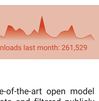
microsoft / Phi-3-mini-128k-instruct



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 fairness 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.

Summary of Model Risks by Example Uses

We identified **14** potential model risks and **17** mitigation strategies (M) for **4** potential model uses

Applicable to the use
Not applicable to the use
Risk resulted in real-world harm*

Risks	Recommending personalized content (U1)	Creating personalized ad campaigns (U2)	Personalizing learning experiences (U3)	Moderating harmful content (U4)
(R1) Produces unfair, unreliable, or offensive outputs in high-risk scenarios harming reputations	■	■	■	■
(R2) Discriminates against marginalized groups by perpetuating harmful biases and stereotypes	■	■	■	■
(R3) Misinterprets or underperforms in response to informal English slang, or non-English languages	■	■	■	■
(R4) Facilitates misinformation by spreading false knowledge	■	■	■	□
(R5) Increases psychological harm by causing fear and anxiety through threatening statements	■	■	■	□
(R6) Produces harmful content if explicitly prompted or instructed	□	■	■	■
(R7) Generates content that mirrors societal biases	□	■	■	■
(R8) Produces inaccurate or outdated content	■	■	□	□
(R9) Facilitates misuse for fraud, spam, or malware production	■	□	■	□
(R10) Violates privacy by exposing internal messages	□	□	□	■
(R11) Violates user trust by bypassing safeguards designed to restrict certain functionalities	□	□	■	□
(R12) Struggles or fails to adhere to intricate or nuanced instructions	□	□	■	□
(R13) Produces problematic outputs when not aligned to human preferences for safety	□	□	■	□
(R14) Uses a mix of Web data and technical sources without transparency	□	□	□	■

Details of the example uses

(U1) Purpose: **Recommending personalized content**
Capability: Analyzing preferences for suggestions
AI User: Streaming platforms
AI Subject: Content consumers
Domain: Recommender Systems and Personalization

(U2) Purpose: **Creating personalized ad campaigns**
Capability: Analyzing user behavior from social media posts
AI User: Marketing agencies
AI Subject: Consumers
Domain: Marketing and Advertising

(U3) Purpose: **Personalizing learning experiences**
Capability: Analyzing student performance and tailoring content
AI User: Educational platforms
AI Subject: Students
Domain: Education and vocational training

(U4) Purpose: **Moderating harmful content**
Capability: Detecting inappropriate posts
AI User: Social media companies
AI Subject: Platform users
Domain: Social Media

Mitigations for Risks

(R1) Produces unfair, unreliable, or offensive outputs in high-risk scenarios harming reputations

Capability risk · Representation and toxicity harms

- (M) Use available safety classifiers or custom solutions
- (M) Assess suitability in high-risk scenarios
- (M) Implement additional mitigations for sensitive contexts
- (M) Perform further assessments and additional debiasing techniques

(R2) Discriminates against marginalized groups by perpetuating harmful biases and stereotypes

Capability risk · Representation and toxicity harms

- (M) Develop ways to reduce model toxicity
- (M) Use available safety classifiers or custom solutions
- (M) Perform further assessments and additional debiasing techniques
- (M) Implement additional mitigations for sensitive contexts

(R3) Misinterprets or underperforms in response to informal English slang, or non-English languages

Capability risk · Misinformation harms

- (M) Assess outputs for their context
- (M) Treat outputs as suggestions or starting points
- (M) Exercise caution and critical thinking when interpreting model outputs

(R4) Facilitates misinformation by spreading false knowledge

Human interaction risk · Representation and toxicity harms

- (M) Inform end-users they are interacting with an AI system
- (M) Build feedback mechanisms and pipelines to ground responses
- (M) Use available safety classifiers or custom solutions
- (M) Exercise caution and critical thinking when interpreting model outputs

(R5) Increases psychological harm by causing fear and anxiety through threatening statements

Human interaction risk · Malicious use

- (M) Use available safety classifiers or custom solutions
- (M) Implement additional safeguards at the application level
- (M) Implement additional mitigations for sensitive contexts
- (M) Develop ways to reduce model toxicity

(R6) Produces harmful content if explicitly prompted or instructed

Capability risk · Information and safety harms

- (M) Use available safety classifiers or custom solutions
- (M) Build feedback mechanisms and pipelines to ground responses
- (M) Follow transparency best practices
- (M) Implement additional safeguards at the application level

(R7) Generates content that mirrors societal biases

Capability risk · Representation and toxicity harms

- (M) Develop ways to reduce model toxicity
- (M) Assess outputs for their context
- (M) Perform further assessments and additional debiasing techniques
- (M) Implement additional mitigations for sensitive contexts

(R8) Produces inaccurate or outdated content

Capability risk · Misinformation harms

- (M) Inform end-users they are interacting with an AI system
- (M) Assess outputs for their context
- (M) Treat outputs as suggestions or starting points
- (M) Exercise caution and critical thinking when interpreting model outputs

(R9) Facilitates misuse for fraud, spam, or malware production

Capability risk · Information and safety harms

- (M) Manually verify all API uses
- (M) Ensure applications do not violate laws and regulations
- (M) Implement additional safeguards at the application level
- (M) Implement additional mitigations or user consent flows

(R10) Violates privacy by exposing internal messages

Capability risk · Misinformation harms

- (M) Inform end-users they are interacting with an AI system
- (M) Ensure applications do not violate laws and regulations
- (M) Implement additional safeguards at the application level
- (M) Follow transparency best practices

(R11) Violates user trust by bypassing safeguards designed to restrict certain functionalities

Systemic risk · Malicious use

- (M) Manually verify all API uses
- (M) Use available safety classifiers or custom solutions
- (M) Implement additional safeguards at the application level
- (M) Follow transparency best practices

(R12) Struggles or fails to adhere to intricate or nuanced instructions

Capability risk · Representation and toxicity harms

- (M) Build feedback mechanisms and pipelines to ground responses
- (M) Assess outputs for their context
- (M) Treat outputs as suggestions or starting points
- (M) Implement additional safeguards at the application level

(R13) Produces problematic outputs when not aligned to human preferences for safety

Capability risk · Representation and toxicity harms

- (M) Use available safety classifiers or custom solutions
- (M) Perform further assessments and additional debiasing techniques
- (M) Implement additional mitigations for sensitive contexts
- (M) Develop ways to reduce model toxicity

(R14) Uses a mix of Web data and technical sources without transparency

Capability risk · Human autonomy and integrity harms

- (M) Inform end-users they are interacting with an AI system
- (M) Follow transparency best practices
- (M) Implement additional safeguards at the application level
- (M) Release the model for research purposes only

Glossary

Risks

- Capability risk — emerges from the technical components of the model
- Human interaction risk — emerges from the experience of people interacting with the model
- Systemic risk — emerges from the impact of the system on the broader systems in which it is embedded, such as society, the economy, and the natural environment

Harms

- * — real-world harm sourced from the AI Incident Database [www.incidentdatabase.ai]
- Malicious use — emerges when the model lowers costs and barriers for harmful actors to engage in illicit activities
- Misinformation harms — emerge when the model under-, over-, or misrepresents certain groups, generating toxic, offensive, abusive, or hateful content
- Information and safety harms — emerge when the model leaks, reproduces, generates, or infers sensitive, private, or hazardous information
- Representation and toxicity harms — emerge when the model generates and spreads inaccurate or misleading information, causing people to develop false beliefs
- Human autonomy and integrity harms — emerge when the model compromises human agency or circumvents meaningful human control
- Socioeconomic and environmental harms — emerge when the model exacerbates inequalities or negatively impacts employment, innovation, or the environment