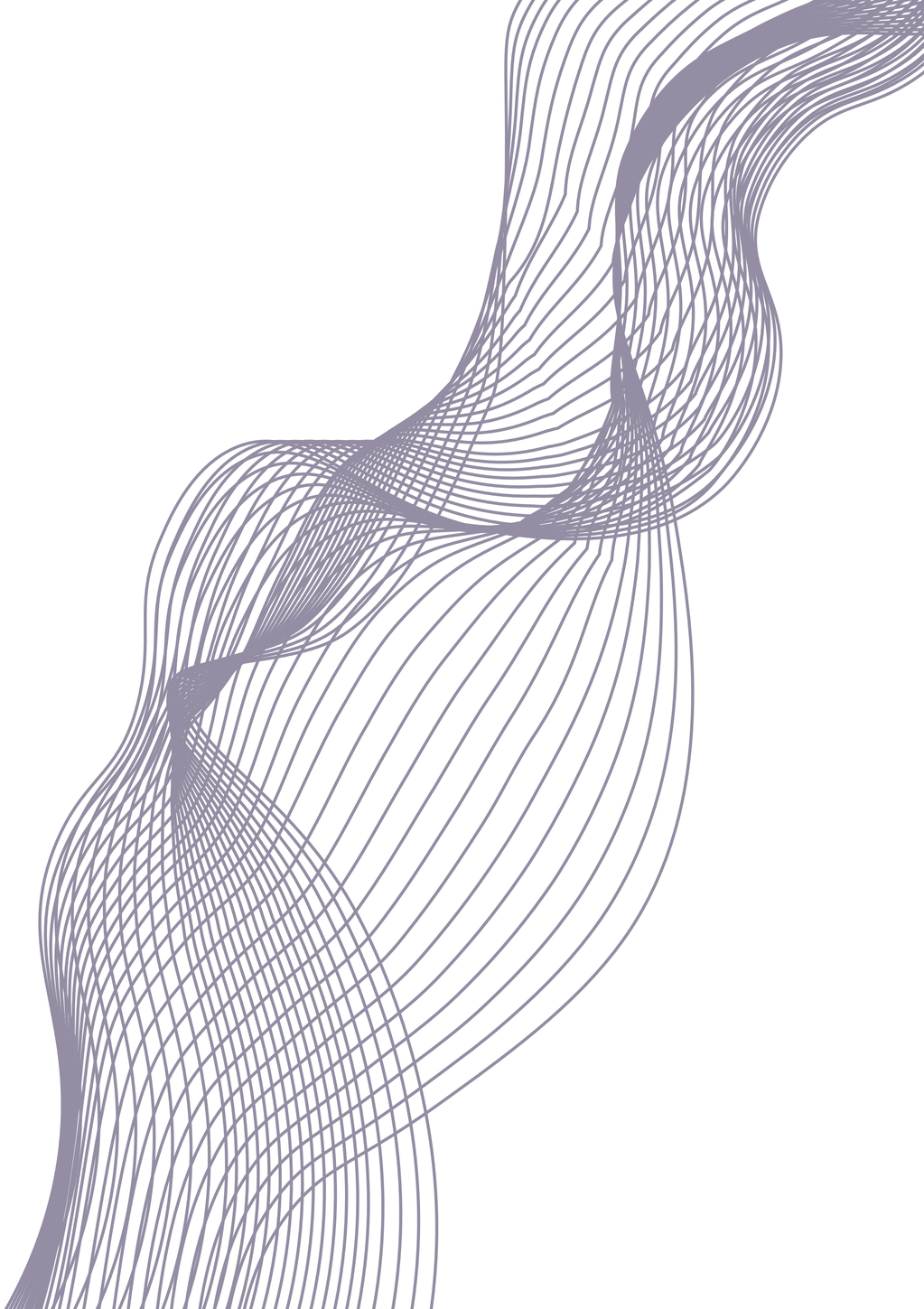
INVISIBLE | aws

Challenges & Opportunities: Small Language Models

The top challenges and insights from tech leaders developing AI applications across industries.



LONG TERM PARTNERSHIP

Foreword

In 2024, the demand for efficient, scalable AI solutions intensified, and Small Language Models (SLMs) emerged as a powerful approach for balancing accuracy and cost-effectiveness in business applications.

In partnership with AWS, Invisible co-hosted technical workshops that brought together industry leaders to explore strategies for fine-tuning and deploying cutting-edge SLMs. Through hands-on sessions with best practices, participants uncovered actionable ways to enhance language model performance while managing costs—a critical balance in today's competitive landscape.

This report captures core insights from these workshops, uncovering shared challenges and knowledge gaps among tech leaders across roles and organizational maturity. The findings in this report enable leaders to benchmark their progress, assess their AI readiness, and understand common hurdles in deploying SLMs.

> For the longest time, we thought the best way to make a model better was to add more parameters and add more data. But we can't always afford the resources necessary to adopt that strategy in perpetuity. Using a **small language model** is one way that you can get the same results for less cost. And using training data with precision can make measurable, meaningful impact for far less.

Lydia Andresen
Director of Data Strategy at Invisible Technologies

Agenda

Emerging Challenges with LLMs

In response to the rise of AI hype, businesses adopted large language models as the go-to solution for use cases across domains, but emerging challenges are tempering optimism.

Gartner forecasts that by 2028, over 50% of enterprises building large AI models from scratch will abandon these efforts due to costs, technical debt, and scaling difficulties [1].

High Costs

LLMs require significant computing power, leading to high expenses for storage, energy, and infrastructure. Executives predict that cost will become an increasingly key factor in decision-making about Generative AI [2].

Slow Processing Speeds

Due to their complexity and size, LLMs struggle with slow processing speeds, especially when handling real-time queries or large datasets.

Inaccuracy and Hallucinations

LLMs are prone to frequent inaccuracies, commonly referred to as "hallucinations." They generate outputs that are confident yet factually incorrect.

The Rise of Small Language Models (SLMs)

In 2024, many enterprise AI teams started to explore options with SLMs. SLMs can be used for domain specific-tasks and offer a more efficient and cost effective solution compared to general purpose LLMs.

Lower Costs

SLMs are significantly more cost-effective as they require fewer computational resources and storage. This makes them an affordable option for businesses, reducing both initial deployment and ongoing maintenance expenses.

Better Performance

SLMs are often more efficient at handling specific queries, offering users more direct and relevant solutions. By avoiding the "everything to everyone" approach of large models, SLMs deliver responses that are better aligned with business needs.

Greater Precision

Since SLMs are trained on smaller, curated, high-quality business datasets, they achieve higher accuracy in specialized applications. Their size allows them to adapt quickly to the nuances of a given task, making them more precise in handling targeted functions.

How Boosted.ai Launched a Better, Faster AI Investment Assistant and Cut Costs by 90%

Boosted.ai, an AI platform for financial analysis, built its portfolio management tool, Boosted Insights, which processed data from 150,000 sources to serve over 180 global asset managers, on a general-purpose LLM. By mid-2023, rising costs, rate limits, and challenges with real-time scaling on their LLM led Boosted.ai to explore a more targeted, cost-effective solution with AWS.

In 2024, Boosted.ai worked with AWS and Invisible to **migrate and fine tune an SLM** on the AWS platform. The results included:

- **Reduced costs by 90%** without sacrificing quality
- Overnight to near **real-time updates**
- **Improved security and personalization** with the ability to run a model in a customer's private cloud
- **Improved accuracy and relevancy** of financial insights for customers through human-in-the-loop training

Read the full story *here*.

INVISIBLE | Boosted.ai | aws

TECHNICAL WORKSHOPS

AWS + Invisible Technical Workshops

In 2024, AWS partnered with Invisible to bridge the gap in understanding the possibilities of small models for domain specific-tasks.

We hosted workshops for tech leaders with AI experts from AWS, data strategy experts from Invisible, and other leading innovators who have deployed custom SLMs with great success.

- July 22, Palo Alto
- August 6, NYC
- September 4, San Francisco

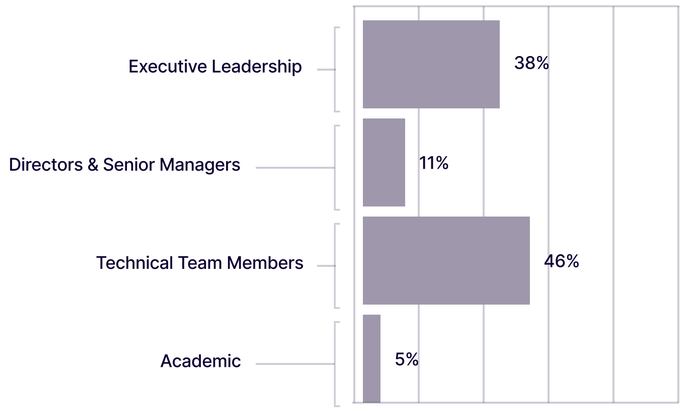
During the workshops, **we spoke to 500+ tech leaders and gathered written feedback from over 100 participants**, gaining insight into their current stage in AI adoption and the challenges they face in model development. This report highlights the key insights we uncovered.

AUDIENCE SNAPSHOT

Tech Leaders Across Roles and Org Sizes Were Curious about SLMs

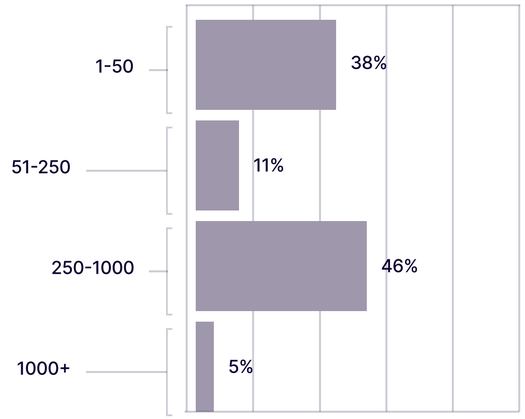
We had a diverse range of tech leaders attend the workshop events, demonstrating similar needs and challenges across organizations regardless of their size.

Tech Leaders & Seniority Levels:



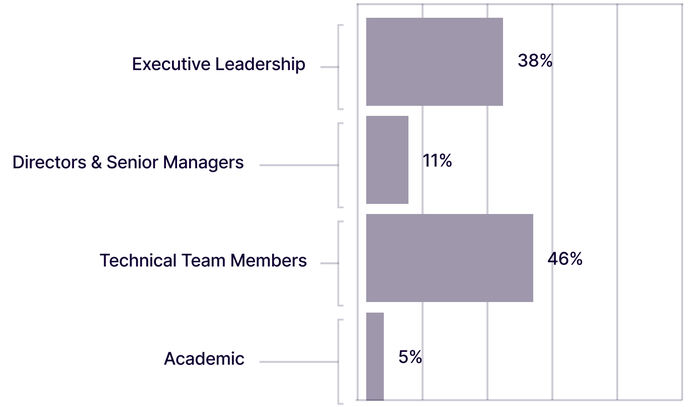
~ 49% were executives, directors, or senior managers

Differences in Company Sizes [employees]:



~ 32% were mid-market & enterprise 250+ employees

Tech Leaders & Seniority Levels:



~ 52% fell into a mix of early to growth-stage startups

AUDIENCE SNAPSHOT

Most Tech Leaders Have Active Generative AI Use Cases

When asked where they were in their fine-tuning journey, most tech leaders said they were actively working on generative AI use cases and were close to production — **yet continue to run into issues with AI deployment**.

Here are some of the issues:

~78%
have built a gen AI prototype

~54%
have tried SLMs

~54%
were close to production

AUDIENCE SNAPSHOT

Few Have Invested in External Expertise for Data Preparation or Training Data

While many companies have active AI initiatives, most **lack experience** in data preparation or developing training datasets, either in-house or with external partners.

~56%

of companies have never done data prep or developed training data internally

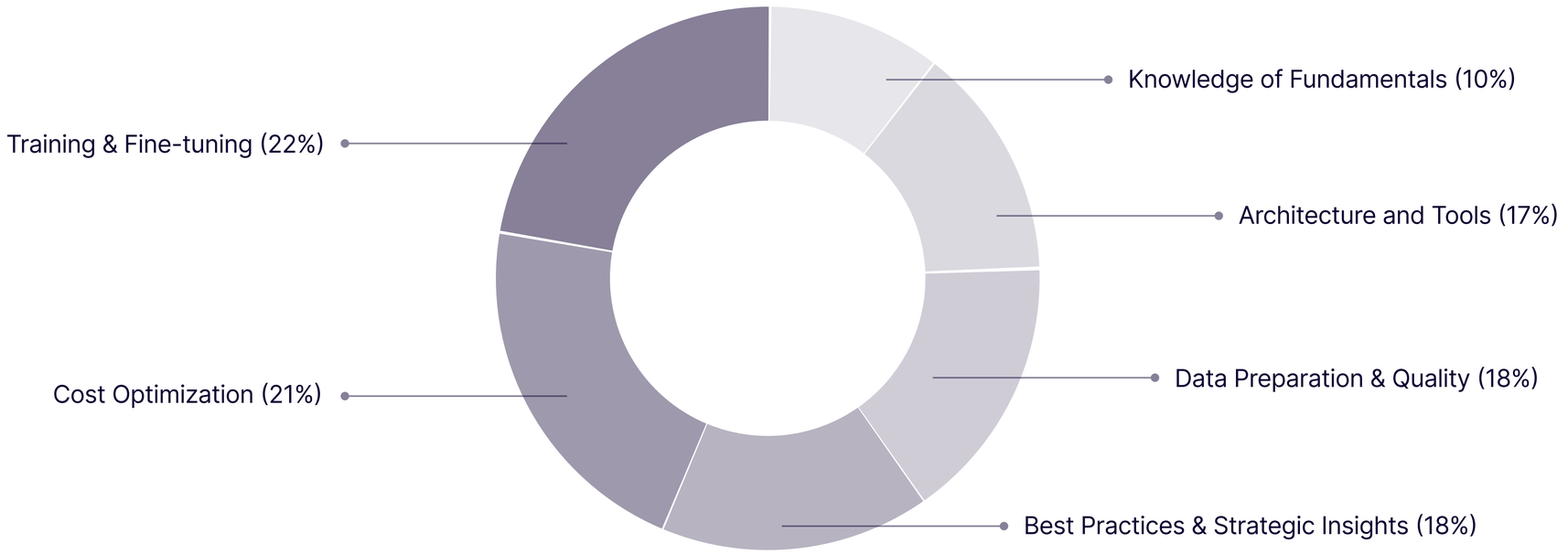
~56%

of companies have never done data prep or developed training data internally

TOP CHALLENGES

Most tech leaders struggled with similar issues across their AI projects

We asked participants what their top challenges were and in what areas they needed more resources and support:



#1 Challenge: Model Training & Fine Tuning

“How many samples are typically needed for “decent” fine tuning?”

Learn Fine Tuning Techniques

Tech leaders want to understand various fine-tuning techniques, including:

- Fine tuning with their own data
- How to iterate based on quantitative feedback
- How to fine tune a model without using production data
- Specific techniques for enterprise AI

RAG Vs. SLM

Teams want to know the differences and tradeoffs of building a model using RAG as opposed to fine tuning an SLM.

Effectively Validate Models

Teams seek to learn most the efficient mechanisms for validating SLMs for a use case.

Prevent Hallucinations

Leaders want to know how to keep their models from hallucinating, and if using SLMs can help prevent or reduce these instances.

#2 Challenge: Budget Constraints

“How much would it cost to deploy a simple AI app by training 1 million data points?”

Understand Costs

Tech leaders seek a clearer understanding of deployment costs, requesting concrete examples to guide their decision-making.

Keep Costs Low

Leaders are struggling to find effective strategies to minimize expenses. They want to learn more about price optimization.

Apply For Credits

Early-stage startups face challenges navigating credit application processes, which slows down progress in their AI initiatives.

Choose Infrastructure

Teams are evaluating the trade-offs between using AWS, other cloud providers, and managing in-house infrastructure to find the most cost-effective option.

#3 Challenge: Understanding SLMs

“I want to understand whether SLMs are worth it and if I should invest my time and money into it.”

Use Cases & Best Practices

Learn about successful use cases and get best practices

Compare SLM to LLM

Understand which kinds of use cases are better for SLMs over LLMs

Understand SLM Benefits, Trade-offs, & Lifecycle

Explore the benefits of using SLMs & learn about the biggest bottle-necks, while understanding the lifecycle of an SLM use case, from design to deployment

Learn Market Trends

Learn the average default tendency for enterprise teams looking to deploy SLMs

Find Quality Resources

Learn how to better source subject matter experts for their use cases

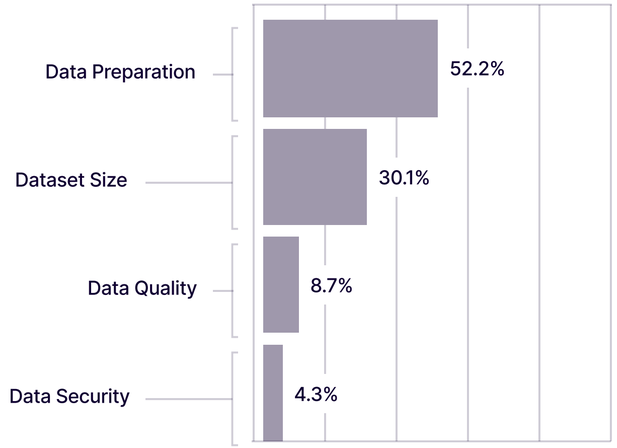
#4 Challenge: Data Preparation & Quality

“How do I know my fine tuning data is sufficient and of good quality?”

Tech leaders seek guidance on data preparation and creating high quality training data for fine-tuning SLMs. Key concerns include ensuring data quality, determining optimal dataset sizes, and preparing for multi-language support.

They also want to learn effective strategies for aligning proof-of-concept architectures with production needs while securely and efficiently deploying SLMs.

Data preparation areas where companies face the most challenges:



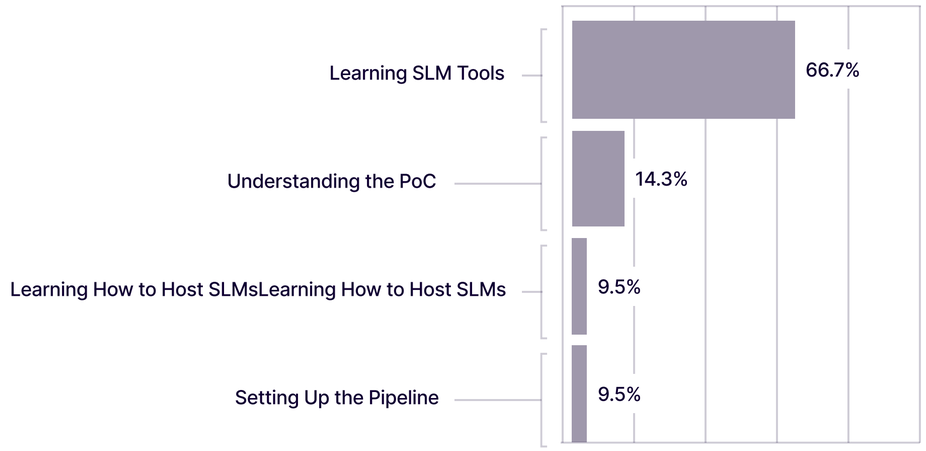
#5 Challenge: Understanding the Tools on the Market

“Learn about the AWS tools for deploying SLMs and routing to a different hierarchy of models.”

Tech leaders want to learn how to host SLMs, focusing on deployment and model versioning.

There is a strong interest in tools for fine-tuning and data preparation, along with guidance on designing proof of concept architectures that align with production needs. The focus is on quickly building prototypes and setting up efficient processes to make everything run smoothly.

SLM architecture & tooling areas where companies face the most challenges:

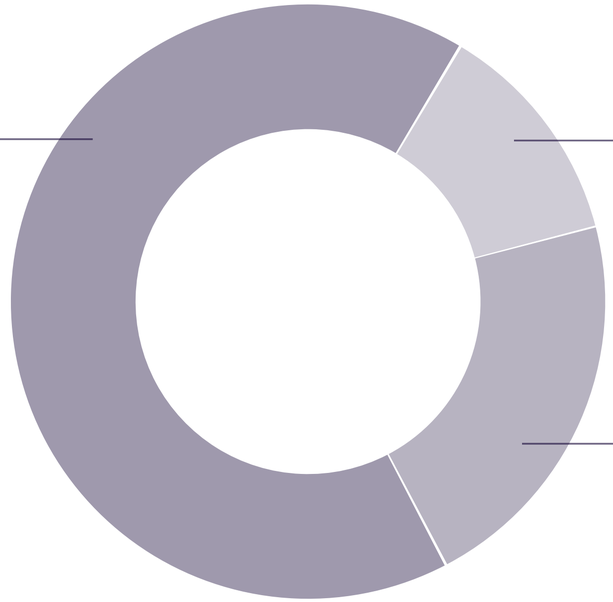


#6 Challenge: Developing Foundational Skills with SLMs

“What are the necessary things to learn to make a working model?”

Understanding SLMs (76.9%)

Tech leaders aim to grasp the essential concepts and skills needed to effectively apply SLMs in their projects. They want to gain insights to create small prototypes and develop hands-on coding skills.



Staying Up-to-Date (15.4%)

They want to keep up with the latest developments.

Selecting Tutorials (7.7%)

Leaders want to understand how to select the right tutorials to start with when there are a lot of vendors.

CONCLUSIONS

Top Takeaways

1 **Model Training & Fine Tuning**
Many tech leaders struggle with fine tuning models, seeking better techniques, validation methods, and solutions for hallucinations while minimizing technical debt.

2 **Budget Contraints**
Many leaders are concerned with finding cost-effective solutions while developing and implementing AI strategies. They seek ways to optimize budgets, allocate resources efficiently, and ensure that investments in AI yield significant returns.

3 **Understanding SLMs**
Tech leaders seek a broader understanding of SLMs, including best practices, SLM vs. LLM comparisons, SLM benefits, and the lifecycle from design to deployment.

4 **Data Preparation & Quality**
Companies struggle with data preparation for fine-tuning SLMs, focusing on quality, dataset sizing, and multi-language support, while seeking secure and efficient deployment strategies.

5 **Understanding The Tools On The Market**
Tech leaders aim to understand SLM tools and architectures, focusing on deployment, model hosting, and creating proof-of-concept designs aligned with production. They prioritize building prototypes and setting up efficient pipelines for smooth operations.

6 **Developing Foundational Skills With SLMs**
Tech leaders seek technical knowledge on applying SLMs, focusing on building prototypes, coding, staying current, and selecting effective tutorials.

CONCLUSIONS

Final Thoughts

This report highlights critical trends and challenges among tech leaders across roles and organizational maturity who are attempting to deploy language-based AI applications. Data emerged as a central theme, as teams face significant hurdles in preparing and leveraging data effectively—often blocking innovation and scalability. Knowledge gaps in areas like best practices, architecture and tooling, and cost management exacerbate these challenges.

Building resilient infrastructures and ensuring compliance remain priorities, but success increasingly depends on addressing data quality issues and fostering operational agility. Strategic partnerships with expert teams like AWS and Invisible are essential to overcoming these barriers, enabling businesses to unlock data's full potential and achieve sustainable growth in a competitive landscape.

> GenAI has huge potential, but many projects struggle to succeed because it's tough to tackle issues like hallucinations, high costs, and data privacy all at once. Through case studies with clients like Boosted.ai, we **proved that fine-tuning small language models for domain-specific tasks can reduce costs by 90%** while improving accuracy and privacy."
>
> Deepam Mishra
> *Senior Advisor AI/ML Startups, AWS*

CONCLUSIONS

About Invisible

Invisible Technologies is redefining operations in the AI era. Our pioneering AI process platform seamlessly fuses cutting-edge AI with elite global human expertise to transform complex processes and deliver rapid results at scale.

Invisible is trusted to train **80% of the world's leading AI models**, including those built by OpenAI, Apple, Microsoft, and Cohere. We have the native AI skills to tackle any enterprise problem, from automating cumbersome KYC/AML processes to digitizing every menu on DoorDash within days.

Pioneering New Frontiers in AI Training:

Our expertise in AI comes from partnering with world-leading AI firms like OpenAI, Amazon, Meta, AI21 Labs and Cohere, developing cutting-edge AI training methods and advanced solutions for model improvement.

Our fully managed expert teams, specialized tools, and continuous collaboration with AI researchers makes Invisible the trusted partner for AI innovators.

Learn more at www.invisible.co

About AWS

Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud, offering over 200 fully featured services from global data centers. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—are using AWS to lower costs, become more agile, and innovate faster.

Learn more at aws.amazon.com

References & Authors

References

[1] Gartner. Take This View to Assess ROI for Generative AI. Read here.

[2] Deloitte. Now decides next: Moving from potential to performance. Read here.

[3] Salesforce. Tiny Titans: How Small Language Models Outperform LLMs for Less. Read here.

Authors

Korina Skhinas, Lydia Andresen, Deepam Mishra