phind

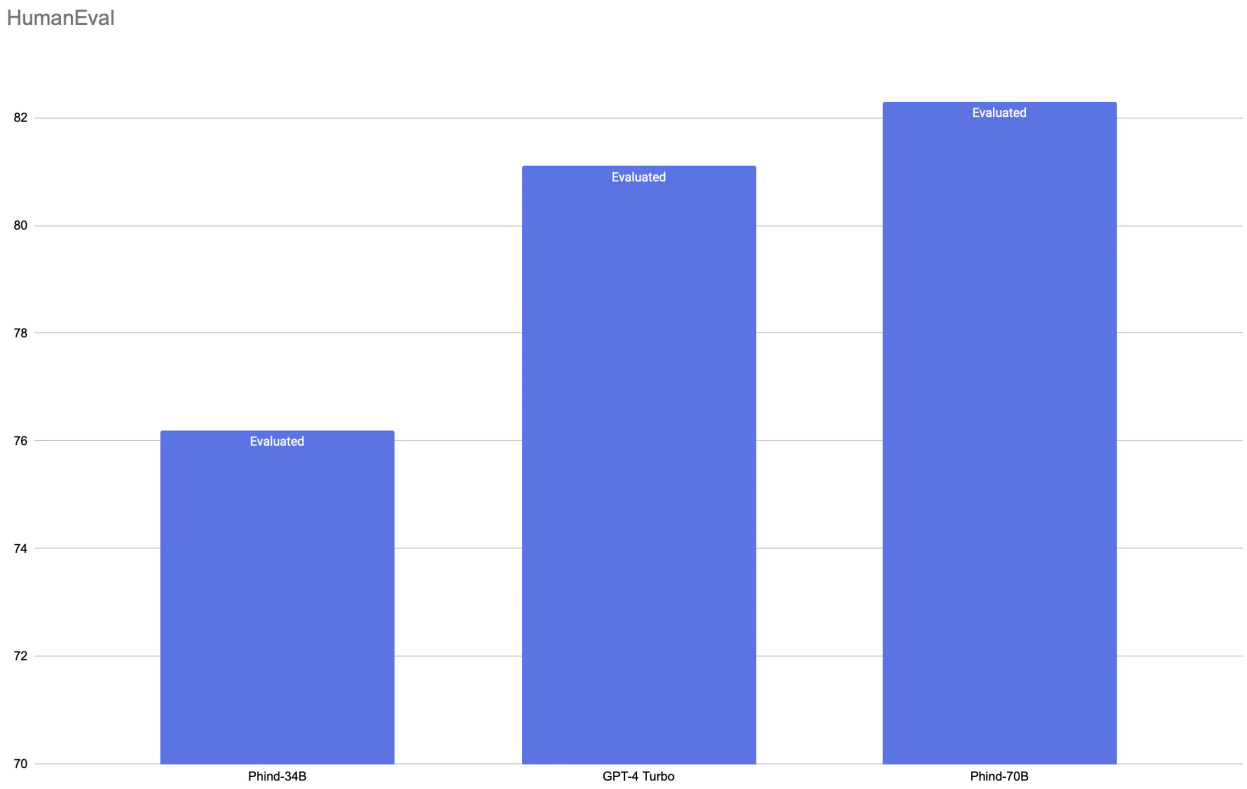
Sign In

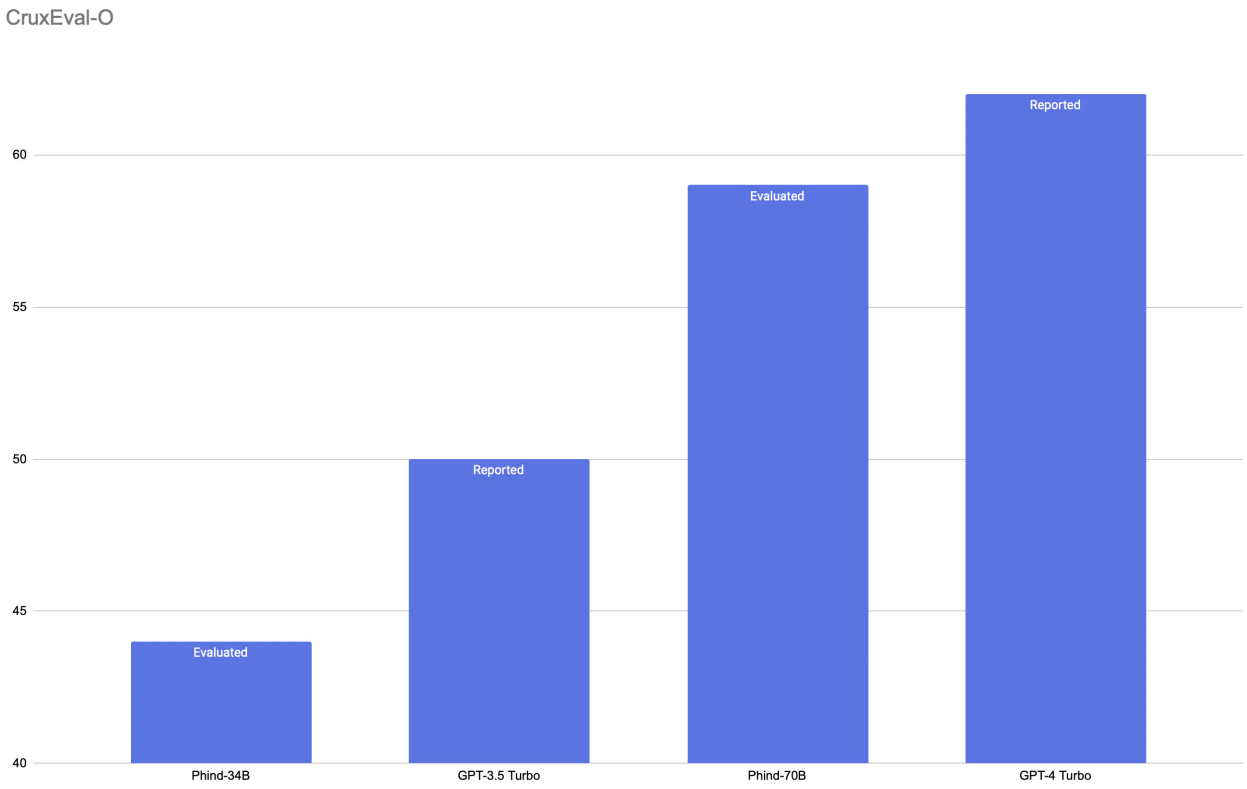
Introducing Phind-70B – closing the code quality gap with GPT-4 Turbo while running 4x faster

We're excited to announce Phind-70B, our largest and most performant model to date. Running at up to 80 tokens per second, Phind-70B gives high-quality answers for technical topics without making users make a cup of coffee while they wait. We think it offers the best overall user experience for developers amongst state-of-the-art models.

Phind-70B is based on the CodeLlama-70B model and is fine-tuned on an additional 50 billion tokens, yielding significant improvements. It also supports a context window of 32K tokens.

Phind-70B scores 82.3% on HumanEval, beating the latest GPT-4 Turbo (gpt-4-0125-preview) score of 81.1% in our evaluation. On Meta's CRUXEval dataset, Phind-70B scores 59% to GPT-4's reported score of 62% on the output prediction benchmark. However, neither of these public datasets fully captures how our users use Phind for real-world workloads. We find that Phind-70B is in the same quality realm as GPT-4 Turbo for code generation and exceeds it on some tasks. Phind-70B is also less “lazy” than GPT-4 Turbo and doesn't hesistate to generate detailed code examples.





Phind-70B is significantly faster than GPT-4 Turbo, running at 80+ tokens per second to GPT-4 Turbo's ~20 tokens per second. We're able to achieve this by running NVIDIA's TensorRT-LLM library on H100 GPUs, and we're working on optimizations to further increase Phind-70B's inference speed.

Phind-70B is available today to try for free and without a login. You can get higher limits by subscribing to Phind Pro.

We love the open-source community and will be releasing the weights for the latest Phind-34B model in the coming weeks. We intend to release the weights for Phind-70B in time as well.

We'd like to thank our cloud partners, SF Compute and AWS, for helping us get the infrastructure right for training and serving Phind-70B. We'd also like to thank our partners at Meta and NVIDIA for their support.

Fun fact: We melted an H100 during Phind-70B's training!

Try Phind-70B