



MARKET REPORT

UPDATED

05/12/2024

Together AI: the \$44M/year Vercel of generative AI

TEAM

Jan-Erik Asplund

Co-Founder

jan@sacra.com

DISCLAIMERS

This report is for information purposes only and is not to be used or considered as an offer or the solicitation of an offer to sell or to buy or subscribe for securities or other financial instruments. Nothing in this report constitutes investment, legal, accounting or tax advice or a representation that any investment or strategy is suitable or appropriate to your individual circumstances or otherwise constitutes a personal trade recommendation to you.

This research report has been prepared solely by Sacra and should not be considered a product of any person or entity that makes such report available, if any.

Information and opinions presented in the sections of the report were obtained or derived from sources Sacra believes are reliable, but Sacra makes no representation as to their accuracy or completeness. Past performance should not be taken as an indication or guarantee of future performance, and no representation or warranty, express or implied, is made regarding future performance. Information, opinions and estimates contained in this report reflect a determination at its original date of publication by Sacra and are subject to change without notice.

Sacra accepts no liability for loss arising from the use of the material presented in this report, except that this exclusion of liability does not apply to the extent that liability arises under specific statutes or regulations applicable to Sacra. Sacra may have issued, and may in the future issue, other reports that are inconsistent with, and reach different conclusions from, the information presented in this report. Those reports reflect different assumptions, views and analytical methods of the analysts who prepared them and Sacra is under no obligation to ensure that such other reports are brought to the attention of any recipient of this report.

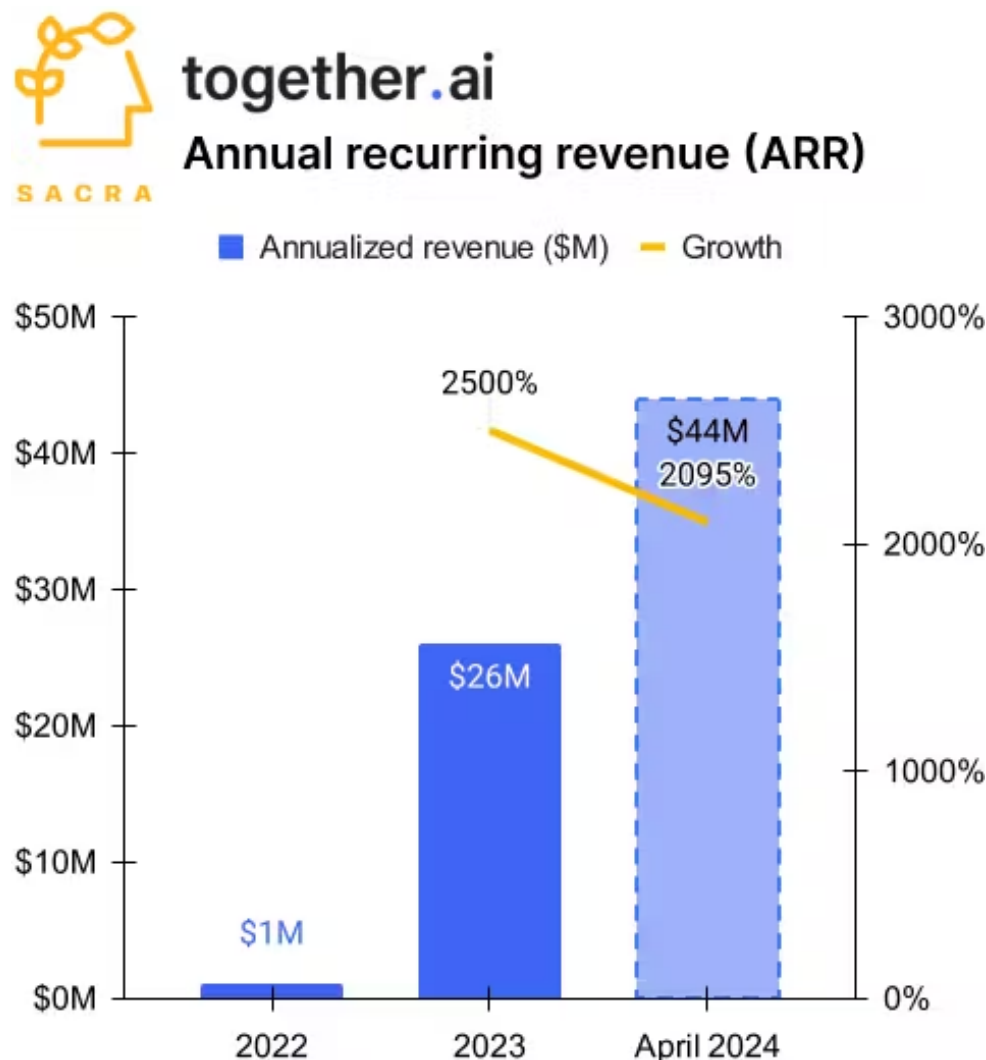
All rights reserved. All material presented in this report, unless specifically indicated otherwise is under copyright to Sacra. Sacra reserves any and all intellectual property rights in the report. All trademarks, service marks and logos used in this report are trademarks or service marks or registered trademarks or service marks of Sacra. Any modification, copying, displaying, distributing, transmitting, publishing, licensing, creating derivative works from, or selling any report is strictly prohibited. None of the material, nor its content, nor any copy of it, may be altered in any way, transmitted to, copied or distributed to any other party, without the prior express written permission of Sacra. Any unauthorized duplication, redistribution or disclosure of this report will result in prosecution.

Published on **May 12th, 2024**

Together AI: the \$44M/year Vercel of generative AI

By Jan-Erik Asplund

TL;DR: Sacra estimates that Together AI hit \$44M in annual recurring revenue (ARR) in April with their developer experience-centric layer built on top of CoreWeave and Lambda Labs's GPU cloud products. For more, check out our full report and dataset on Together AI.



Key points from our research:

- **Companies use cloud GPU hosts Together AI, CoreWeave, and Lambda Labs to use Nvidia (NASDAQ: NVDA) graphics processing units (GPUs) to train AI models on their datasets, fine-tune them, and deploy them into production. Together AI differentiated itself as a GPU cloud platform early by indexing on open source, allowing its customers access to 100+ open models from Mistral to Llama-**



2 to rapidly experiment with training different LLMs on their data.

- **Together AI found product-market fit charging per-token, based on the number of API calls, as a developer experience-centric layer on top of CoreWeave and Lambda Labs's per-hour pricing, allowing them to win early-stage startups and individual developers who can mitigate the risk of paying for idle GPU time.** While CoreWeave and Lambda Labs focus on locking in multi-year reservations to recoup the fixed capex costs of their data centers and GPUs, Together AI operates a layer above, aligning their pricing with the spiky API volumes of startups training new models and launching new products.
- **Sacra estimates that Together AI hit \$44M in annualized revenue in April 2024, up 2095% year-over-year, with 90% of revenue coming from sales of bundled on-demand GPU compute and training via Together Inference and ~45% gross margin.** Compare to ~80% gross margin GPU cloud providers like CoreWeave with \$440M of revenue in 2023, up 1,660% from \$25M in 2022, and Lambda Labs at \$250M of revenue in 2023, up 1,150% from \$20M in 2022, and fellow reseller Crusoe Energy at \$100M of revenue in 2023, up 400% from ~\$20M in 2022.

For more, check out this other research from our platform:

- OpenAI vs. Anthropic vs. Cohere
- Perplexity: the \$11M/year Cliff Notes for the web growing 4,272%
- Perplexity (dataset)
- Anthropic (dataset)
- OpenAI (dataset)
- Scale (dataset)
- Hugging Face (dataset)
- CoreWeave (dataset)
- Lambda Labs (dataset)
- Scale (dataset)
- Apple vs. Limitless vs. Gong
- Jenni AI: the \$5M/year Chegg of generative AI
- David Park, CEO and co-founder of Jenni AI, on prosumer generative AI apps post-ChatGPT



- AI writing goes enterprise
- Photoroom: the \$65M/year background removal app
- Hugging Face: the \$70M/year anti-OpenAI growing 367% year-over-year