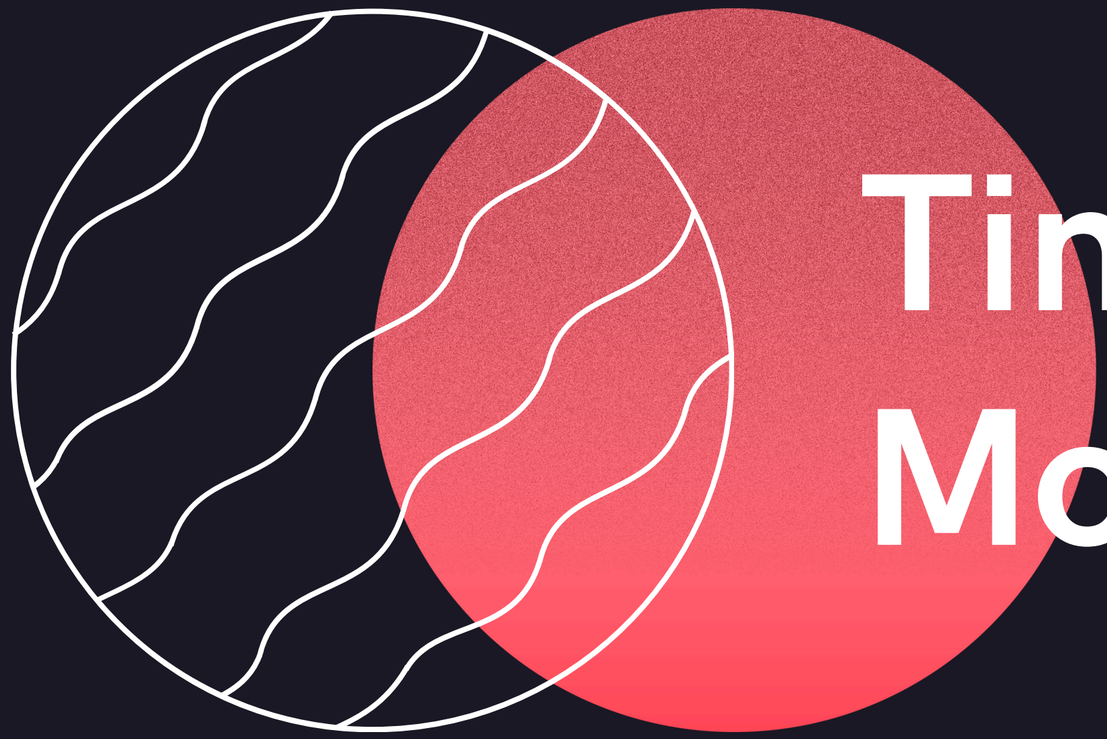


Tiny Language Models

Group 5:
Aditi Sonawane, Michelle Lie,
Rania Soetirto, Sherleen Lee

WHAT ARE TINY LANGUAGE MODELS?

Tiny language models, often referred to as "micro" or "small-scale" language models, are versions of language models that have been **optimized for deployment on resource-constrained devices**.

Compared to large language models (LLMs), they are **designed to perform well for simpler tasks** and is more accessible and easier to use for organizations with limited resources.

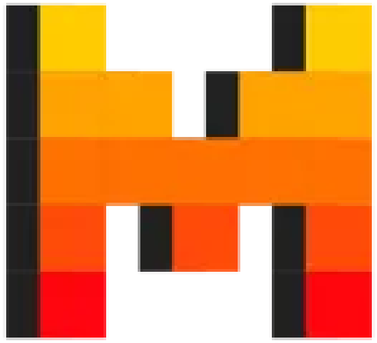
They can also be more easily fine-tuned to meet specific needs.

Source: Microsoft

"less is indeed more"

POPULAR TINY LANGUAGE MODELS

Mistral 7B



Microsoft Phi-2

Google’s Gemma

SMALL LANGUAGE MODELS

- **Size and Complexity:**
  - SLMs have fewer parameters and simpler architecture than LLMs, reducing computational demands.
- **Capabilities:**
  - They may struggle with context and complex linguistic tasks, impacting text coherence.
- **Training Data:**
  - Often trained on smaller, less diverse datasets, affecting performance and generalization.
- **Resource Requirements:**
  - SLMs need fewer computational resources - suitable for standard or low-power devices.
- **Applications:**
  - Primarily used for simpler language tasks like sentiment analysis and basic chatbots.

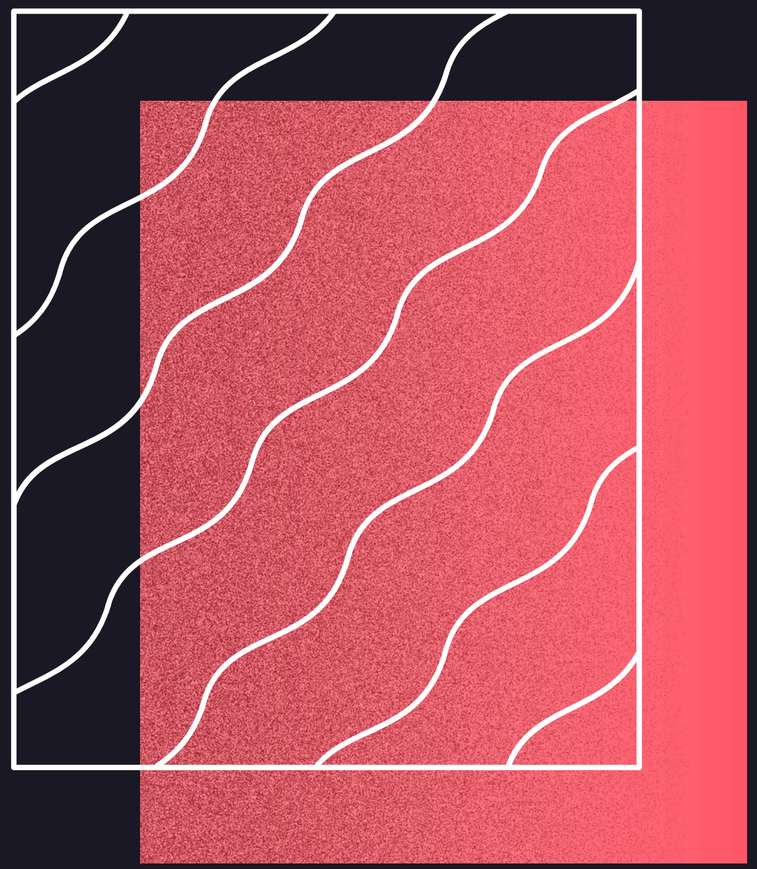
VS

LARGE LANGUAGE MODELS

- **Size and Complexity:**
  - Large models have millions or billions of parameters and complex architecture, requiring significant computational resources.
- **Capabilities:**
  - They excel at understanding context and generating coherent responses, leveraging vast and diverse training data.
- **Training Data:**
  - Trained on extensive textual data from the internet, capturing a wide range of language patterns and nuances.
- **Resource Requirements:**
  - Demand high computational resources, often relying on specialized hardware like GPUs or TPUs.
- **Applications:**
  - Power advanced NLP tasks such as machine translation, text summarization, and question answering.

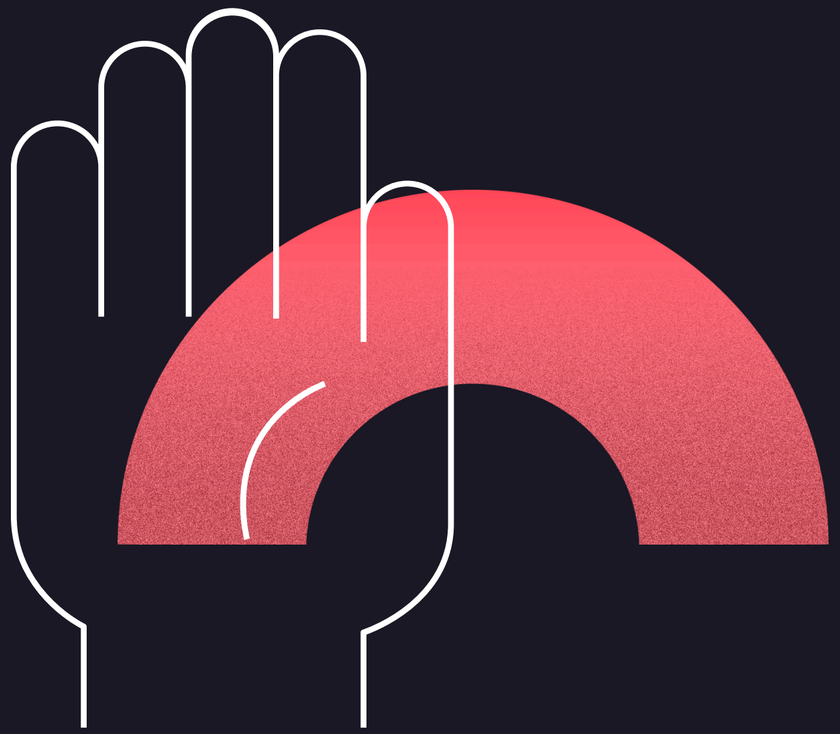
EMERGENCE OF DOMAIN-SPECIFIC SLM

FINANCIAL SMALL LANGUAGE MODELS	• Analyze earnings statements, asset valuations, risk modeling, and more • Softbank-owned Fortia built a custom SLM using client data to forecast currency exchange rates and arbitrage trading opportunities.
LEGAL SMALL LANGUAGE MODELS	• Automating document review, contract analysis and providing legal advice • JPMorgan Chase uses an AI program, COIN, to interpret commercial loan agreements, which has drastically reduced the number of man-hours needed for document review.
ADVANTAGES	• Superior Accuracy • Confidentiality • Responsiveness • Cost Efficiency
CHALLENGES	• Data Sufficiency • Model Governance • Maintenance Costs



ADVANTAGES OF TINY LANGUAGE MODELS

- Efficiency
  - faster in inference speed/throughput - fewer parameters
  - Less memory and storage - smaller model size
  - Smaller dataset
- Cost
  - LLM requires substantial computational resources
  - SLM can readily be trained, deployed, and run on commodity hardware
- Customizability
  - Can be more easily fine-tuned to meet specific needs
  - Shift to a portfolio of models where customers get to decide on what is best for their model scenario



LIMITATIONS OF TINY LANGUAGE MODELS

- Not designed for in-depth knowledge retrieval
  - LLMs excel due to their greater capacity and training using much larger data sets
  - LLMs are better at complex reasoning over large amounts of information due to their size and processing power
- Security and data theft concerns
  - SLM codes are open-source, which is especially risky if they are fine-tuned on proprietary and confidential data
- Limited performance
  - Their limited understanding and contextual awareness means they struggle with complex or niche topics

SLM USE CASE EXAMPLES

Mobile Apps

- **Offline Translation:** Travel apps could embed SLMs for translating menus, signs, or basic conversational phrases without internet access.
- **Grammar and Spelling Checkers**
- **Text Summarization:** News or productivity apps could summarize long articles locally using SLMs.

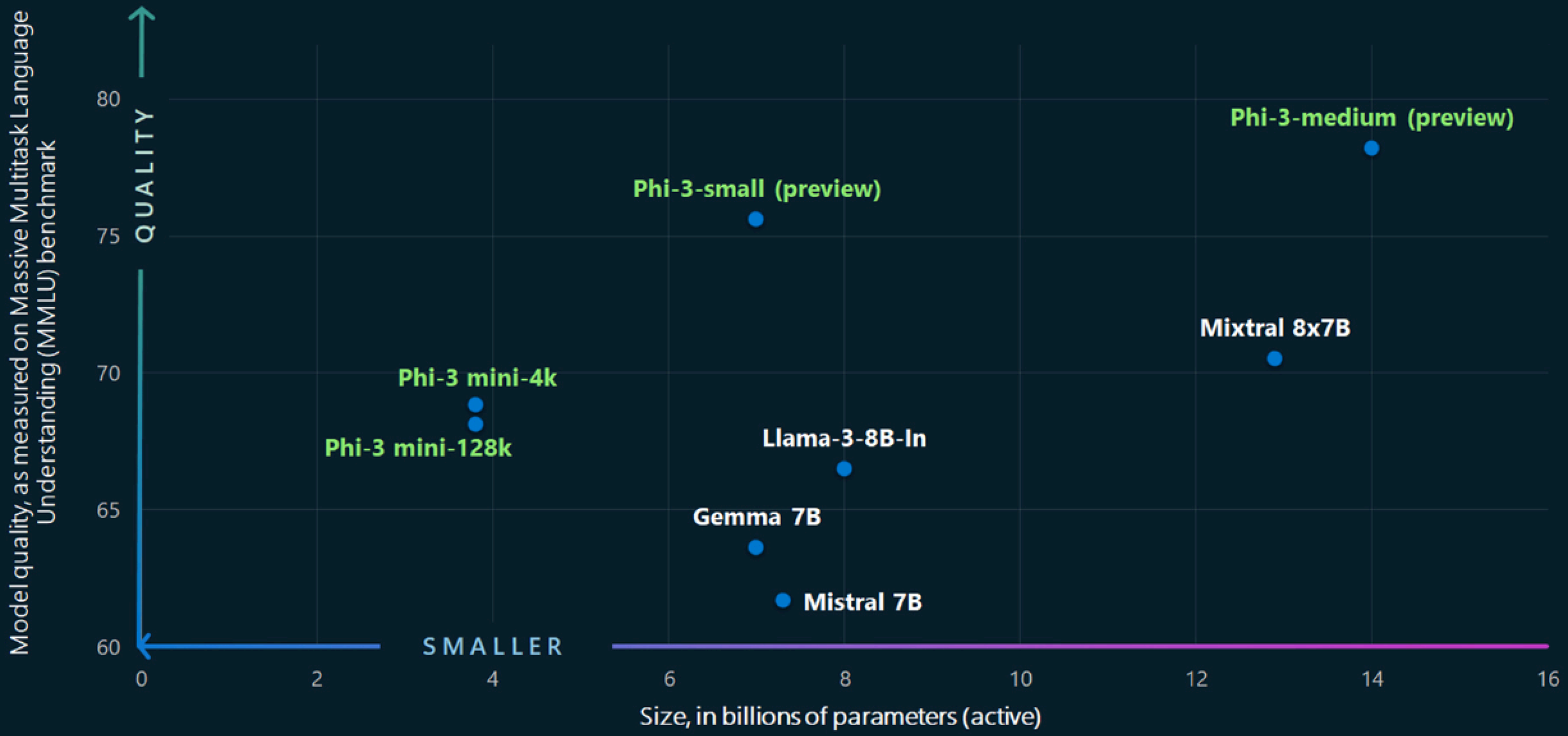
Video Games

- **Dynamic Dialogue Generation:** SLMs could generate more varied and context-aware dialogue for non-player characters (NPCs), leading to more immersive conversations in offline games.
- **Procedural Text Generation:** Open-world games could use SLMs to generate descriptions of items, quests, or even environmental details.

Other Devices

- **Smartwatches:** Voice assistants on smartwatches could use SLMs for basic commands and responses offline.
- **Smart Appliances:** SLMs could power voice recognition and simple command processing in offline modes.

Quality vs Size in Small Language Models (SLMs)



LATEST RESEARCH

- Microsoft Phi-3 is a lightweight and efficient model, achieved through a unique training method inspired by children’s stories, enabling effective learning from smaller datasets.

- Language learning by neural networks involves processing vast internet text. Drawbacks of generative models include expensive training and complex inner workings.

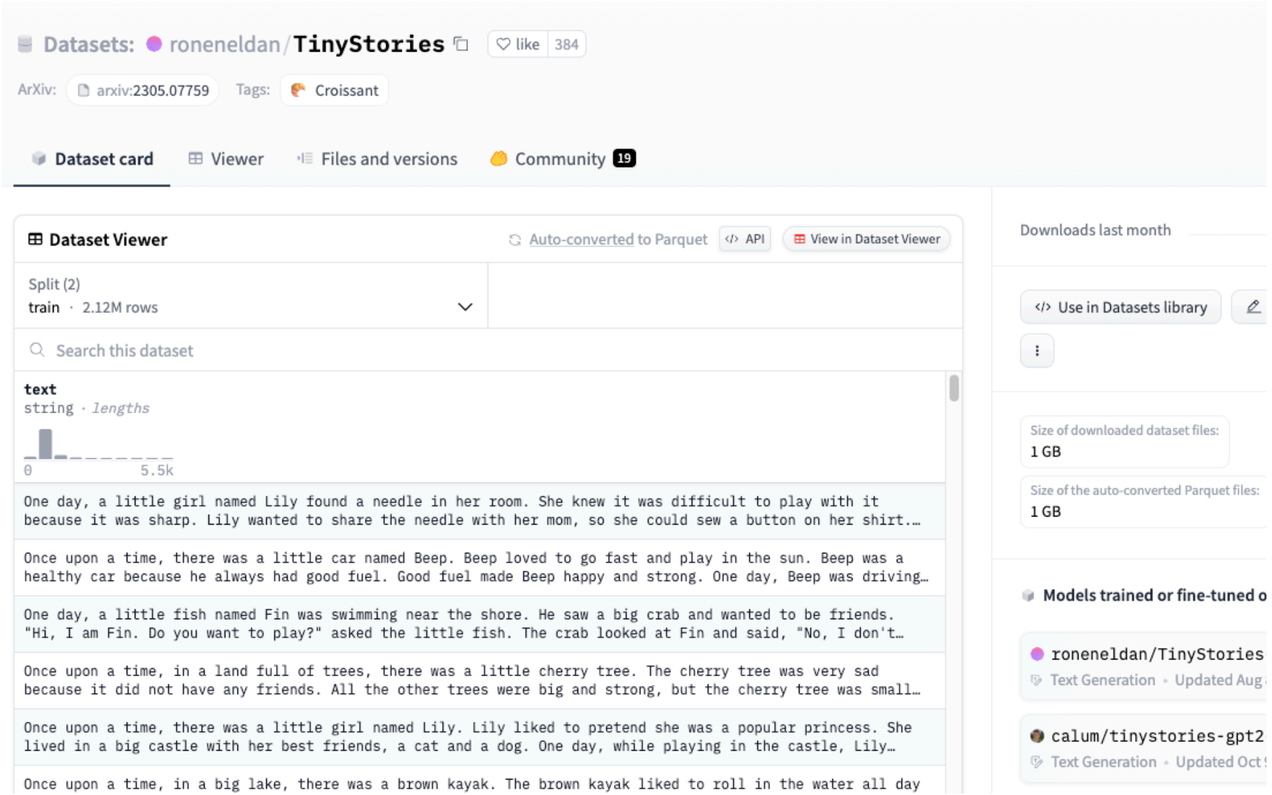
- Microsoft researchers devise a method using children’s stories to train small language models. This aims to understand large model behavior, showing that even tiny models can produce coherent stories.

- Large models generate synthetic children's stories as training data for smaller models. Children's stories' simplicity aids small model learning.

- Small models with 1-30 million parameters are trained and tested on the TinyStories dataset.

Test Results:
Larger models, even thousands of times smaller than GPT-3.5, performed well, while smaller ones struggled.

For example, a 28-million-parameter model produced a coherent story, contrasting with GPT-2's poor performance on a similar task.



High quality training dataset with 3k words: equal no. of noun, verbs & adjectives

LLM creates a story using 1 noun, 1 verb and 1 adjective from the list iterated millions of times

Millions of tiny children's stories "Tiny Stories" dataset

"Tiny Stories" dataset used as Training dataset for SLM with 10 million parameters

Fluent narratives with perfect grammar

CONCLUSION

- Unlike their larger counterparts, SLMs are designed for efficiency, leveraging advanced techniques to maintain or even enhance performance while significantly reducing computational demands.
- SLMs are increasingly favored for applications where computational efficiency, speed, and adaptability are crucial.
- They are ideal for deployment in edge devices, mobile platforms, and situations requiring rapid inference capabilities.
- They offer a sustainable alternative, with reduced carbon and water footprints compared to larger models, making them a more environmentally friendly choice for AI applications.

References

- https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/#:~:text=Small%20language%20models%20are%20designed,tuned%20to%20meet%20specific%20needs.
- https://medium.com/@bijit211987/the-rise-of-small-language-models-efficient-customizable-cb48ddee2aad
- https://www.arthur.ai/blog/the-beginners-guide-to-small-language-models
- https://arxiv.org/abs/2305.07759
- https://www.linkedin.com/pulse/small-language-models-slms-santiago-santa-mar%C3%ADa-morales-7gxee/