Claude 3

Anthropic announces Claude 3, their new family of models that include Claude 3 Haiku, Claude 3 Sonnet, and Claude 3 Opus.

Claude 3 Opus (the strongest model) is reported to outperform GPT-4 and all other models on common benchmarks like MMLU and HumanEval.

Results and Capabilities

Claude 3 capabilities include advanced reasoning, basic mathematics, analysis, data extraction, forecasting, content creation, code generation, and converting in non-English languages like Spanish, Japanese, and French. The table below demonstrates how Claude 3 compares with other models on several benchmarks with Claude 3 Opus outperforming all the mentioned models:

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Claude 3 Haiku is the fastest and most cost-effective model of the series. Claude 3 Sonnet is 2x faster than previous iterations of Claude and Opus is as fast as Claude 2.1 with more superior capabilities.

The Claude 3 models offer support for 200K context windows but can be extended to 1M tokens to select customers. Claude 3 Opus achieved near-perfect recall on the Needle In A Haystack (NIAH) evaluation which measures the model's ability to recall information in a large corpus and effectively process long context prompts.

The models also have strong vision capabilities for processing formats like photos, charts, and graphs.

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4V	Gemini 1.0 Ultra	Gemini 1.0 Pro
Math & reasoning *MMMU (val)*	59.4%	53.1%	50.2%	56.8%	59.4%	47.9%
Document visual Q&A *ANLS score, test*	89.3%	89.5%	88.8%	88.4%	90.9%	88.1%
Math *MathVista (testmini)*	50.5% CoT	47.9% CoT	46.4% CoT	49.9%	53.0%	45.2%
Science diagrams *AI2D, test*	88.1%	88.7%	86.7%	78.2%	79.5%	73.9%
Chart Q&A *Relaxed accuracy (test)*	80.8% 0-shot CoT	81.1% 0-shot CoT	81.7% 0-shot CoT	78.5% 4-shot CoT	80.8%	74.1%

Anthropic also claim that these models have a more nuanced understanding of requests and make fewer refusals. Opus also shows significant improvements in factual question answering in open-ended questions while reducing incorrect answers or hallucinations. Claude 3 models are also better than the Claude 2 models at producing structured outputs like JSON objects.

References

- Claude 3 Haiku, Claude 3 Sonnet, and Claude 3 Opus
- The Claude 3 Model Family: Opus, Sonnet, Haiku

Last updated on September 19, 2024

Copyright © 2024 DAIR.AI