

APRIL 2025

# Open Source LLMs for Everyone at Scale: Red Hat Acquires Neural Magic

Torsten Volk, Principal Analyst

## Overview

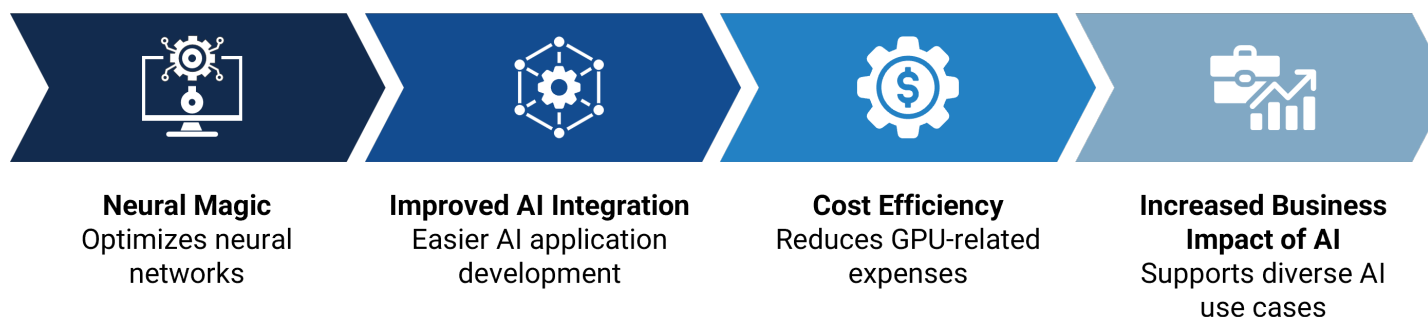
Red Hat's acquisition of Neural Magic addresses enterprise demand for greater performance and cost efficiency when building, deploying, and managing AI-driven applications, enabling organizations to select and optimize open source large language models (LLMs) according to their individual requirements. By integrating Neural Magic's expertise in inference performance engineering and model optimization, Red Hat strengthens its AI portfolio, complementing existing capabilities for scalable AI lifecycle orchestration across hybrid cloud environments. This combination has the potential to notably increase Red Hat's differentiation in the rapidly evolving generative AI landscape.

## Business Importance

The high cost of specialized hardware (e.g. GPUs, TPUs, NPUs, and FPGAs) that enables the rapid training, tuning and inference of LLMs, combined with high electricity consumption, demanding cooling requirements, and complex compliance standards and regulations, prevent application development teams from harnessing the advantages of AI at scale. In a nutshell, instead of being able to harness AI capabilities to ensure optimal user experiences and the best possible application functionality, organizations are often deterred by the large price tags of obtaining the required infrastructure resources.

Figure 1. Red Hat and Neural Magic Combination

### Red Hat Enhances AI Capabilities With Neural Magic



Source: Enterprise Strategy Group, now part of Omdia

The quickly increasing popularity of agentic AI adds cost and performance pressure, due to these agents solving problems collaboratively, while at the same time consuming large numbers of costly LLM tokens for communication and coordination. Different agents can leverage different open source LLMs to contribute to this iterative problem-solving approach, each LLM with its own set of performance characteristics. These challenges are often

compounded by a lack of developer skill and experience in configuring and optimizing complex AI workflows, leading to increased operational costs and delays.

Red Hat acquired Neural Magic to make it more cost efficient and easier for development teams to add LLM capabilities to their applications. Adding an “AI accelerator” to their Red Hat AI deployments could significantly improve Red Hat’s position in the marketplace for cloud-native application platforms.

## Background

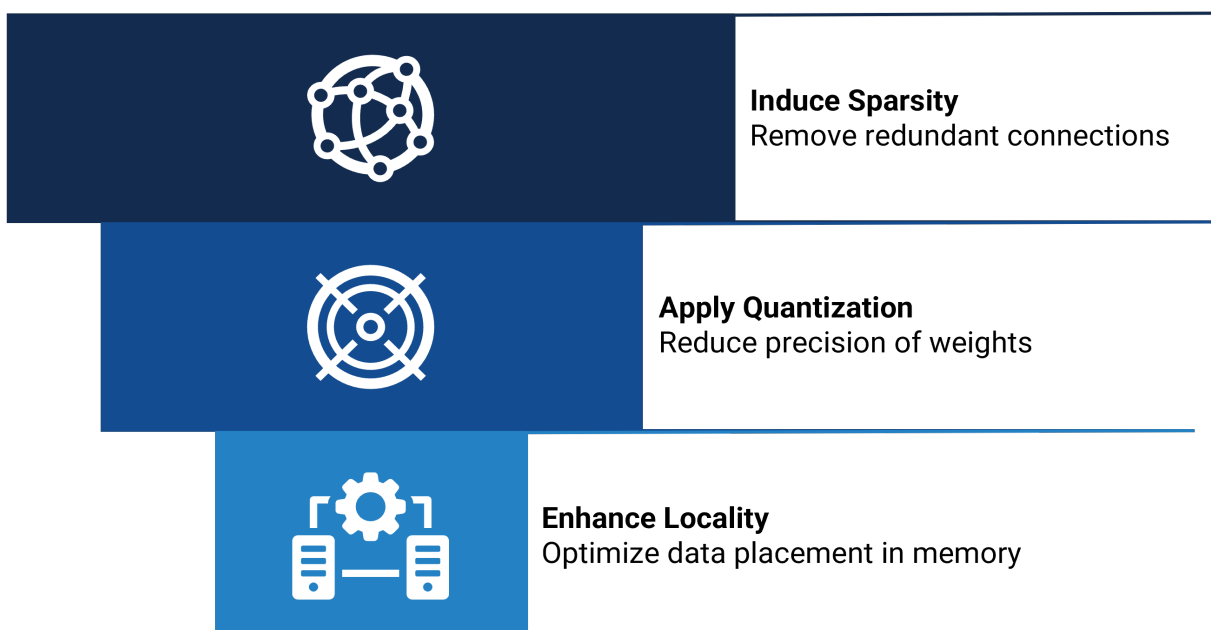
Red Hat’s acquisition of Neural Magic, while undisclosed in terms of financials, brings on board a team of approximately 40 employees with deep expertise in optimizing, deploying and managing LLMs. Neural Magic secured \$35 million in its latest round of venture funding in late 2021, bringing its total funding to \$55 million. The company is renowned for its contributions to the vLLM open-source project, dedicated to improving the efficiency, scalability, management, accuracy and performance of LLMs. With a portfolio of 60 patents and a wealth of expertise reflected in over 200 research papers on AI model optimization, and general machine learning and high-performance computing research, Neural Magic holds significant intellectual property and engineering talent. This infusion of talent and intellectual property in the AI arena could become an important asset for Red Hat.

## Key Technologies Explained

Neural Magic’s ability to optimize the resource efficiency of open source LLMs—including Llama, Mistral, Gemma, Granite and Qwen—has clear benefits for organizations looking to leverage Red Hat’s hybrid cloud product portfolio. By employing advanced mathematical operations and algorithms to reduce the hardware footprint of these LLMs, Neural Magic helps businesses deploy more LLM workloads without having to purchase more GPUs, TPUs, IPUs, RDUs, etc. This improved efficiency could become a compelling factor when choosing Red Hat’s hybrid cloud application stack over competing platforms.

**Figure 2.** Optimizing LLMs for Efficient Deployment

### Optimizing LLMs for Efficient Deployment



Source: Enterprise Strategy Group, now part of Omdia

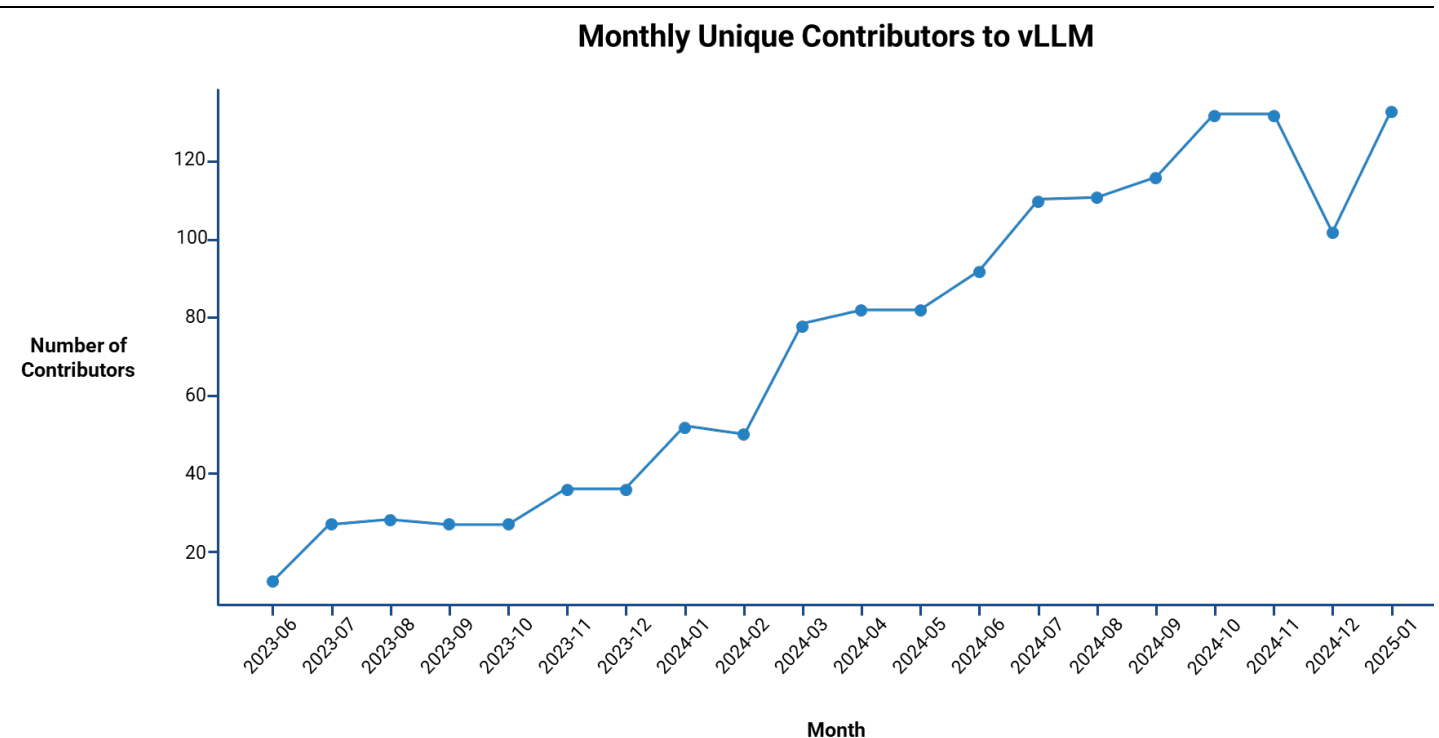
In addition to broad open source LLM support, Neural Magic has been the top commercial contributor to the vLLM open-source inferencing server, originally started at University of California, Berkeley. This community project focuses on simplifying the deployment and lifecycle management of LLMs, while at the same time optimizing performance and resource efficiency. This lets users run already optimized turnkey LLMs from the HuggingFace Hub. More experienced users with more specific requirements, such as advanced domain adaptation or integrating proprietary data sets, can fine-tune the base models for specialized use cases (e.g., medical text analysis or financial forecasting). Neural Magic is taking an active role in supporting vLLM inferencing across heterogeneous accelerators, including Nvidia, AMD, Google TPUs, and Intel, bringing choice and optionality to inference deployments.

Key to Neural Magic's approach is the incorporation of sparsity and quantization into vLLM, an open source inferencing server. Sparsity applies a number of mathematical principles (weight pruning, sparse activation, low-rank factorization, and regularization techniques) to only use the parts of the AI model that are critical for accurate inferencing. Not only do sparse models require less memory, they can also move the data closer together (locality of reference) to ensure more efficient use of processor cache, further accelerating inference speed.

Quantization reduces the resolution (number of dimensions) of the input weights of the model, enabling more efficient computation and reduced memory usage while maintaining performance. Combined, sparsity and quantization reduce memory usage and computational demands, enhancing resource efficiency for LLMs in real-life use cases.

## Open Source

Neural Magic's emphasis on open-source software and AI models aligns closely with Red Hat's own commitment. By contributing to the rapidly growing vLLM open source project, Neural Magic not only showcases its ability to optimize LLMs for performance and resource consumption, but also benefits from community feedback and collaboration. In addition to 14 regular contributors from Neural Magic, vLLM draws support from University of California, Berkeley and, interestingly enough, from IBM/Red Hat.

**Figure 3.** Number of Unique Code Committers on GitHub per Month

Source: GitHub API

Open source is central to Neural Magic's strategy: It enables developers, researchers, and enterprises to freely access and modify its tools, fueling rapid innovation in areas such as model integration, production monitoring (e.g., Prometheus metrics and Grafana dashboards), multi-accelerator support for scalable inference, Kubernetes integration, and advanced mathematical techniques for performance optimization. Meanwhile, Red Hat, long synonymous with enterprise open source, expands its AI footprint by incorporating Neural Magic's open source projects into its broader hybrid cloud ecosystem. This alignment underlines Red Hat's commitment to delivering transparent, scalable, community-driven solutions—empowering customers to adopt AI more easily and with full visibility into the underlying technology stack.

All of this reinforces the argument that Red Hat's hybrid cloud application platform, combined with Neural Magic's optimizations, can deliver both cost and performance advantages for organizations looking to maximize LLM potential in their hybrid cloud environments.

## Conclusion

Complementing Red Hat's already existing ability to accelerate the deployment and management of AI-driven applications on Red Hat AI, Neural Magic aims at helping Red Hat customers significantly expand the range of AI use cases by tearing down key obstacles such as high cost and limited accessibility of specialized hardware. Providing ready-made LLMs and enterprise-grade support helps customers address skill gaps and ensure the desired level of uptime for business-critical AI models.

We will continue to watch Red Hat's efforts of wrapping the Neural Magic capabilities into turnkey components for Red Hat AI, providing DevOps teams with 'training wheels' for adding AI-driven capabilities to their software projects. Considering that Red Hat's strategy is all about helping enterprises develop, deploy, and manage high-

value applications faster and more economically, this acquisition is spot-on and we look forward to seeing the first platform integrations soon.

©2025 TechTarget, Inc. All rights reserved. The Informa TechTarget name and logo are subject to license. All other logos are trademarks of their respective owners. Informa TechTarget reserves the right to make changes in specifications and other information contained in this document without prior notice.

Information contained in this publication has been obtained by sources Informa TechTarget considers to be reliable but is not warranted by Informa TechTarget. This publication may contain opinions of Informa TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent Informa TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, Informa TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.


Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of Informa TechTarget, is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact Client Relations at [cr@esg-global.com](mailto:cr@esg-global.com).

---

**About Enterprise Strategy Group**

Enterprise Strategy Group, now part of Omdia, provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

 [contact@esg-global.com](mailto:contact@esg-global.com)

 [www.esg-global.com](http://www.esg-global.com)