Report of 'Phi-3 Technical Report'

Tilian

Original article : Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, Microsoft

Model Details

	Nb of parameters (B = billions)	**Tokens (T = trillions)**
Phi3 mini	3.8 B	3.3 T
Phi3 small	7B	4.8T
Phi3 medium	14B	4.8T

Table 1: Model sizes and token counts

Phi 3 mini

Transformer decoder architecture. Built upon a similar block structure as llama2 and uses the same tokenizer with vocabulary size 32064. All packages developed for Llama-2 family of models can be directly adapted to phi-3-mini. The model uses 3072 hidden dimension, 32 heads, and 32 layers. Already chat-finetuned. Thanks to its small size phi3-mini can be quantized to 4-bits so that it only occupies 1.8GB of memory. They tested the quantized model by deploying phi-3-mini on iPhone 14 with A16 Bionic chip running natively on-device and fully offline achieving more than 12 tokens per second.

Phi 3 small

The phi-3-small model (7B parameters) leverages the tiktoken tokenizer (for better multilingual tokenization) with a vocabulary size of 1003522 and has default context length 8192. It follows the standard decoder architecture of a 7B model class having 32 heads, 32 layers, and a hidden size of 4096. We switched from GELU activation to GEGLU and used Maximal Update Parametrization (muP) to tune hyperparameters on a small proxy model and transfer them to the target 7B model. These helped ensure better performance and training stability.

Phi 3 medium

Embedding dimension: 5120. Uses the same tokenizer and architecture of phi-3-mini and trained on the same data for slightly more epochs.

Training Methodology

High-Quality Training Data

Utilizes high-quality data to enhance the performance of small language models diverging from standard scaling laws.

Standard Scaling Laws

Traditionally, it is believed that increasing the number of parameters in a model (i.e., making the model larger) leads to better performance. This principle is referred to as scaling laws in the field of machine learning.

Alternative Approach

Instead of following this conventional route of increasing model size, the methodology uses superior quality data to train smaller models. This approach shows that smaller models can achieve high performance if trained on well-curated, high-quality datasets.

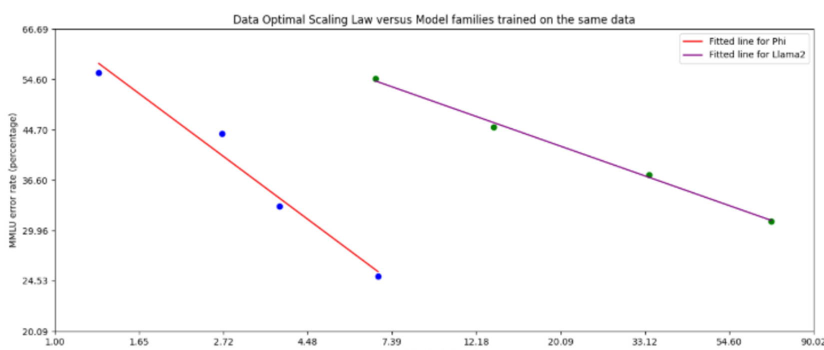


Figure 1: Scaling law close to the "Data Optimal Regime" (from left to right: phi-1.5, phi-2, phi-3-mini, phi-3-small) versus Llama-2 family of models (7B, 13B, 34B, 70B) that were trained on the same fixed data. We plot the log of MMLU error versus the log of model size.

Efficient Model Size

Achieves performance comparable to highly capable models like GPT-3.5 or Mixtral with only 3.8B parameters, whereas Mixtral has 45B parameters.

Data Sources

Combines heavily filtered public web data (based on ”educational level”) and synthetic LLM-generated data. They filter the publicly available web data to contain the correct level of “knowledge” and keep more web pages that could potentially improve the “reasoning ability” for the model. As an example, the result of a game in premier league on 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 mini size models.

Two-Phase Pre-Training

Phase 1

Focuses on web sources for general knowledge and language understanding.

Phase 2

Uses a more refined subset of Phase-1 data and synthetic data to teach logical reasoning and niche skills.

Two-Stage Process

Post-training involves supervised finetuning (SFT) and direct preference optimization (DPO).

Supervised Finetuning (SFT)

- Curated High-Quality Data: Utilizes carefully selected high-quality data across various domains (math, coding, reasoning, conversation, model identity, and safety).
- English-Only Examples: The initial data mix includes only English-language examples.

Direct Preference Optimization (DPO)

- Chat Format and Responsible AI (RAI): Focuses on chat format data, reasoning, and efforts in responsible AI.
- Behavioral Steering: Uses outputs as ”rejected” responses to guide the model away from unwanted behaviors.

Outcome Improvements

- Enhanced Capabilities: Improves the model's performance in math, coding, reasoning, robustness, and safety.
- AI Assistant Transformation: Transforms the language model into an AI assistant for efficient and safe user interaction.

Evaluation of Phi3

- Consistent Evaluation Pipeline: All reported numbers are produced using the same pipeline to ensure comparability across models. These numbers may differ from other published numbers due to variations in evaluation choices.
- Few-Shot Prompts: Evaluations are performed using few-shot prompts at temperature 0, a standard approach. The specifics of the prompts and the number of shots are part of a Microsoft internal tool.
- No Pipeline Optimization: No specific optimizations were made to the evaluation pipeline for the phi-3 models.

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b	Mixtral 8x7b	GPT-3.5 version 1106
MMLU (5-Shot) [HBK+21]	68.8	75.7	78.0	56.3	61.7	63.6	66.5	70.5	71.4
HellaSwag (5-Shot) [ZHB+19]	76.7	77.0	82.4	53.6	58.5	49.8	71.1	70.4	78.8
ANLI (7-Shot) [NWD+20]	52.8	58.1	55.8	42.5	47.1	48.7	57.3	55.2	58.1
GSM-8K (8-Shot; CoT) [CKB+21]	82.5	89.6	91.0	61.1	46.4	59.8	77.4	64.7	78.1
MedQA (2-Shot) [JPO+20]	53.8	65.4	69.9	40.9	50.0	49.6	60.5	62.2	63.4
AGIEval (0-Shot) [ZCG+23]	37.5	45.1	50.2	29.8	35.1	42.1	42.0	45.2	48.4
TriviaQA (5-Shot) [JCWZ17]	64.0	58.1	73.9	45.2	75.2	72.3	67.7	82.2	85.8
Arc-C (10-Shot) [CCE+18]	84.9	90.7	91.6	75.9	78.6	78.3	82.8	87.3	87.4
Arc-E (10-Shot) [CCE+18]	94.6	97.0	97.7	88.5	90.6	91.4	93.4	95.6	96.3
PIQA (5-Shot) [BZGC19]	84.2	86.9	87.9	60.2	77.7	78.1	75.7	86.0	86.6
SociQA (5-Shot) [BZGC19]	76.6	79.2	80.2	68.3	74.6	65.5	73.9	75.9	68.3
BigBench-Hard (3-Shot; CoT) [SRR+22, SSS+22]	71.7	79.1	81.4	59.4	57.3	59.6	51.5	69.7	68.32
WinoGrande (5-Shot) [SLBBC19]	70.8	81.5	81.5	54.7	54.2	55.6	65.0	62.0	68.8
OpenBookQA (10-Shot) [MCKS18]	83.2	88.0	87.4	73.6	79.8	78.6	82.6	85.8	86.0
BoolQ (2-Shot) [CLC+19]	77.2	84.8	86.5	–	72.2	66.0	80.9	77.6	79.1
CommonSenseQA (10-Shot) [THLB19]	80.2	80.0	82.8	69.3	72.6	76.2	79.0	78.1	79.6
TruthfulQA (10-Shot; MC2) [LHE22]	65.0	70.2	75.1	–	53.0	52.1	63.2	60.1	85.8
HumanEval (0-Shot) [CTJ+21]	58.5	61.0	62.2	59.0	28.0	34.1	60.4	37.8	62.2
MBPP (3-Shot) [AON+21]	70.0	71.7	75.2	60.6	50.8	51.5	67.7	60.2	77.8
Average	71.2	75.7	78.5	–	61.2	61.7	69.4	69.8	74.3
GPQA (2-Shot; CoT) [RHS+23]	32.8	34.3	–	–	–	–	–	–	29.0
MT Bench (2 round ave.) [ZCS+23]	8.38	8.70	8.91	–	–	–	–	–	8.35

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b	Mixtral 8x7b	GPT-3.5 version 1106
MMLU (5-Shot) [HBK+21]	68.8 +	75.7	78.0	56.3	61.7	63.6	66.5 −	70.5	71.4
HellaSwag (5-Shot) [ZHB+19]	76.7 +	77.0	82.4	53.6	58.5	49.8	71.1 −	70.4	78.8
ANLI (7-Shot) [NWD+20]	52.8 −	58.1	55.8	42.5	47.1	48.7	57.3 +	55.2	58.1
GSM-8K (8-Shot; CoT) [CKB+21]	82.5 +	89.6	91.0	61.1	46.4	59.8	77.4 −	64.7	78.1
MedQA (2-Shot) [JPO+20]	53.8 −	65.4	69.9	40.9	50.0	49.6	60.5 +	62.2	63.4
AGIEval (0-Shot) [ZCG+23]	37.5 −	45.1	50.2	29.8	35.1	42.1	42.0 +	45.2	48.4
TriviaQA (5-Shot) [JCWZ17]	64.0 −	58.1	73.9	45.2	75.2	72.3	67.7 +	82.2	85.8
Arc-C (10-Shot) [CCE+18]	84.9 +	90.7	91.6	75.9	78.6	78.3	82.8 −	87.3	87.4
Arc-E (10-Shot) [CCE+18]	94.6 +	97.0	97.7	88.5	90.6	91.4	93.4 −	95.6	96.3
PIQA (5-Shot) [BZGC19]	84.2 +	86.9	87.9	60.2	77.7	78.1	75.7 −	86.0	86.6
SociQA (5-Shot) [BZGC19]	76.6 +	79.2	80.2	68.3	74.6	65.5	73.9 −	75.9	68.3
BigBench-Hard (3-Shot; CoT) [SRR+22, SSS+22]	71.7 +	79.1	81.4	59.4	57.3	59.6	51.5 −	69.7	68.32
WinoGrande (5-Shot) [SLBBC19]	70.8 +	81.5	81.5	54.7	54.2	55.6	65.0 −	62.0	68.8
OpenBookQA (10-Shot) [MCKS18]	83.2 +	88.0	87.4	73.6	79.8	78.6	82.6 −	85.8	86.0
BoolQ (2-Shot) [CLC+19]	77.2 −	84.8	86.5	–	72.2	66.0	80.9 +	77.6	79.1
CommonSenseQA (10-Shot) [THLB19]	80.2 +	80.0	82.8	69.3	72.6	76.2	79.0 −	78.1	79.6
TruthfulQA (10-Shot; MC2) [LHE22]	65.0 +	70.2	75.1	–	53.0	52.1	63.2 −	60.1	85.8
HumanEval (0-Shot) [CTJ+21]	58.5 −	61.0	62.2	59.0	28.0	34.1	60.4 +	37.8	62.2
MBPP (3-Shot) [AON+21]	70.0 +	71.7	75.2	60.6	50.8	51.5	67.7 −	60.2	77.8
Average	71.2 +	75.7	78.5	–	61.2	61.7	69.4 −	69.8	74.3
GPQA (2-Shot; CoT) [RHS+23]	32.8	34.3	–	–	–	–	–	–	29.0
MT Bench (2 round ave.) [ZCS+23]	8.38	8.70	8.91	–	–	–	–	–	8.35

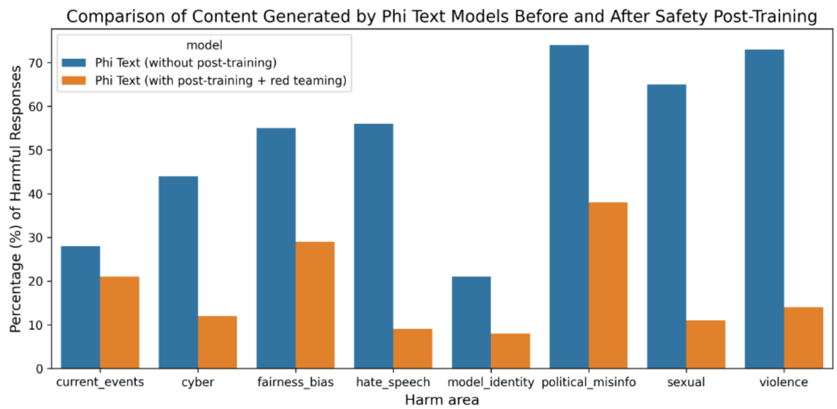
Safety

Key Points Regarding the Safety Measures for phi-3-mini

- Alignment with Responsible AI Principles: Phi-3-mini was developed following Microsoft's responsible AI principles to ensure ethical and safe use.
- Post-Training Safety Alignment: Safety considerations were integrated during the post-training phase using various techniques and datasets.
- Red-Teaming: An independent red team at Microsoft continuously tested the model to identify potential safety issues and areas for improvement.
- Automated Testing and Evaluations: The model was subjected to automated testing and evaluations across multiple responsible AI (RAI) harm categories.
- Use of Diverse Datasets: Safety post-training utilized helpfulness and harmlessness preference datasets as well as several in-house generated datasets modified based on relevant research.

- Iterative Improvement: Feedback from the red team led to the curation of additional datasets which were used to further refine the model's post-training process.

- Reduction of Harmful Responses: The iterative process and refined datasets significantly reduced the rate of harmful responses generated by the model.



For phi3 medium and Phi3 small

- Similar Safety Process: Safety alignment for phi-3-small and phi-3-medium followed the same red-teaming process as phi-3-mini.

- Identical Datasets: The same datasets were used for safety alignment with a slightly larger number of samples incorporated for these models.

Weaknesses

- Limited Factual Knowledge: The model has limited capacity to store factual knowledge, leading to low performance on tasks like TriviaQA.

- Search Engine Augmentation: This weakness can be addressed by augmenting the model with a search engine as demonstrated using the HuggingFace default Chat-UI with phi-3-mini.

- Language Restriction: The model is mostly restricted to English. Expanding to multilingual capabilities is an important next step with initial promising results seen in phi-3-small using more multilingual data.

- Challenges: Despite responsible AI efforts, challenges remain including:
  - Factual inaccuracies (hallucinations)
  - Reproduction or amplification of biases

- Inappropriate content generation
- Safety issues

- Mitigation Efforts: Issues are significantly mitigated through:
  - Carefully curated training data
  - Targeted post-training
  - Red-teaming insights

Phi3 – Vision

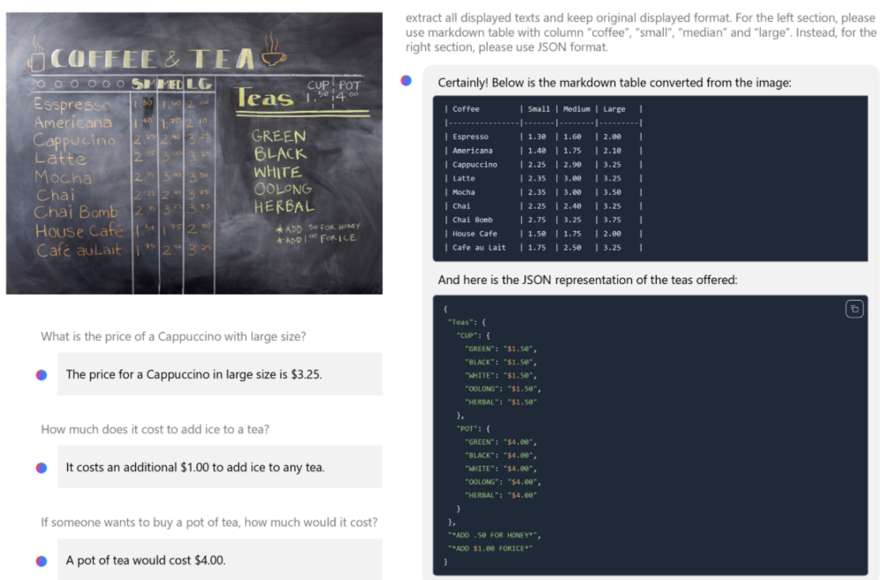


Figure 6: The demo case shows Phi-3-Vision's capability in natural image understanding and reasoning.

Architecture

The Phi-3-Vision (4.2B parameters) is a multimodal model designed to process an image and a textual prompt as inputs and subsequently generate textual outputs. This model is composed of two primary components: an image encoder (Clip : Contrastive Language-Image Pre-training, which encodes images into visual tokens) and a transformer decoder (phi-3-mini-128K-instruct).

Pre-training

- **Supervised Finetuning (SFT):** This stage involves finetuning the model using several datasets:

- These datasets cover various domains and tasks such as natural image understanding, chart/table/diagram understanding/reasoning, PowerPoint understanding, and model safety.
- The multimodal SFT data collectively amounts to approximately 15 billion tokens.

- **Direct Preference Optimization (DPO):** In this stage, the model is further optimized using:
  - A text DPO dataset.
  - A smaller-scale multimodal DPO dataset.
  - The focus here is on enhancing preferences directly related to task-specific criteria.

Academic evaluation

Evaluation Setup:

- Phi-3-Vision was evaluated on nine open-source academic benchmarks that test reasoning and perceptual capabilities using visual and text inputs.

- Benchmarks are categorized into Science, Charts, and Generic knowledge.

	Phi-3-Vision 4.2b	MM1-3B-Chat 3.6b [MGF+24]	MM1-7B-Chat 7.6b [MGF+24]	LLaVA-1.6 Vicuna-7b [LLLL23]	LLaVA-Next LLama3-8b [LLL+24]	Qwen-VL-Chat 9.6b [BBY+23]	Claude 3 haiku [Ant24]	Gemini 1.0 Pro V [TAB+23]	GPT-4V-Turbo turbo-2024-04-09
MMMU (val) [YNZ+23]	40.4	33.9	37.0	34.2	36.4	39.0	40.7	42.0	55.5
ScienceQA (test) [LMX+22]	90.8	69.4	72.6	70.6	73.7	67.2	72.0	79.7	75.7
MathVista (testmini) [LBX+24]	44.5	32.0	35.9	31.5	34.8	29.4	33.2	35.0	47.5
Inter-GPS (test) [LGJ+21]	38.1	-	-	20.5	24.6	22.3	32.1	28.6	41.0
MMBench (dev-en) [LDZ+24]	80.5	75.9	79.0	76.3	79.4	75.8	62.4	80.0	86.1
POPE (test) [LDZ+23]	85.8	87.4	86.6	87.2	87.0	82.6	74.4	84.2	83.7
AI2D (test) [KSK+16]	76.7	-	-	63.1	66.9	59.8	60.3	62.8	74.7
ChartQA (test) [MLT+22]	81.4	-	-	55.0	65.8	50.9	59.3	58.0	62.3
TextVQA (test) [SNS+19]	70.9	71.9	72.8	64.6	55.7	59.4	62.7	64.7	68.1

Safety

- **Integration with Responsible AI Principles:** The integration of Phi-3-Vision aligns with Microsoft's Responsible AI (RAI) principles by incorporating safety measures during both Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) stages.

- **Safety Training Datasets:** Safety training datasets were created using text-only RAI datasets and in-house Multi-Modal (MM) RAI datasets. These datasets cover various harm categories identified in both public and internal benchmarks.

- **Evaluation Methodology:** Evaluation of Phi-3-Vision's RAI performance involved rigorous quantitative assessments on both public and internal benchmarks. Human evaluations were conducted by Microsoft's internal red team to complement the quantitative assessments.

- **Benchmark Results:** Phi-3-Vision's performance on three MM RAI benchmarks (one internal and two public benchmarks: RTVLM and VLGuard) was compared with results from other open-source models such as Llava-1.5, Llava-1.6, Qwen-VL-Chat, and GPT4-V. The results show that safety post-training significantly enhances Phi-3-Vision's RAI performance across all benchmarks.

	Phi-3-Vision 3.8b+0.3b	Phi-3-Vision w/o safety 3.8b+0.3b	Llava-1.6 Vicuna 7b+0.3b	Qwen-VL-Chat 7.7b+1.9b	GPT4-V N/A
Internal (private)	8.30	7.06	5.44	7.27	8.55
RTVLM (public)	4.64	3.56	3.86	4.78	6.81
VLGuard (public)	9.12	4.66	5.62	8.33	8.90

Table 3: Comparison results on public and private multi-modal RAI benchmarks. Note that all metrics in the table are [0,10] and a higher value indicates a better performance.

Conclusion on Phi-3-Vision

- **Performance:** Phi-3-Vision demonstrates strong multi-modal language modeling (LLM) capabilities across various domains but faces limitations in tasks requiring high-level reasoning abilities. It also occasionally produces ungrounded outputs, particularly in sensitive domains like finance.

- **Mitigation Strategy:** To address these issues, future plans include incorporating more data focused on reasoning and hallucination-related scenarios using Direct Preference Optimization (DPO) post-training.

- **Responsible AI Concerns:** Despite advancements in safety post-training, Phi-3-Vision sometimes fails to avoid answering harmful or sensitive queries.

Examples include deciphering specific types of captchas and describing scam images containing disinformation or hallucination.

- **Trade-off:** Issues like these stem partly from the capabilities acquired during training, such as OCR, which present a trade-off between usefulness and potential harm. Achieving a better balance between these aspects is a key focus for future development.

Index: Explanation of terms from evaluation of Phi-3-Vision

- **MMLU:** Massive Multitask Language Understanding. A benchmark that tests a language model's ability to handle a wide range of tasks across different domains.
- **HellaSwag:** A benchmark for commonsense reasoning in natural language understanding, where models are evaluated on their ability to choose the most plausible continuation of a given story or situation.
- **ANLI:** Adversarial Natural Language Inference. A benchmark for natural language inference (NLI) that includes adversarially crafted examples designed to be challenging for models.
- **GSM-8K:** A dataset for evaluating models on grade-school math word problems. The "8K" refers to the size of the dataset, which includes 8,000 problems.
- **MedQA:** A benchmark for medical question answering, where models are evaluated on their ability to answer questions from medical exams.
- **AGIEval:** A benchmark for evaluating a model's ability on academic exams, designed to test general intelligence and education-level knowledge.
- **BoolQ:** Boolean Questions. A benchmark where models are evaluated on their ability to answer yes/no questions based on given passages.
- **Arc:** A benchmark from the Allen Institute for AI, evaluating models on multiple-choice science questions from grade-school exams. It has two subsets: Arc C (Challenge) and Arc E (Easy).
- **PIQA:** Physical Interaction QA. A benchmark for commonsense reasoning about physical interactions, where models are evaluated on their ability to choose the most plausible physical interaction.
- **SIQA:** Social Interaction QA. A benchmark for commonsense reasoning about social interactions, where models are evaluated on their ability to understand and predict social situations.
- **BigBench-Hard:** A subset of the BigBench benchmark, which includes diverse and challenging tasks across different domains.
- **WinoGrande:** A benchmark for commonsense reasoning, where models are evaluated on their ability to resolve ambiguities in sentences.
- **OpenBookQA:** A benchmark for science question answering, where models are evaluated on their ability to answer questions based on a small "open book" of facts.

- **CommonSenseQA:** A benchmark for commonsense reasoning, where models are evaluated on their ability to answer multiple-choice questions requiring commonsense knowledge.
- **TruthfulQA:** A benchmark for evaluating the truthfulness of language models, where models are tested on their ability to avoid generating false or misleading information.
- **HumanEval:** A benchmark for evaluating the performance of models on programming tasks, where models are tested on their ability to generate correct and efficient code.
- **MBPP:** The Multi-lingual Bible Parallel Project, a dataset for evaluating the translation quality of models across multiple languages.
- **GPQA:** General Physical QA. A benchmark for evaluating models on general physical reasoning tasks.
- **MT Bench:** A benchmark for evaluating the performance of models on multitasking scenarios, where models are tested on their ability to handle multiple tasks simultaneously.