Unveiling Hermes 3: The First Full-Parameter Fine-Tuned Llama 3.1 405B Model is on Lambda's Cloud

Try Hermes 3 for free with the New Lambda Chat Completions API and Lambda Chat.

Introducing Hermes 3: A new era for Llama fine-tuning

We are thrilled to announce our partner Nous Research's launch of Hermes 3 —the first full-parameter fine-tune of Meta's groundbreaking Llama 3.1 405B model, trained on Lambda's 1-Click Cluster. Designed for the open-source community, Hermes 3 is a neutrally-aligned generalist model with exceptional reasoning capabilities, now available for free through the new Lambda Chat Completions API and Lambda Chat interface.

Powered by an 8-node Lambda 1-Click Cluster, Nous Research achieved outstanding results in just a few short weeks. Hermes 3 meets or exceeds Llama 3.1 Instruct on Open Source LLM benchmarks (see table below).

"Lambda's 1-Click Clusters make the experience of renting and using a multi-node cluster as simple and easy as renting and using a single node,"

-Jeffrey Quesnelle, co-founder of Nous Research

Hermes 3: A uniquely unlocked, uncensored, and steerable model

Hermes 3 is the latest advancement in Nous Research's series of models, which have been downloaded over 33 million times. This instruct-tuned model is specifically designed to be flexible and adept at following instructions. It excels in complex role-playing and creative writing, offering users more immersive character portrayals, deeper simulations, and unexpected fictional experiences.

Metric	Hermes 3 405B	Llama 3.1 Instruct 405B	Hermes 3 70B	Llama 3.1 Instruct 70B	Hermes 3 8B	Llama 3.1 Instruct 8B
AGIEval *0-shot*	**61.84**	58.60	**56.18**	48.26	**41.26**	40.49
ARC-C *0-shot*	**69.45**	66.04	**65.53**	63.40	**58.11**	55.12
ARC-E *0-shot*	**86.24**	85.40	82.95	**83.67**	**80.05**	79.71
BoolQ *0-shot*	88.93	**89.52**	**88.04**	87.76	**84.95**	84.01
BBH *3-shot*	75.37	**76.25**	67.82	**69.24**	**52.94**	48.83
GPQA *0-shot*	**44.84**	42.66	37.67	**40.09**	29.36	**30.62**
Hellaswag *10-shot*	**90.19**	88.34	**88.19**	86.42	**82.83**	80.01
IFEval *Strict*	84.87	**87.09**	81.21	**87.25**	62.25	**80.15**
MATH Lvl 5 *4-shot*	30.85	**35.98**	20.80	**29.24**	7.48	**8.91**
MMLU *5-shot*	85.02	**86.14**	79.09	**82.27**	64.79	**68.05**
MMLU-PRO *5-shot*	54.14	**63.51**	47.24	**52.94**	32.08	**35.77**
MT-Bench *Avg.*	8.93	**9.17**	**8.99**	8.93	8.27	**8.39**
MuSR *0-shot*	**48.26**	47.58	**50.67**	47.08	**43.52**	38.23
OpenbookQA *0-shot*	**48.80**	48.60	**49.40**	47.20	**47.80**	43.20
PiQA *0-shot*	**85.96**	84.93	**84.44**	83.73	80.25	**81.01**
TruthfulQA *MC2 0-shot*	**65.57**	64.83	**63.29**	59.91	**58.69**	53.99
Winogrande *5-shot*	86.27	**86.82**	83.19	**85.00**	77.74	**77.90**

In addition to its creative capabilities, Hermes 3 is an invaluable tool for professionals requiring advanced reasoning and decision-making abilities. Its strategic planning and operational decision-making features include function-calling, step-labeled reasoning, and more.

Optimized for efficiency

Hermes 3 was meticulously trained using synthesized data and supervised fine-tuning on Meta's Llama 3.1 405B base model. This was followed by reinforcement learning from human feedback (RLHF) and finally, quantization using Neural Magic's FP8 method.

This optimization effectively reduces the model's VRAM and disk requirements by approximately 50%, allowing it to run on a single node.

"Since the start of my journey in AI I wanted to bring about the realization of an open source frontier level model that aligns to you, the user - not some

corporation or higher authority before the user. Today, with Hermes 3 405B, we've achieved that goal, a model that is frontier level, but truly aligned to you.

Thanks to our hard work on data synthesis and post training research, we were able to make a dataset that is fully synthetic over almost a year in the making to train Hermes 3 - and will be releasing much more to come.”

-Teknium, cofounder of Nous Research

For those seeking dedicated access and flexibility, Hermes 3 can run on a single node (available on-demand on Lambda’s Cloud), or quickly scale to a multi-node 1-Click Cluster for further fine-tuning using Lambda's scalable cluster infrastructure.