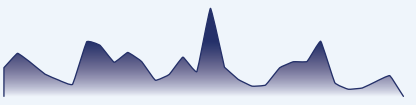
microsoft / Phi-3-mini-128k-instruct



Downloads last month: 261,529

Model Description

The Phi-3-Mini-128K-Instruct is a 3.8 billion-parameter, lightweight, state-of-the-art open model trained using the Phi-3 datasets. This dataset includes both synthetic data and filtered publicly available website data, with an emphasis on high-quality and reasoning-dense properties. The model belongs to the Phi-3 family with the Mini version in two variants 4K and 128K which is the context length (in tokens) that it can support.

After initial training, the model underwent a post-training process that involved supervised fine-tuning and direct preference optimization to enhance its ability to follow instructions and adhere to safety measures. When evaluated against benchmarks that test common sense, language understanding, mathematics, coding, long-term context, and logical reasoning, the Phi-3 Mini-128K-Instruct demonstrated robust and state-of-the-art performance among models with fewer than 13 billion parameters.

Datasets

Our training data includes a wide variety of sources, totaling 4.9 trillion tokens, and combines:
- Publicly available documents filtered rigorously for quality, selected high-quality educational data, and code;
- Newly created synthetic, “textbook-like” data for the purpose of teaching math, coding, common sense reasoning, general knowledge of the world (science, daily activities, theory of mind, etc.);
- High quality chat format supervised data covering various topics to reflect human preferences on different aspects such as instruct-following, truthfulness, honesty and helpfulness.

We are focusing on the quality of data that could potentially improve the reasoning ability for the model, and we filter the publicly available documents to contain the correct level of knowledge. As an example, the result of a game in premier league in a particular day might be good training data for frontier models, but we need to remove such information to leave more model capacity for reasoning for the small size models. More details about data can be found in the Phi-3 Technical Report.

Intended Uses

Primary use cases

The model is intended for commercial and research use in English. The model provides uses for applications which require:
- Memory/compute constrained environments
- Latency bound scenarios
- Strong reasoning (especially code, math and logic)

Our model is designed to accelerate research on language and multimodal models, for use as a building block for generative AI powered features.

Use case considerations

Our models are not specifically designed or evaluated for all downstream purposes. Developers should consider common limitations of language models as they select use cases, and evaluate and mitigate for accuracy, safety, and fariness before using within a specific downstream use case, particularly for high risk scenarios. Developers should be aware of and adhere to applicable laws or regulations (including privacy, trade compliance laws, etc.) that are relevant to their use case.
Nothing contained in this Model Card should be interpreted as or deemed a restriction or modification to the license the model is released under.

Responsible AI Considerations

Like other language models, the Phi series models can potentially behave in ways that are unfair, unreliable, or offensive. Some of the limiting behaviors to be aware of include:
- **Quality of Service:** the Phi models are trained primarily on English text. Languages other than English will experience worse performance. English language varieties with less representation in the training data might experience worse performance than standard American English.
- **Representation of Harms & Perpetuation of Stereotypes:** These models can over- or under-represent groups of people, erase representation of some groups, or reinforce demeaning or negative stereotypes. Despite safety post-training, these limitations may still be present due to differing levels of representation of different groups or prevalence of examples of negative stereotypes in training data that reflect real-world patterns and societal biases.
- **Inappropriate or Offensive Content**: these models may produce other types of inappropriate or offensive content, which may make it inappropriate to deploy for sensitive contexts without additional mitigations that are specific to the use case.
- **Information Reliability:** Language models can generate nonsensical content or fabricate content that might sound reasonable but is inaccurate or outdated.
- **Limited Scope for Code:** Majority of Phi-3 training data is based in Python and use common packages such as "typing, math, random, collections, datetime, itertools". If the model generates Python scripts that utilize other packages or scripts in other languages, we strongly recommend users manually verify all API uses.

Developers should apply responsible AI best practices and are responsible for ensuring that a specific use case complies with relevant laws and regulations (e.g. privacy, trade, etc.). Important areas for consideration include:
- **Allocation:** Models may not be suitable for scenarios that could have consequential impact on legal status or the allocation of resources or life opportunities (ex: housing, employment, credit, etc.) without further assessments and additional debiasing techniques.
- **High-Risk Scenarios:** Developers should assess suitability of using models in high-risk scenarios where unfair, unreliable or offensive outputs might be extremely costly or lead to harm. This includes providing advice in sensitive or expert domains where accuracy and reliability are critical (ex: legal or health advice). Additional safeguards should be implemented at the application level according to the deployment context.
- **Misinformation:** Models may produce inaccurate information. Developers should follow transparency best practices and inform end-users they are interacting with an AI system. At the application level, developers can build feedback mechanisms and pipelines to ground responses in use-case specific, contextual information, a technique known as Retrieval Augmented Generation (RAG).
- **Generation of Harmful Content:** Developers should assess outputs for their context and use available safety classifiers or custom solutions appropriate for their use case.
- **Misuse:** Other forms of misuse such as fraud, spam, or malware production may be possible, and developers should ensure that their applications do not violate applicable laws and regulations.