

Domain-Specific SLMs: A Strategic Alternative to Fine-Tuning Large Language Models

[Assem Hijazi – Experience, Knowledge & Ideas](#)

Introduction: Should Organizations Create a Private Small Language Model (SLM)?

As organizations explore Generative AI, the idea of creating a private Small Language Model (SLM) has gained traction. However, this approach requires careful consideration due to high costs and significant challenges. Unlike Large Language Models (LLMs) like ChatGPT, SLMs are not suitable for general-purpose use and should instead focus on domain-specific or subdomain-specific applications.

This article discusses the key factors, challenges, and considerations for creating an SLM, helping organizations determine whether this approach aligns with their needs.

Key Points to Consider for SLM Creation

1. The Idea of a Private SLM for Organizations:

- Private SLMs enable organizations to leverage AI while maintaining control over sensitive data.
- Creating an SLM is complex and costly, making it suitable only in specific scenarios.

2. Costs and Challenges:

- Building an SLM involves substantial expenses, including dataset acquisition, hiring experts, and infrastructure setup.
- Even an SLM, while smaller than an LLM, requires significant investment to achieve meaningful results.

3. Why Generic SLMs are Not Recommended:

- Generic SLMs cannot compete with general-purpose LLMs in handling diverse prompts and broad functionality.
- Creating a generic SLM is impractical and inefficient, as it lacks the scale and dataset diversity of LLMs.

4. The Focus of SLMs:

- SLMs are most effective when tailored to specific domains or subdomains.
- Examples of suitable use cases include:
 1. Mission-critical organizations (e.g., military, police).
 2. Financial institutions (e.g., banks seeking custom fintech solutions).
 3. Organizations handling sensitive, proprietary data.

5. Key Questions for Organizations to Evaluate:

- What are the goals and purpose of the SLM?
- How does the SLM align with the organization's broader AI strategy?
- What architectural considerations and data privacy measures are required?
- What are the cost implications compared to fine-tuning an LLM?

6. The Role of an AI Center of Excellence:

- **An AI Center of Excellence can guide strategic decisions, ensuring alignment with organizational objectives.**
- **Such a framework can evaluate whether fine-tuning an LLM or creating an SLM is the optimal approach. “we will cover this in future articles in different series of AI Strategy for organizations”**

7. Important to consider:

- **The decision to create an SLM or fine-tune an LLM is not one-size-fits-all.**
- **Organizations must evaluate their goals, resources, and data privacy needs to ensure the chosen approach aligns with their long-term strategy.**

The main Considerations for Creating or Adopting an SLM

1. Cost Implications of Building an SLM from Scratch

- **Dataset Acquisition Challenges:**
 - Acquiring high-quality, domain-specific datasets is expensive and time-intensive.
 - Requires partnerships with data providers or significant in-house data generation efforts.
 - **Hiring Expertise:**
 - Building an SLM requires specialized AI/ML engineers, domain experts, and data scientists.
 - Costs for assembling and maintaining such a team can exceed initial estimates.
 - **High Initial Costs:**
 - Training infrastructure (GPUs/TPUs), algorithm development, and ongoing iteration make building an SLM from scratch significantly costlier than leveraging existing solutions.
 - **Recommendation:**
 - Building an SLM from scratch should only be considered when there's a highly specific scenario that cannot be met by existing models.
-

2. Domain and Subdomain Specificity

- **Narrow Focus is Essential:**
 - Building an effective SLM requires targeting specific domains or even subdomains.
 - General-purpose SLMs are inefficient and dilute the advantages of specialization.
 - **Requirement for Domain Expertise:**
 - Deciding the domain or subdomain and curating relevant datasets require collaboration with subject matter experts (SMEs).
 - SMEs must also help validate the performance of the SLM in practical use cases.
 - **Impact on Costs and Feasibility:**
 - Without clarity on the domain, development costs and timelines can spiral out of control.
 - Misaligned datasets or domain definitions can render the SLM ineffective.
-

3. Challenges with Fine-Tuning on LLMs

- **Transparency Issues:**
 - Companies cannot easily monitor or control how fine-tuned knowledge interacts with the LLM's base parameters.
 - Providers often claim no access to fine-tuning data, but there's little visibility into what happens during fine-tuning or model updates.

- **Knowledge Exchange Risks:**
 - While fine-tuning claims to isolate proprietary data, organizations can't fully audit what the major LLM "learns" from the fine-tuning.
 - **Fine-Tuning Suitability:**
 - While fine-tuning works well for some scenarios, it may not align with organizational requirements for privacy, control, or specificity.
-

4. Importance of Starting with Pre-Existing SLMs

- **Leverage Existing Models:**
 - Instead of building from scratch, organizations can start with pre-existing SLMs designed for specific domains (e.g., financial services, healthcare).
 - These models often come with curated datasets and pre-tuned capabilities.
 - **Transparent Datasets:**
 - The transparency of datasets used to train pre-existing SLMs is critical.
 - Organizations must have access to and control over the training datasets to ensure alignment with their use cases and data governance policies.
 - **Customization Potential:**
 - Ready-made SLMs should allow easy integration of organizational datasets to fine-tune or expand capabilities without retraining the entire model.
-

5. Compatibility of Algorithms

- **Generic vs. Domain-Specific Algorithms:**
 - Most pre-trained SLMs rely on generic algorithms optimized for broad applications.
 - Organizations may require specific algorithms tailored to niche requirements or compliance needs.
 - **Flexibility for Algorithmic Changes:**
 - The SLM must support adding or modifying algorithms, including integrating public or custom algorithms for domain-specific tasks.
 - **Technical Expertise Required:**
 - Modifying algorithms in pre-existing SLMs requires skilled developers who understand both the model's architecture and the targeted domain.
-

6. Integration of Organizational Datasets

- **Ease of Adding Data:**

- The ease with which organizational datasets can be added to the SLM is a critical factor for success.
- Organizations should prioritize models that offer intuitive interfaces for integrating proprietary datasets.
- **Data and Model Alignment:**
 - Not all datasets are inherently compatible with the existing model’s training data or algorithms.
 - Mismatched datasets can degrade model performance or require additional preprocessing efforts.

Key Considerations in Decision-Making

Factor	SLM Approach	Fine-Tuning on LLM
Transparency of Datasets	Full control and visibility over training datasets.	Limited or no access to the datasets or training process of the LLM.
Algorithm Customization	Allows organizations to add or modify algorithms for specific needs.	Fine-tuning often relies on the LLM’s existing algorithms, with limited flexibility for changes.
Domain-Specificity	Targets narrow domains or subdomains with focused datasets and algorithms.	General-purpose fine-tuning is broader but lacks domain-specific precision.
Privacy and Security	Sensitive data remains localized, ensuring compliance and control.	Data must be shared externally for fine-tuning, increasing risks.
Initial Cost	Higher for building an SLM from scratch unless leveraging pre-existing domain-specific models.	Lower upfront cost as the LLM is already built; costs arise primarily from fine-tuning efforts.
Integration Ease	Depends on the architecture of the pre-existing SLM.	Easier integration but limited transparency over internal processes.

Recommendations

1. **Start with Pre-Existing SLMs:** Choose a pre-existing SLM tailored to your target domain or subdomain. This minimizes costs and accelerates deployment.
2. **Ensure Dataset Transparency:** Verify the transparency and accessibility of training datasets for any chosen SLM.
3. **Evaluate Algorithm Flexibility:** Confirm that the SLM supports customization of algorithms for niche requirements.
4. **Plan Data Integration:** Ensure your datasets can be easily integrated without extensive preprocessing or compatibility issues.
5. **Consider Costs vs. Value:** Assess whether building an SLM aligns with long-term ROI and privacy goals. Only opt for scratch-built SLMs in highly specific scenarios.

Ready-made Small Language Models (SLMs) available in the market

Note: Based on quick searching

Model Name	Description	Key Features
TinyLlama	A compact 1.1-billion-parameter model pretrained on ~1 trillion tokens.	- Based on Llama 2 architecture. - Uses FlashAttention for efficiency. - Excellent for downstream tasks.
SmolLM	A family of small models (135M, 360M, 1.7B parameters) designed for high performance on devices.	- High-quality dataset. - Optimized for fast deployment on local devices.
Phi-3-Mini	A 3.8-billion-parameter model optimized for reasoning and language understanding.	- Supports up to 128K tokens. - Matches performance of much larger models.
Cerebras-GPT	A family of compute-efficient models ranging from 111M to 13B parameters.	- Implements Chinchilla scaling laws. - Scalable and resource-efficient.
Mistral 7B	A 7-billion-parameter dense model designed for efficiency.	- High performance for NLP tasks. - Suitable for various applications.

This table highlights the features and resources for each SLM, making it easier for organizations to assess which model aligns with their specific needs. Let me know if you'd like to add more details or refine the table further!

Understanding the Costs of Building or Adopting a Small Language Model (SLM)

Estimating the costs of building or adopting an SLM provides a general sense of the financial commitment required. However, the figures presented here are based on quick searches and general market research, not detailed consultation with experts. Actual costs may vary significantly depending on factors such as domain specificity, dataset availability, and infrastructure requirements.

1. Dataset Acquisition and Curation Costs

- **Acquiring High-Quality Data:**
 - For domain-specific or subdomain-specific SLMs, obtaining relevant datasets is essential.
 - Costs can range from \$50,000 to \$200,000, depending on the domain, size, and exclusivity of the dataset.
 - **Data Cleaning and Preprocessing:**
 - Ensures that datasets are usable and high-quality.
 - Requires data engineers and domain experts, costing an additional \$10,000 to \$50,000.
 - **Synthetic Data Generation:**
 - In cases where existing data is insufficient, synthetic data generation might be necessary, adding \$50,000 to \$150,000.
-

2. Expert Hiring Costs

- **Domain Experts:**
 - Specialists are needed to define the domain or subdomain and validate the SLM's performance.
 - Salaries for subject matter experts (SMEs) range from \$80,000 to \$150,000 per year.
 - **AI/ML Engineers and Data Scientists:**
 - Experienced professionals to design, train, and maintain the model.
 - Salaries for AI/ML engineers can range from \$100,000 to \$200,000 annually.
-

3. Infrastructure and Training Costs

- **On-Premise Infrastructure:**
 - Setting up GPUs or TPUs for training can cost \$50,000 to \$200,000, depending on model size.
- **Cloud-Based Infrastructure:**
 - Using cloud services for training and hosting can cost \$10,000 to \$30,000 per month, depending on usage.

- **Training Costs:**
 - Training an SLM, even at a smaller scale (e.g., 1 billion parameters), requires significant compute power.
 - Training costs alone can range from \$20,000 to \$100,000, depending on the complexity of the model.
-

4. Fine-Tuning and Iteration Costs

- **Periodic Fine-Tuning:**
 - Adjustments to the model to incorporate new data or improve performance can cost \$20,000 to \$50,000 per iteration.
 - **Ongoing Updates and Monitoring:**
 - Maintenance and monitoring require dedicated resources, costing \$50,000 to \$100,000 annually.
-

5. Pre-Existing SLM Customization Costs

Using pre-existing SLMs reduces initial costs but still involves customization:

- **Model Licensing:**
 - Open-source models (e.g., Mistral, TinyLlama) may be free, but some require licensing fees, which can range from \$10,000 to \$100,000.
- **Data Integration:**
 - Integrating organizational data into pre-existing SLMs requires preprocessing and adaptation, costing \$10,000 to \$50,000.
- **Algorithm Customization:**
 - Modifying or adding domain-specific algorithms can cost \$20,000 to \$100,000, depending on complexity.

Total estimated cost

Scenario	Estimated Initial Cost	Estimated Annual Cost
Basic Domain-Specific SLM (Pre-Existing Model)	\$150,000–\$300,000	\$50,000–\$100,000
Custom SLM (Moderate Scale)	\$300,000–\$500,000	\$100,000–\$200,000
High-Specificity SLM (Built from Scratch)	\$500,000–\$1,000,000	\$200,000–\$400,000

Factors Influencing Costs

1. **Domain Complexity:**
 - Specialized domains like healthcare or defense increase data acquisition and expertise costs.
2. **Scale of the Model:**
 - Larger models (more parameters) increase training and deployment expenses.
3. **Customization Needs:**
 - Adding or modifying algorithms and datasets impacts development and maintenance costs.
4. **Infrastructure Choice:**
 - Cloud-based systems offer scalability but incur recurring costs, while on-premise setups require significant upfront investment.

Recommendations to Optimize Costs

1. **Leverage Pre-Existing SLMs:**
 - Start with open-source or commercially available SLMs to minimize initial investment.
2. **Focus on Domain-Specificity:**
 - Narrowing the scope to a specific domain or subdomain reduces dataset and algorithm customization costs.
3. **Prioritize Dataset Transparency:**
 - Use models with accessible and validated training datasets to avoid compatibility issues.
4. **Evaluate Long-Term ROI:**
 - Ensure that the investment aligns with the organization’s strategic goals and offers measurable benefits.