**The Rise of Small Language Models— Efficient & Customizable**

<https://medium.com/@bijit211987/the-rise-of-small-language-models-efficient-customizable-cb48ddee2aad>

**Summary**

The article discusses the emergence and growing significance of Small Language Models (SLMs) in the field of artificial intelligence, contrasting them with Large Language Models (LLMs) like OpenAI's GPT-3. While LLMs have demonstrated impressive capabilities, their massive size leads to drawbacks in efficiency, cost, and customizability. SLMs address these issues by providing a more efficient, cost-effective, and customizable alternative without significantly compromising performance.

**What Are Small Language Models?**

Small Language Models are AI systems trained on large text datasets but with significantly fewer parameters than LLMs. Typically, SLMs have fewer than 100 million parameters, sometimes even as few as 1 million. In contrast, models like GPT-3 contain over 100 billion parameters. Despite their smaller size, SLMs aim to perform similar tasks such as text generation, summarization, translation, and question-answering. The challenge lies in balancing model size with performance and flexibility.

**Motivations for Developing Small Language Models**

**Efficiency**

* **Faster Inference:** With fewer parameters, SLMs process inputs more quickly.
* **Lower Resource Requirements:** They need less memory and storage space.
* **Smaller Training Datasets:** SLMs can be effectively trained on less data, reducing the data acquisition burden.

**Cost**

* **Reduced Training Costs:** Training an LLM can cost millions of dollars, whereas SLMs can be trained economically on commodity hardware.
* **Affordable Deployment:** SLMs can run on lower-powered devices, making them suitable for edge computing and offline applications.

**Customizability**

* **Domain Specialization:** SLMs can be tailored to specific domains or tasks more easily than LLMs.
* **Quicker Iteration:** Smaller models allow for faster experimentation and refinement.
* **Adaptable Architectures:** They can be modified to suit niche tasks through various techniques:
  + **Pretraining:** Using domain-specific datasets to prime the model.
  + **Fine-tuning:** Continuing training on specialized end-task data.
  + **Prompt-based Learning:** Optimizing prompts for specific applications.
  + **Architecture Modifications:** Adjusting the model's structure to better suit certain tasks.

**How Small Can Useful Language Models Be?**

Research indicates that models with as few as 1–10 million parameters can demonstrate basic language competencies. An 8-million-parameter model achieved 59% accuracy on the GLUE benchmark, a standard for evaluating natural language understanding. Performance generally improves with increased parameters up to a point, but gains diminish after around 200–300 million parameters. This suggests a "sweet spot" where SLMs balance efficiency and capability effectively.

**Methods for Training Efficient Small Language Models**

**Transfer Learning**

* **Pretraining on Broad Datasets:** Models learn general language patterns before specializing.
* **Specialization Phases:** Refining the model for specific applications without increasing size.

**Self-Supervised Learning**

* **Predictive Tasks:** Models learn by predicting masked or corrupted text segments.
* **Effective for SLMs:** Encourages deeper generalization from limited data.

**Architecture Choices**

* **Efficient Transformers:** Streamlined models that focus on essential components like attention mechanisms.
* **Knowledge Distillation:** Training smaller models to mimic larger ones, retaining performance while reducing size.
* **Parameter Efficiency:** Selecting architectures that maximize performance per parameter.

**Example Applications Where Small Language Models Shine**

**Finance**

* **Transaction Classification:** Automating the categorization of financial transactions.
* **Sentiment Analysis:** Extracting opinions from earnings calls to inform trading strategies.
* **Data Standardization:** Converting unstructured financial data into standardized formats for analysis.

**Entertainment**

* **Content Generation:** Assisting in scriptwriting or prose creation to enhance productivity.
* **Dynamic Gaming Dialogue:** Creating interactive conversations in games tailored to player context.
* **Metadata Enrichment:** Analyzing content themes to improve recommendation systems.

These applications demonstrate how SLMs can deliver value in specialized contexts where efficiency and customization are crucial.

**Developer Frameworks Enabling Creation of Custom SLMs**

* **Hugging Face Hub:** Offers tools for hosting datasets, training models, and deploying them via APIs.
* **Anthropic Claude:** Provides an easy interface for training custom models with built-in safety features.
* **Cohere for AI:** Enables building small language models with as few as 1 million parameters, with options for client-hosted deployment.
* **Assembler:** Simplifies the development of specialized models through a user-friendly web interface.

These platforms lower the barrier to entry for businesses looking to develop and deploy custom SLMs.

**Emergence of Domain-Specific SLMs**

Organizations are increasingly developing proprietary SLMs tailored to their specific domains, such as finance, healthcare, and education.

**Advantages**

* **Superior Accuracy:** Training on niche datasets improves performance in specialized tasks.
* **Confidentiality:** In-house models prevent exposure of sensitive data, ensuring compliance.
* **Responsiveness:** Control over the development lifecycle allows for rapid iteration and alignment with organizational needs.
* **Cost Efficiency:** Long-term savings by reducing reliance on external models and services.

**Challenges**

* **Data Sufficiency:** Limited access to large, high-quality datasets can hinder model performance.
* **Model Governance:** Requires investment in monitoring, maintenance, and responsible AI practices.
* **Maintenance Costs:** Models need regular updates to address data drift and maintain reliability.

**Future Opportunities and Challenges**

**Opportunities**

* **Customization:** Targeting specialized industry gaps with tailored AI solutions.
* **Hybrid Intelligence:** Combining human expertise with AI to enhance productivity.
* **Advancements in Training Techniques:** Improved methods will expand SLM capabilities.
* **Faster Adoption:** Tangible commercial returns may accelerate the adoption of SLMs across industries.

**Challenges**

* **Responsible Implementation:** Ensuring ethical use, especially in sensitive areas.
* **Data Quality and Drift:** Maintaining model accuracy over time as data evolves.
* **Transparency:** Difficulty in understanding model decisions as they become more widespread.
* **Security Concerns:** Preventing misuse or exploitation of language models.

**Conclusion**

Small Language Models offer a promising avenue for making language AI more accessible, efficient, and customizable. They provide practical solutions tailored to specific applications, enabling businesses to leverage AI without the significant costs associated with LLMs. While challenges such as responsible deployment and maintenance exist, the potential benefits position SLMs to drive the next phase of AI innovation and productivity across various industries.

**Full Article**

**The Rise of Small Language Models**

Over the past few year, we have seen an explosion in artificial intelligence capabilities, much of which has been driven by advances in large language models (LLMs). Models like GPT-3, which contains 175 billion parameters, have shown the ability to generate human-like text, answer questions, summarize documents, and more. However, while the capabilities of LLMs are impressive, their massive size leads to downsides in efficiency, cost, and customizability. This has opened the door for an emerging class of models called Small Language Models (SLMs).

Let’s dive deep into the rise of Small Language Models:

* What are small language models?
* Motivations for developing small language models — Efficiency, Cost, Customizability
* How small can useful language models be?
* Methods for training efficient small language models
* Example applications where small language models shine
* Developer frameworks enabling creation of custom SLMs
* Future opportunities and challenges for the development and deployment of SLMs

By the end, you’ll understand the promise that small language models hold in bringing the power of language AI to more specialized domains in a customizable and economical manner.

**What Are Small Language Models?**

Language models are AI systems trained on large text datasets which enable capabilities like generating text, summarizing documents, translating between languages, and answering questions. Small language models fill much of the same niche but with notably smaller model sizes. But what constitutes a small language model?

Researchers typically consider language models under 100 million parameters to be relatively small, with some cutting off at even lower thresholds like 10 million or 1 million parameters. For comparison, models considered huge on today’s scale top over 100 billion parameters, like the aforementioned GPT-3 model from OpenAI.

The smaller model sizes allow small language models to be more efficient, economical, and customizable than their largest counterparts. However, they achieve lower overall capabilities since model capacity in language models has been shown to correlate with size. Determining optimal model size for real-world applications involves navigating the tradeoffs between flexibility & customizability and sheer model performance.

**Motivations for Small Language Models**

As mentioned above, small language models come with inherent advantages in efficiency, cost, and customizability over their larger counterparts. Let’s break down each of these motivations in more detail:

**Efficiency**

Due to having fewer parameters, small language models are significantly more computationally efficient than massive models like GPT-3 in several regards:

* They are faster in inference speed / throughput since fewer parameters need executing per input
* They require less memory and storage space also owing to the overall smaller model size
* Smaller datasets suffice for training small language models. As model capacity grows, the appetite for data grows with it.

These efficiency advantages directly cascade into cost savings:

**Cost**

Large language models require substantial computational resources to train and deploy. It’s estimated that developing GPT-3 cost OpenAI somewhere in the tens of millions of dollars accounting for hardware and engineering costs. Many of today’s publicly available large language models are not yet profitable to run due to their resource requirements.

Meanwhile, small language models can readily be trained, deployed, and run on commodity hardware available to many businesses without breaking the bank. Their reasonable resource requirements open up applications in edge computing where they can run offline on lower-powered devices. Overall there’s greater potential to find profitable applications of small language models in the short-term.

**Customizability**

A key advantage that small language models maintain over their largest counterparts is customizability. While models like GPT-3 demonstrate strong versatility across many tasks, their capabilities still represent a compromise solution that balances performance across domains.

Small language models on the other hand can readily be adapted to more narrow domains and specialized applications. With quicker iteration cycles, small language models make it practical to experiment with tailoring models to particular types of data through approaches like:

* Pretraining — priming small models on domain-specific datasets
* Fine-tuning — continuing training to optimize for end-task data
* Prompt-based learning — optimizing model prompts for specialized applications
* Architecture modifications — adapting model structure for niche tasks

These sorts of customization processes become increasingly arduous for large models. Combined with their accessibility, small language models provide a codex that developers can mold to their particular needs.

**How Small Can Useful Language Models Be?**

Given the motivations to minimize model size covered above, a natural question arises — how far can we shrink down language models while still maintaining compelling capabilities? Recent research has continued probing the lower bounds of model scale required to complete different language tasks.

Many investigations have found that modern training methods can impart basic language competencies in models with just 1–10 million parameters. For example, an 8 million parameter model released in 2023 attained 59% accuracy on the established GLUE natural language understanding benchmark.

Performance continues rising as model capacity grows. A 2023 study found that across a variety of domains from reasoning to translation, useful capability thresholds for different tasks were consistently passed once language models hit about 60 million parameters. However, returns diminished after the 200–300 million parameter scale — adding additional capacity only led to incremental performance gains.

These findings suggest even mid-sized language models hit reasonable competence across many language processing applications provided they are exposed to enough of the right training data. Performance then reaches a plateau where the vast bulk of compute and data seemingly provides little additional value. The sweet spot for commercially deployable small language models likely rests around this plateau zone balancing wide ability with lean efficiency.

Of course, specialized small language models tuned deeply rather than broadly may require much less capacity to excel at niche tasks. We’ll cover some of those applied use cases later on. But first, let’s overview popular techniques for effectively training compact yet capable small language models.

**Training Methods for Efficient Small Language Models**

The active progress training increasingly proficient small language models relies on methods that augment data efficiency and model utilization during the learning process. These techniques end up imparting more capability per parameter relative to naive training of larger models. We’ll break down some of the popular approaches here:

**Transfer Learning**

Most modern language model training leverages some form of transfer learning where models bootstrap capability by first training on broad datasets before specializing to a narrow target domain. The initial pretraining phase exposes models to wide-ranging language examples useful for learning general linguistic rules and patterns.

Small language models can capture much of this broad competency during pretraining despite having limited parameter budgets. Specialization phases then afford refinement towards specific applications without needing to expand model scale. Overall, transfer learning greatly improves data efficiency in training small language models.

**Self-Supervised Learning**

Transfer learning training often utilizes self-supervised objectives where models develop foundational language skills by predicting masked or corrupted portions of input text sequences. These self-supervised prediction tasks serve as pretraining for downstream applications.

Recent analysis has found that self-supervised learning appears particularly effective for imparting strong capabilities in small language models — more so than for larger models. By presenting language modelling as an interactive prediction challenge, self-supervised learning forces small models to deeply generalize from each data example shown rather than simply memorizing statistics passively. This engages fuller model capacity during training.

**Architecture Choices**

Not all neural network architectures are equivalently parameter-efficient for language tasks. Careful architecture selection focuses model capacity in areas shown to be critical for language modelling like attention mechanisms while stripping away less essential components.

For example, Efficient Transformers have become a popular small language model architecture employing various techniques like knowledge distillation during training to improve efficiency. Relative to baseline Transformer models, Efficient Transformers achieve similar language task performance with over 80% fewer parameters. Effective architecture decisions amplify the ability companies can extract from small language models of limited scale.

The techniques above have powered rapid progress, but there remain many open questions around how to most effectively train small language models. Identifying the best combinations of model scale, network design, and learning approaches to satisfy project needs will continue keeping researchers and engineers occupied as small language models spread to new domains. Next we’ll highlight some of those applied use cases starting to adopt small language models and customized AI.

**Example Applications Where Small Language Models Shine**

While excitement around AI often focuses on massive models grabbing headlines, an array of companies have already found utility by deploying small language models customized to their specific needs. I’ll highlight some representative examples like finance and entertainment domains where compact, specialized models are creating business value:

Finance

Financial organizations generate troves of numeric data and documents ripe for extracting insights using small, tailored language models. Use cases with strong return-on-investment include:

* Transaction classifiers automatically code invoice line-items with accounting categories to speed entry into bookkeeping systems.
* Sentiment models extract opinions from earnings call transcripts to develop trading signals by detecting management tone shifts.
* Custom entities aid in systematizing unstructured bank statements into standardized data reporting business revenue for lending risk analysis.

These applications translate language AI into direct process automation and improved analytics within established financial workflows — accelerating profitable models rather than speculating on technology promises alone. Risk management remains imperative in financial services, favoring narrowly-defined language models versus general intelligence.

Entertainment

Media, gaming, and related entertainment verticals constitute some of the most forward-leaning adopters of language AI-infused solutions as creative processes meld with advanced technology:

* Employing natural language generation, small language models automatically create first draft scripts or prose for animations that creators later refine, exponentially boosting individual productivity.
* In open world gaming, dialogue models produce dynamic conversation trees tailored to user context — expanding interactive freedom within virtual reality expanses.
* More capable language analysis enriches entertainment metadata, for instance identifying movie themes by patterns in subtitle content so recommendation engines better connect viewers to their unique interests.

Entertainment’s creative latitude provides an ideal testbed for exploring small language models generative frontiers. Though current applications still warrant oversight given model limitations, small language models efficiency grants developers ample space to probe creative potential.

The applications above highlight just a snippet of the use cases embracing small language models customized to focused needs. Our next section provides tools empowering many more such implementations.

**Developer Frameworks for Building Custom SLMs**

So with all the promise, how can developers actually start building specially-tailored small language models? Open source technologies have brought customized language AI within reach for enterprises across domains and scales. Full-service platforms below enable creating and deploying customized small language models cost-effectively:

🤗 Hugging Face Hub — Hugging Face provides a unified machine learning ops platform for hosting datasets, orchestrating model training pipelines, and efficient deployment for predictions via APIs or apps. Their Clara Train product specializes in state-of-the-art self-supervised learning for creating compact yet capable small language models.

Anthropic Claude — From the makers of ConstitutionalAI focused on model safety, Claude enables easily training custom classifiers, text generators, summarizers, and more with just a few lines of code. Built-in safety constraints and monitoring curb potential risks during deployment.

✨ Cohere for AI — Cohere offers a developer-friendly platform for building language models down to 1 million parameters drawing from their own training data or imported custom sets. Client-hosted options provide end-to-end privacy compliance.

Assembler — Assembler delivers tools for developing reader, writer, and classifier small language models specialized to niche data inputs. Their simple web interface masks infrastructure complexity for model creation and monitoring.

The services above exemplify the turnkey experience now realizable for companies ready to explore language AI’s possibilities. Expertise with machine learning itself is helpful but no longer a rigid prerequisite with the right partners. This brings more industries within reach to create value from AI specialization.

**Emergence of Domain-Specific SLMs**

Up to this point we have covered the general capabilities of small language models and how they confer advantages in efficiency, customization, and oversight compared to massive generalized LLMs. However, SLMs also shine for honing in on specialized use cases by training on niche datasets.

As large language models scale up, they become jacks-of-all-trades but masters of none. Their knowledge and performance gets diluted across domains. What’s more, exposing sensitive data to external LLMs poses security, compliance, and proprietary risks around data leakage or misuse.

These limitations motivate organizations across industries to develop their own small, domain-specific language models using internal data assets. Customization better serves their specific accuracy and security needs. We highlight leading examples next.

**Financial Small Language Models**

Financial corporations also deploy SLMs for needs around analyzing earnings statements, asset valuations, risk modeling and more. Domain fluency is mandatory yet sensitive data can’t leak externally.

For example, Softbank-owned Fortia built a custom SLM using client data to forecast currency exchange rates and arbitrage trading opportunities. Tight focus lets it outperform generalized LLMs and low latency enables automation. Data security was also imperative.

**Advantages of Specialized Domain SLMs**

What drives development of proprietary domain-specific SLMs across organizations and industries? Several key advantages stand out:

**Superior Accuracy:** Specialized model training on niche datasets with peculiarities simply not captured in general corpora yields dramatic accuracy gains over external LLMs. Augmenting models with domain adaptation techniques like weight imprinting, adapter modules and self-training provide further boosts.

**Confidentiality:** Relying on generalized external models forces exposing sensitive IP, financial, healthcare or other confidential data externally. But tightly scoped architectural boundaries of SLMs trained in-house mitigate risks of data leakage or misuse. This also provides compliance benefits.

**Responsiveness:** Organizations own the full model development lifecycle fine-tuning for precise alignment with specialized use cases like customer support flows. Direct control enables modifying and redeploying models in hours or days rather than month-long coordination with external LLM providers. Agility speeds iteration.

**Cost Efficiency:** Large language models carry not just hefty training costs but also per-query pricing for access that adds up. Building own models amortizes expenses long-term. Though some exorbitantly large datasets still benefit from pre-trained LLM foundations transfer learned into specialized heads.

**Challenges with Specialized SLMs**

Specialized SLMs do face adoption hurdles to balance against the advantages:

**Data Sufficiency:** Many organizations lack sizable structured datasets to train robust models from scratch. Alternative approaches like few-shot learning adapters atop foundation models help, but some data hungry applications still benefit from external general models. Though data leverage techniques like augmentation help.

**Model Governance:** Developing reliable performing SLMs requires investments into developer workflows, instrumentation, model ops and oversight well beyond many teams’ capabilities today. Responsible specialized AI still demands governance scaling expertise even for small models. Progress relies on DevOps maturation.

**Maintenance Costs:** Even compact models demand upkeep as data drifts. But monitoring burdens and rebuild requirements are much lighter for SLMs than LLMs. Still, treated as expendable, model rot can undermine reliability over time. Commitment to lifecycle management is key.

**Future Opportunities and Challenges for Small Language Models**

The efficiency, versatility and accessibility small language models introduce signifies just the start of a new wave of industrial AI adoption tailored to vertical needs rather than one-size-fits-all solutions. There remains enormous headroom for innovation as developers grasp the implications these new customizable codebases unlock.

However, practices and diligence around responsible implementation will be critical given language models’ inherent sociotechnical complexities even at small scale. Let’s conclude by briefly highlighting promising opportunities and key challenges ahead:

**Opportunities**:

* Customization enables targeting specialized gaps in industries underserved by general AI where language assistance can boost outcomes.
* Hybrid intelligence combinations with human domain experts will likely prove most constructive near-term, designing AI to enhance rather than replace jobs.
* Continued progress in efficient training techniques and multitask model architectures will further expand small language models’ capabilities.
* Adoption momentum of tailored language AI could surpass general adoption, as more targeted commercial returns become tangible sooner across verticals.

**Challenges:**

* Inappropriate reliance on AI for sensitive use cases risks excluding human expertise and oversight needed to make socially informed decisions.
* Data quality and concept drift issues compound quickly when applied models make unsupported inferences beyond their specialized training distributions.
* Transparency will grow more elusive as multiplying small models obfuscate why particular outputs were produced, especially when personalized per industry data.
* Malicious exploitation remains a concern with any proliferating technology, mandating safeguards against language models directly or indirectly causing harm.

With attentiveness to responsible development principles, small language models have potential to transform a great number of industries for the better in the years ahead. We’re just beginning to glimpse the possibilities as specialized AI comes within reach.

**Conclusion**

* What constitutes small language models and how they compare functionally to the largest models scaling today
* Motivations like efficiency, cost savings, and customizability driving adoption of small language models over generalized language AI
* The frontier of model miniaturization — how small can language models shrink while preserving compelling capabilities via modern training techniques?
* Real-world examples of companies putting specialized small language models to work across verticals like education, medicine, finance, and entertainment
* Developer resources enabling enterprises to move from AI interest to implementation with customized language models

The experiential technology of small language models distills broad excitement around language AI down to practical building blocks deliverable in the hands of commercial teams and users. Still an industry in its infancy, unlocking new applications harnesses both developer creativity and thoughtfulness on impacts as specialized models spread. But tailorable language intelligence now arriving on the scene appears poised to drive the next phase of AI productivity.