4 reasons to use open source small language models

Small language models are transforming enterprise AI strategies

Proprietary large language models (LLMs) excel in general-purpose applications, but they are not always the best fit for enterprise artificial intelligence (AI) solutions. Their significant computational requirements, opaque decision-making processes, and high licensing costs can limit flexibility and increase operational complexity. Small language models (SLMs), particularly those based on open source principles, offer an alternative for organizations looking to develop customized AI solutions, maintain control over data, and manage costs effectively.

Here are 4 reasons why open source SLMs may be the right choice for your next AI project.

1 Access community innovation

Combining flexibility, collaboration, and innovation, open source SLMs offer a foundation for building highly adaptable and specialized AI applications. By providing access to both software components and pretrained model weights, open source AI projects let you collaborate with a global community of developers and researchers to continuously refine and improve generative AI (gen AI) technologies. This shared innovation empowers you to work with modern, advanced tools and tailor them to meet the technical needs of your enterprise AI solutions.

With open source SLMs like the IBM Granite family of gen AI models, you can directly contribute knowledge and domain expertise to foundation models. Rather than waiting for updates to proprietary LLMs, you can actively customize open source SLMs to increase their relevance and performance in your AI applications. This interactive approach lets you iterate faster and keep your models current with evolving business needs.

Open source SLMs offer essential flexibility for deployment in dynamic environments across on-site datacenters and public cloud infrastructure. With full control over your models, you can optimize them for a range of deployment scenarios, from high-compliance environments to real-time AI processing. And by helping you maintain control over your AI technology stack, open source SLMs ensure that your innovative AI solutions remain adaptable and scalable as your technical and business needs change.

2 Gain control over training data

Open source SLMs offer greater transparency compared to proprietary alternatives. Because trusted providers disclose the data used to pretrain these models, you can thoroughly assess model quality and confirm that no harmful or biased information is included. This transparency allows you to make informed decisions about adapting and deploying models. As a result, you can ensure that your AI solutions meet ethical standards and business objectives before incorporating your own proprietary, confidential data.

Additionally, because you can deploy SLMs in on-site datacenters and private cloud resources across your enterprise IT environments, you can maintain full control over your training data. This control is crucial for organizations handling highly confidential or regulated data, as you can ensure that proprietary information is never exposed to external providers. And by managing your gen AI models within your own environment, you can control access, streamline regulatory compliance, enhance data security, and maintain greater transparency across your AI solutions.

Finally, the IBM Granite family of models offers assurance policies that indemnify their customers from claims that the open source software or AI models provided violate a third party's intellectual property rights. Choosing these models and vendors can help further protect your organization in a complex and changing AI technology landscape.

3 Customize your AI solutions

Open source SLMs let you rapidly and efficiently develop AI solutions tailored to your specific business requirements. Designed and built for targeted use cases, these models let you address domain-specific challenges with precision while avoiding the complexity and resource demands of general-purpose LLMs.

By tuning SLMs with your enterprise data, you can embed organizational knowledge and domain expertise directly into model parameters. This approach improves the relevance of SLM responses, reduces the frequency and cost of retraining, and shortens development timelines for critical AI applications and services.

With compact sizes and reduced data requirements compared to LLMs, SLMs are easier to customize, allowing you to develop accurate, efficient models optimized for specific tasks or domains. And in resource-constrained environments and edge deployments, SLMs allow real-time applications to run directly on user devices, simplifying development and eliminating the need for external cloud infrastructure.

SLMs like IBM Granite models also streamline the transition from experimentation to production. Simplified integration of SLMs with diverse hardware and software infrastructure gives you the ability to tailor your gen AI solutions to your enterprise IT environment. This adaptability helps reduce operational complexity while maintaining control over deployment and performance.

4 Reduce AI model costs

For many enterprise organizations, reducing the computational demands of AI is critical to effectively managing expenses. Open source SLMs deliver the performance needed for advanced gen AI solutions while lowering the cost of training and inferencing, as well as the required computing power, compared to LLMs.

With a reduced size that is often thousands of times smaller than leading LLMs, SLMs require far less compute resources, data, and energy. This efficiency supports faster training times, easier fine-tuning, and a more sustainable approach to AI development.

Furthermore, open source SLMs scale efficiently across multiple projects and organizations without the need for costly hardware upgrades. By deploying these models within existing IT infrastructure, you can create customized AI solutions without compromising performance or exceeding budget constraints.

The cost savings extend beyond infrastructure. Open source SLMs also eliminate the licensing fees associated with proprietary models, offering cost-effective access to advanced gen AI capabilities without vendor-imposed restrictions or limitations.

Innovate with open source SLMs from Red Hat and IBM

The Granite family of open source gen AI models—developed by IBM and included with Red Hat® Enterprise Linux® AI—addresses the specific demands of enterprise AI applications.

Learn more about open source gen AI models

Read the Maximize AI innovation with open source models e-book to find out more about task-specific SLMs and open source gen AI solutions.

About Red Hat

Red Hat helps customers standardize across environments, develop cloud-native applications, and integrate, automate, secure, and manage complex environments with award-winning support, training, and consulting services.

North America	Europe, Middle East, and Africa	Asia Pacific	Latin America
1 888 REDHAT1 www.redhat.com	00800 7334 2835 europe@redhat.com	+65 6490 4200 apac@redhat.com	+54 11 4329 7300 info-latam@redhat.com

facebook.com/redhatinc
twitter.com/RedHat
linkedin.com/company/red-hat

redhat.com
0225_KVM