 12:42:10 PM	Synthetic data	1 h 46 m	Completed	
granite-8b-starter v2	March 21, 2025 at 12:42:25 PM	Taxonomy	—	Completed	

[Source IBM InstructLab Watsonx ai, beta version: Final MMLU, MT Bench Score]

- Inference speed: On CPU (8 cores), each inference (answer generation) took ~20+ seconds for a ~150-word answer. On an NVIDIA V100 GPU (which we tested in a cloud environment), this dropped to around five seconds for a similar answer. So, deployment on GPU can achieve near real-time response.
- Memory footprint: We used an **8-bit parameter model** with **FP16 (float16)** precision for computation. The base model weights required approximately **15–16 GB** to load in FP16. However, **total GPU memory usage** also depends on the **KV cache** and **context length** during inference. For a **4k–8k token context**, the end-to-end GPU requirement is typically **~28–30 GB** in FP16. This is an important consideration for IT teams planning deployment, as it determines the GPU resources needed. For example, an **NVIDIA A100 (40 GB)** or an equivalent GPU with sufficient VRAM can comfortably host and run the model for these context sizes.

- **Model maintenance:** An internal model card was developed to support responsible deployment.

Field	Details
<b>Intended Use</b>	Assist with answering questions specifically related to Dodd-Frank Act Section 1033. Optimized for regulatory interpretation and compliance workflows
<b>Not Intended For</b>	General-purpose Q&A outside Section 1033. While it can respond to other topics, these outputs have not been validated for accuracy.
<b>Training Data</b>	Original Section 1033 regulatory text, plus curated and synthetic Q&A pairs designed for coverage of obligations, rights, consent rules, and compliance definitions.
<b>Performance</b>	Achieved high accuracy (>80% correct) on in-domain queries; no formal evaluation on out-of-domain topics.
<b>Ethical Considerations</b>	Content reflects the formal, factual language of regulation. Minimal bias risk given domain specificity. No inappropriate outputs observed during testing; however, safeguards should prevent prompts unrelated to the regulation
<b>Benchmarking</b>	Evaluated against Industry-standard benchmark MMLU, MT along with manual task-specific evaluation.

# Case Studies:

## Applying the Trained Granite Model in Banking Compliance

To illustrate practical value, consider this scenario from a mid-sized U.S. bank preparing for Section 1033 compliance.

### Scenario

A project manager is leading a cross-functional initiative involving legal experts, IT architects, and business analysts. The team integrates the Trained Granite Model into planning meetings to clarify regulatory requirements in real time and prevent oversight.

#### Example interaction 1

The team is discussing what data needs to be shared.

- **Prompt (project manager)**

“AI Assistant, what data does Section 1033 require us to share with customers or third parties?”

- **Trained Granite Model response**

“Under Section 1033, the bank must make available ‘covered data’ which includes any information that the bank has about a consumer’s use of a financial product or service obtained from the bank. This typically includes transaction data, account balances, payment history, and any other information the bank holds that is related to the consumer’s account or transactions. The data must be provided in an electronic, machine-readable format that can be used by the consumer or a third-party service the consumer authorizes.

- In summary, banks must share account information, transaction details, and related customer data that they possess.

## Impact

The team prioritizes identifying and cataloguing transaction and account data in their internal systems. They also flag the “machine-readable format” requirement, prompting IT to plan API-based delivery mechanisms.

### Example interaction 1

The legal expert wants to confirm consent requirements.

- **Legal SME**

“AI Assistant, what does Section 1033 say about customer consent for data sharing?”

- **Trained Granite Model response**

“Section 1033 requires that a bank obtains a consumer’s explicit consent before sharing their data with a third party. The bank must present the consumer with a clear and conspicuous authorization disclosure, which outlines what data will be shared, with whom, and for what purpose. Only after the consumer agrees (for example, by signing an authorization or checking an approval box) can the data be shared. Additionally, the consumer should have the ability to revoke consent at any time, and the bank must stop sharing data if consent is withdrawn.”

**Key takeaway:** By embedding a fine-tuned regulatory model into daily compliance activities accelerates decision-making, reduces misinterpretation risks, and streamlines the path from regulation to operational implementation.

# Conclusion and Future Directions

Fine-tuning IBM's Granite 3-8B model for Section 1033 compliance demonstrates the value of adapting LLMs to high-stakes, domain-specific tasks.

By grounding the model in regulation-specific content and curated Q&A, it evolved into a precision legal assistant—reducing manual research, accelerating interpretation, and enabling faster, more confident compliance decisions.

This project was delivered in weeks, with modest resources, and produced measurable benefits such as:

- Significant time savings for compliance and legal teams
- Improved accessibility** to dense statutory language
- Scalable AI support** extendable across multiple compliance domains

Verification and governance remain non-negotiable in regulated industries, ensuring accuracy, trust, and defensibility of outputs.

Looking ahead, we will:

- Continuously retrain the model to reflect legal changes and user feedback.
- Integrate with IBM watsonx tools for seamless enterprise deployment.
- Expand coverage to other financial regulations and jurisdictions.
- Explore optimizations such as model compression for efficient scaling.

The methodology, including taxonomy creation, seed example design, and synthetic data generation will be published as a reusable framework for other teams to adapt to their own regulatory or domain-specific needs.

**Bottom line:** Regulatory interpretation no longer needs to rely solely on manual effort. With targeted fine-tuning, AI delivers scalable, trusted, and operationally relevant expertise transforming compliance from a reactive obligation into a strategic advantage.

**Note:** The above project was implemented using a beta version of InstructLab on IBM watsonx.ai

# References

---

What is InstructLab and why do developers need it?, Syeda Aameena Begum, IBM Developer, August 3, 2025:

<https://developer.ibm.com/articles/awb-instructlab-why-developers-need-it>

A new way to collaboratively customize LLMs., IBM, IBM Research, June 7, 2024:

<https://research.ibm.com/blog/instruct-lab>

From data to decisions: Aligning models for enterprise use cases with InstructLab in watsonx.ai. , Suhas Kashyap, Syeda Aameena Begum, IBM News, October 21, 2024:

<https://www.ibm.com/new/announcements/from-data-to-decisions-aligning-models-for-enterprise-use-cases-with-instructlab-in-watsonx-ai>

LAB: Large-Scale Alignment for ChatBots, Shivchander Sudalairaj, Abhishek Bhandwaladar, Aldo Pareja, Kai Xu, David D. Cox, Akash Srivastava, MIT-IBM Watson AI Lab and IBM Research, March 2, 2024:

<https://arxiv.org/abs/2403.01081>

MMLU Overview (measuring massive multitask language understanding), Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt, arxiv, January 12, 2021:

<https://arxiv.org/pdf/2009.03300v3>

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica., arxiv, December 24, 2023:

<https://arxiv.org/abs/2306.05685v4>

Enterprise Benchmarks for Large Language Model , Bing Zhang, Mikio Takeuchi , Ryo Kawahara, Shubhi Asthana, Maruf Hossain, Guang-Jie Ren , Kate Soule, Yada Zhu , IBM, 2025:

<https://research.ibm.com/publications/enterprise-benchmarks-for-large-language-model-evaluation--1>

## Authors' bio



### **Rahul Bhave**

*Senior Specialist- Data & Analytics, LTIMindtree*

Rahul Bhave brings over 15 years of experience in Software Automation Testing, Data and Analytics, and AI Engineering. In his current role at LTIMindtree, Rahul is part of the IBM watsonx Center of Excellence, where he is actively engaged in various watsonx product suites, including IBM watsonx.ai, IBM watsonx.data, and IBM InstructLab. His work focuses on leveraging cutting-edge AI technologies to drive innovation and deliver impactful business solutions.

### **Syeda Aameena Begum**

*AI Engineer, Ecosystem Engineering Lab, IBM*

Syeda is an AI engineer at Ecosystem Engineering SI Labs IBM, specializing in Generative AI and Neural Networks. She brings extensive hands-on experience in Generative AI, watsonx.ai tech stack and InstructLab, focusing on end-to-end AI development and scalable deployment. Her core expertise lies in building intelligent systems that blend research-grade innovation with production-ready engineering, driving advanced solutions that transform industries.



**LTIMindtree** is a global technology consulting and digital solutions company that enables enterprises across industries to reimagine business models, accelerate innovation, and maximize growth by harnessing digital technologies. As a digital transformation partner to more than 700 clients, LTIMindtree brings extensive domain and technology expertise to help drive superior competitive differentiation, customer experiences, and business outcomes in a converging world. Powered by 83,000+ talented and entrepreneurial professionals across more than 40 countries, LTIMindtree — a Larsen & Toubro Group company — solves the most complex business challenges and delivers transformation at scale. For more information, please visit <https://www.ltimindtree.com/>