**Granite-3.3-8B-Instruct**

**Model Summary:** Granite-3.3-8B-Instruct is a 8-billion parameter 128K context length language model fine-tuned for improved reasoning and instruction-following capabilities. Built on top of Granite-3.3-8B-Base, the model delivers significant gains on benchmarks for measuring generic performance including AlpacaEval-2.0 and Arena-Hard, and improvements in mathematics, coding, and instruction following. It also supports Fill-in-the-Middle (FIM) for code completion tasks and structured reasoning through <think></think> and <response></response> tags, providing clear separation between internal thoughts and final outputs. The model has been trained on a carefully balanced combination of permissively licensed data and curated synthetic tasks.

* **Developers:** Granite Team, IBM
* **Website**: [Granite Docs](https://www.ibm.com/granite/docs/)
* **Release Date**: April 16th, 2025
* **License:** [Apache 2.0](https://www.apache.org/licenses/LICENSE-2.0)

**Supported Languages:** English, German, Spanish, French, Japanese, Portuguese, Arabic, Czech, Italian, Korean, Dutch, and Chinese. However, users may finetune this Granite model for languages beyond these 12 languages.

**Intended Use:** This model is designed to handle general instruction-following tasks and can be integrated into AI assistants across various domains, including business applications.

**Capabilities**

* **Thinking**
* Summarization
* Text classification
* Text extraction
* Question-answering
* Retrieval Augmented Generation (RAG)
* Code related tasks
* Function-calling tasks
* Multilingual dialog use cases
* **Fill-in-the-middle**
* Long-context tasks including long document/meeting summarization, long document QA, etc.

**Generation:** This is a simple example of how to use Granite-3.3-8B-Instruct model.

Install the following libraries:

pip install torch torchvision torchaudio

pip install accelerate

pip install transformers

Then, copy the snippet from the section that is relevant for your use case.

from transformers import AutoModelForCausalLM, AutoTokenizer, set\_seed

import torch

model\_path="ibm-granite/granite-3.3-8b-instruct"

device="cuda"

model = AutoModelForCausalLM.from\_pretrained(

model\_path,

device\_map=device,

torch\_dtype=torch.bfloat16,

)

tokenizer = AutoTokenizer.from\_pretrained(

model\_path

)

conv = [{"role": "user", "content":"Redesign a common household item to make it more sustainable and user-friendly. Explain the changes and their benefits."}]

input\_ids = tokenizer.apply\_chat\_template(conv, return\_tensors="pt", thinking=True, return\_dict=True, add\_generation\_prompt=True).to(device)

set\_seed(42)

output = model.generate(

\*\*input\_ids,

max\_new\_tokens=8192,

)

prediction = tokenizer.decode(output[0, input\_ids["input\_ids"].shape[1]:], skip\_special\_tokens=True)

print(prediction)

**Example Outputs**

* thinking=True

<think>To address this prompt, I'll focus on redesigning a common household item: the plastic water bottle. The goal is to make it more sustainable and user-friendly. Here are the proposed changes and their benefits:

1. \*\*Material\*\*: Replace plastic with biodegradable or recyclable materials like glass or plant-based plastics (PLA).

2. \*\*Design\*\*: Introduce a modular design for easy refilling and cleaning.

3. \*\*Functionality\*\*: Incorporate a built-in water filter for purification.

4. \*\*Durability\*\*: Enhance the bottle's lifespan through robust construction.

5. \*\*User-friendliness\*\*: Simplify the cap mechanism for easy opening and closing.

These changes would significantly reduce plastic waste, promote recycling, ensure access to clean drinking water, and enhance the overall user experience.

Now, I will detail each change and its benefits in a structured format.</think>

<response>### Redesign of the Plastic Water Bottle for Sustainability and User-Friendliness

#### 1. Material: Biodegradable or Recyclable Alternatives

\*\*Change\*\*: Transition from traditional PET plastic to glass or plant-based plastics (Polylactic Acid - PLA).

\*\*Benefits\*\*:

- \*\*Sustainability\*\*: Glass and PLA are biodegradable or highly recyclable, drastically reducing plastic pollution.

- \*\*Health\*\*: Unlike PET, these materials do not leach chemicals into the water, ensuring safer consumption.

#### 2. Modular Design for Easy Refilling and Cleaning

\*\*Change\*\*: Implement a detachable, interchangeable inner liner within a reusable outer casing.

\*\*Benefits\*\*:

- \*\*User-Friendliness\*\*: Users can easily remove the inner liner for thorough cleaning, reducing the risk of bacterial growth.

- \*\*Sustainability\*\*: The outer casing remains intact, encouraging reuse, while the inner liner can be replaced or recycled when worn out.

#### 3. Built-in Water Filter

\*\*Change\*\*: Integrate a removable, replaceable water filter into the bottle cap.

\*\*Benefits\*\*:

- \*\*Accessibility\*\*: Provides clean drinking water directly from any source (tap, river, etc.), eliminating the need for additional filtration devices.

- \*\*Health\*\*: Removes contaminants, bacteria, and impurities, ensuring safe hydration.

- \*\*Sustainability\*\*: Reduces the use of single-use plastic water bottles by offering a purification solution within a reusable container.

#### 4. Enhanced Durability

\*\*Change\*\*: Use high-quality, durable materials and robust construction techniques for both the outer casing and inner liner.

\*\*Benefits\*\*:

- \*\*Longevity\*\*: A more robust design increases the bottle's lifespan, reducing the frequency of replacements.

- \*\*Cost-effectiveness\*\*: Although the initial cost might be higher, the long-term savings from reduced replacements make it economical.

- \*\*Sustainability\*\*: Less frequent disposal means lower environmental impact.

#### 5. Simplified Cap Mechanism

\*\*Change\*\*: Adopt a one-touch, easy-to-open and close cap design, possibly inspired by sports bottle caps.

\*\*Benefits\*\*:

- \*\*User-Friendliness\*\*: Especially beneficial for children, elderly, or those with limited hand dexterity.

- \*\*Convenience\*\*: Quick access to water encourages more frequent hydration, promoting healthier habits.

### Conclusion

This redesigned water bottle addresses key sustainability concerns by minimizing plastic use and waste, while also enhancing user-friendliness through improved functionality, durability, and accessibility. By integrating biodegradable materials, a modular design, a built-in filter, robust construction, and a simple cap mechanism, the bottle not only reduces environmental footprint but also offers a superior user experience.</response>

* thinking=False

Item for Redesign: Plastic Bottle (e.g., water or beverage bottle)

Redesign Concept: Eco-Friendly, Modular, and Multi-Functional Reusable Bottle

Changes:

1. Material: Replace the conventional plastic with food-grade, BPA-free, and recyclable materials such as stainless steel or glass. These materials are durable, non-toxic, and have a longer lifespan, reducing waste.

2. Modular Design: Introduce a modular design that allows users to customize their bottle according to their needs. This could include interchangeable caps, mouthpiece sizes, and carrying loops. Users can purchase additional modules as needed, promoting a more sustainable approach by reducing the need to buy an entirely new bottle for different purposes.

3. Integrated Filter: Incorporate a built-in, washable, and reusable filter that can remove impurities and improve the taste of water. This eliminates the need for single-use disposable filters or bottled water, further reducing plastic waste.

4. Smart Cap: Develop a smart cap with a built-in digital display and temperature sensor. This feature allows users to track their daily water intake, set hydration goals, and monitor the temperature of their beverage. The smart cap can be synced with a mobile app for additional functionality, such as reminders and progress tracking.

5. Easy-to-Clean Design: Ensure the bottle has a wide mouth and smooth interior surfaces for easy cleaning. Include a brush for hard-to-reach areas, making maintenance simple and encouraging regular use.

6. Collapsible Structure: Implement a collapsible design that reduces the bottle's volume when not in use, making it more portable and convenient for storage.

Benefits:

1. Sustainability: By using recyclable materials and reducing plastic waste, this redesigned bottle significantly contributes to a more sustainable lifestyle. The modular design and reusable filter also minimize single-use plastic consumption.

2. User-Friendly: The smart cap, easy-to-clean design, and collapsible structure make the bottle convenient and user-friendly. Users can customize their bottle to suit their needs, ensuring a better overall experience.

3. Healthier Option: Using food-grade, BPA-free materials and an integrated filter ensures that the beverages consumed are free from harmful chemicals and impurities, promoting a healthier lifestyle.

4. Cost-Effective: Although the initial investment might be higher, the long-term savings from reduced purchases of single-use plastic bottles and disposable filters make this reusable bottle a cost-effective choice.

5. Encourages Hydration: The smart cap's features, such as hydration tracking and temperature monitoring, can motivate users to stay hydrated and develop healthier habits.

By redesigning a common household item like the plastic bottle, we can create a more sustainable, user-friendly, and health-conscious alternative that benefits both individuals and the environment.

**Evaluation Results:**

| **Models** | **Arena-Hard** | **AlpacaEval-2.0** | **MMLU** | **PopQA** | **TruthfulQA** | **GSM8K** | **HumanEval** | **HumanEval+** | **IFEval** | **AttaQ** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Granite-3.1-2B-Instruct | 23.3 | 27.17 | 57.11 | 20.55 | 59.79 | 67.55 | 79.45 | 75.26 | 63.59 | 84.7 |
| Granite-3.2-2B-Instruct | 24.86 | 34.51 | 57.18 | 20.56 | 59.8 | 67.02 | 80.13 | 73.39 | 61.55 | 83.23 |
| **Granite-3.3-2B-Instruct** | 28.86 | 43.45 | 55.88 | 18.4 | 58.97 | 72.48 | 80.51 | 75.68 | 65.8 | 87.47 |
| Llama-3.1-8B-Instruct | 36.43 | 27.22 | 69.15 | 28.79 | 52.79 | 83.24 | 85.32 | 80.15 | 79.10 | 83.43 |
| DeepSeek-R1-Distill-Llama-8B | 17.17 | 21.85 | 45.80 | 13.25 | 47.43 | 72.18 | 67.54 | 62.91 | 66.50 | 42.87 |
| Qwen-2.5-7B-Instruct | 25.44 | 30.34 | 74.30 | 18.12 | 63.06 | 84.46 | 93.35 | 89.91 | 74.90 | 81.90 |
| DeepSeek-R1-Distill-Qwen-7B | 10.36 | 15.35 | 50.72 | 9.94 | 47.14 | 78.47 | 79.89 | 78.43 | 59.10 | 42.45 |
| Granite-3.1-8B-Instruct | 37.58 | 30.34 | 66.77 | 28.7 | 65.84 | 79.15 | 89.63 | 85.79 | 73.20 | 85.73 |
| Granite-3.2-8B-Instruct | 55.25 | 61.19 | 66.79 | 28.04 | 66.92 | 81.65 | 89.35 | 85.72 | 74.31 | 84.7 |
| **Granite-3.3-8B-Instruct** | 57.56 | 62.68 | 65.54 | 26.17 | 66.86 | 80.89 | 89.73 | 86.09 | 74.82 | 88.5 |
| **Comparison with different models over various benchmarks**[**1**](https://dataplatform.cloud.ibm.com/wx/agents?context=wx&project_id=af537d36-78af-4d04-bf18-7bc42221c5b9#fn1)**. Scores of AlpacaEval-2.0 and Arena-Hard are calculated with thinking=True** | | | | | | | | | | |

| **Models** | **AIME24** | **MATH-500** |
| --- | --- | --- |
| Granite-3.1-2B-Instruct | 0.89 | 35.07 |
| Granite-3.2-2B-Instruct | 0.89 | 35.54 |
| **Granite-3.3-2B-Instruct** | 3.28 | 58.09 |
| Granite-3.1-8B-Instruct | 1.97 | 48.73 |
| Granite-3.2-8B-Instruct | 2.43 | 52.8 |
| **Granite-3.3-8B-Instruct** | 8.12 | 69.02 |
| **Math Benchmarks** | | |

**Training Data:** Overall, our training data is largely comprised of two key sources: (1) publicly available datasets with permissive license, (2) internal synthetically generated data targeted to enhance reasoning capabilites.

**Infrastructure:** We train Granite-3.3-8B-Instruct using IBM's super computing cluster, Blue Vela, which is outfitted with NVIDIA H100 GPUs. This cluster provides a scalable and efficient infrastructure for training our models over thousands of GPUs.

**Ethical Considerations and Limitations:** Granite-3.3-8B-Instruct builds upon Granite-3.3-8B-Base, leveraging both permissively licensed open-source and select proprietary data for enhanced performance. Since it inherits its foundation from the previous model, all ethical considerations and limitations applicable to [Granite-3.3-8B-Base](https://huggingface.co/ibm-granite/granite-3.3-8b-base) remain relevant.

**Resources**

* â­ï¸ Learn about the latest updates with Granite: <https://www.ibm.com/granite>
* ð Get started with tutorials, best practices, and prompt engineering advice: <https://www.ibm.com/granite/docs/>
* ð¡ Learn about the latest Granite learning resources: <https://ibm.biz/granite-learning-resources>

[[1]](https://dataplatform.cloud.ibm.com/wx/agents?context=wx&project_id=af537d36-78af-4d04-bf18-7bc42221c5b9#fnref1) Evaluated using [OLMES](https://github.com/allenai/olmes) (except AttaQ and Arena-Hard scores)