SCALExM.ai

YOUR TRUSTED PARTNER IN AI JOURNEY

WHITE PAPER

A Compact Guide to Large Language Models

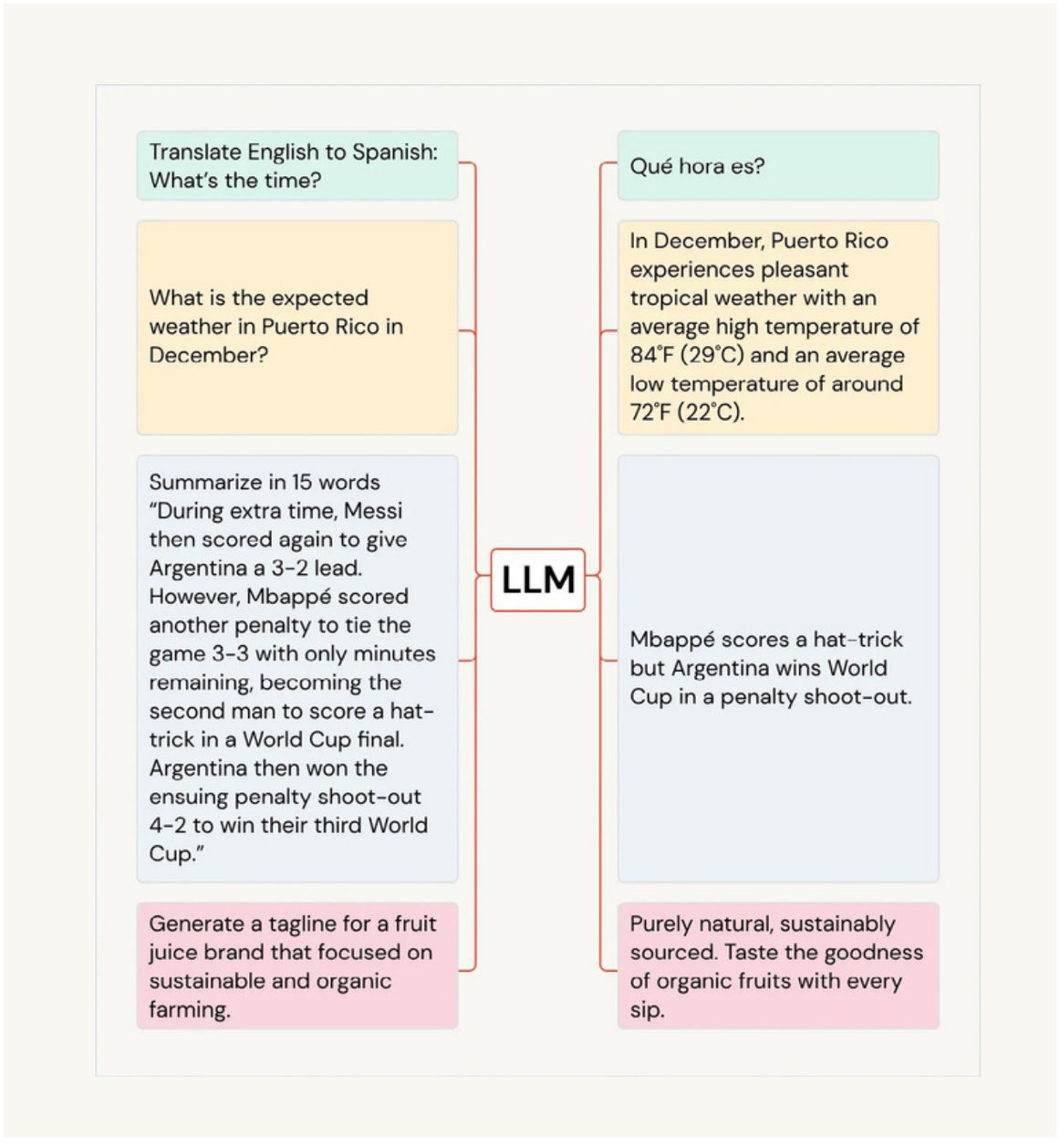
SECTION 1

Introduction

Definition of large language models (LLMs)

Large language models are AI systems that are designed to process and analyze vast amounts of natural language data and then use that information to generate responses to user prompts. These systems are trained on massive data sets using advanced machine learning algorithms to learn the patterns and structures of human language, and are capable of generating natural language responses to a wide range of written inputs. Large language models are becoming increasingly important in a variety of applications such as natural language processing, machine translation, code and text generation, and more.

While this guide will focus on language models, it's important to understand that they are only one aspect under a larger generative AI umbrella. Other noteworthy generative AI implementations include projects such as art generation from text, audio and video generation, and certainly more to come in the near future.



Extremely brief historical background and development of LLMs

1950s–1990s

Initial attempts are made to map hard rules around languages and follow logical steps to accomplish tasks like translating a sentence from one language to another.

While this works sometimes, strictly defined rules only work for concrete, well-defined tasks that the system has knowledge about.

1990s

Language models begin evolving into statistical models and language patterns start being analyzed, but larger-scale projects are limited by computing power.

2000s

Advancements in machine learning increase the complexity of language models, and the wide adoption of the internet sees an enormous increase in available training data.

2012

Advancements in deep learning architectures and larger data sets lead to the development of GPT (Generative Pre-trained Transformer).

2018

Google introduces BERT (Bidirectional Encoder Representations from Transformers), which is a big leap in architecture and paves the way for future large language models.

2020

OpenAI releases GPT-3, which becomes the largest model at 175B parameters and sets a new performance benchmark for language-related tasks.

2022

ChatGPT is launched, which turns GPT-3 and similar models into a service that is widely accessible to users through a web interface and kicks off a huge increase in public awareness of LLMs and generative AI.

2023

Open source LLMs begin showing increasingly impressive results with releases such as Dolly 2.0, LLaMA, Alpaca and Vicuna.
GPT-4 is also released, setting a new benchmark for both parameter size and performance.

SECTION 2

Understanding Large Language Models

What are language models and how do they work?

Large language models are advanced artificial intelligence systems that take some input and generate humanlike text as a response. They work by first analyzing vast amounts of data and creating an internal structure that models the natural language data sets that they're trained on. Once this internal structure has been developed, the models can then take input in the form of natural language and approximate a good response.

If they've been around for so many years, why are they just now making headlines?

A few recent advancements have really brought the spotlight to generative AI and large language models:

ADVANCEMENTS IN TECHNIQUES

Over the past few years, there have been significant advancements in the techniques used to train these models, resulting in big leaps in performance. Notably, one of the largest jumps in performance has come from integrating human feedback directly into the training process.

INCREASED ACCESSIBILITY

The release of ChatGPT opened the door for anyone with internet access to interact with one of the most advanced LLMs through a simple web interface. This brought the impressive advancements of LLMs into the spotlight, since previously these more powerful LLMs were only available to researchers with large amounts of resources and those with very deep technical knowledge.

GROWING COMPUTATIONAL POWER

The availability of more powerful computing resources, such as graphics processing units (GPUs), and better data processing techniques allowed researchers to train much larger models, improving the performance of these language models.

IMPROVED TRAINING DATA

As we get better at collecting and analyzing large amounts of data, the model performance has improved dramatically. In fact, our team showed that you could get amazing results training a relatively small model with a high-quality data set with Dolly 2.0.

So what are organizations using large language models for?

Here are just a few examples of common use cases for large language models:

CHATBOTS AND VIRTUAL ASSISTANTS

One of the most common implementations, LLMs can be used by organizations to provide help with things like customer support, troubleshooting, or even having open-ended conversations with user-provided prompts.

CODE GENERATION AND DEBUGGING

LLMs can be trained on large amounts of code examples and give useful code snippets as a response to a request written in natural language. With the proper techniques, LLMs can also be built in a way to reference other relevant data that it may not have been trained with, such as a company's documentation, to help provide more accurate responses.

SENTIMENT ANALYSIS

Often a hard task to quantify, LLMs can help take a piece of text and gauge emotion and opinions. This can help organizations gather the data and feedback needed to improve customer satisfaction.

TEXT CLASSIFICATION AND CLUSTERING

The ability to categorize and sort large volumes of data enables the identification of common themes and trends, supporting informed decision-making and more targeted strategies.

LANGUAGE TRANSLATION

Globalize all your content without hours of painstaking work by simply feeding your web pages through the proper LLMs and translating them to different languages. As more LLMs are trained in other languages, quality and availability will continue to improve.

SUMMARIZATION AND PARAPHRASING

Entire customer calls or meetings could be efficiently summarized so that others can more easily digest the content. LLMs can take large amounts of text and boil it down to just the most important bytes.

CONTENT GENERATION

Start with a detailed prompt and have an LLM develop an outline for you. Then continue on with those prompts and LLMs can generate a good first draft for you to build off. Use them to brainstorm ideas, and ask the LLM questions to help you draw inspiration from.

Note: Most LLMs are *not* trained to be fact machines. They know how to use language, but they might not know who won the big sporting event last year. It's always important to fact check and understand the responses before using them as a reference.

SECTION 3

Applying Large Language Models

There are a few paths that one can take when looking to apply large language models for their given use case. Generally speaking, you can break them down into two categories, but there's some crossover between each. We'll briefly cover the pros and cons of each and what scenarios fit best for each.

Proprietary services

As the first widely available LLM powered service, OpenAI's ChatGPT was the explosive charge that brought LLMs into the mainstream. ChatGPT provides a nice user interface (or API) where users can feed prompts to one of many models (GPT-3.5, GPT-4, and more) and typically get a fast response. These are among the highest-performing models, trained on enormous data sets, and are capable of extremely complex tasks both from a technical standpoint, such as code generation, as well as from a creative perspective like writing poetry in a specific style.

The downside of these services is the absolutely enormous amount of compute required not only to train them (OpenAI has said GPT-4 cost them over $100 million to develop) but also to serve the responses. For this reason, these extremely large models will likely always be under the control of organizations, and require you to send your data to their servers in order to interact with their language models. This raises privacy and security concerns, and also subjects users to "black box" models, whose training and guardrails they have no control over. Also, due to the compute required, these services are not free beyond a very limited use, so cost becomes a factor in applying these at scale.

In summary: Proprietary services are great to use if you have very complex tasks, are okay with sharing your data with a third party, and are prepared to incur costs if operating at any significant scale.

Open source models

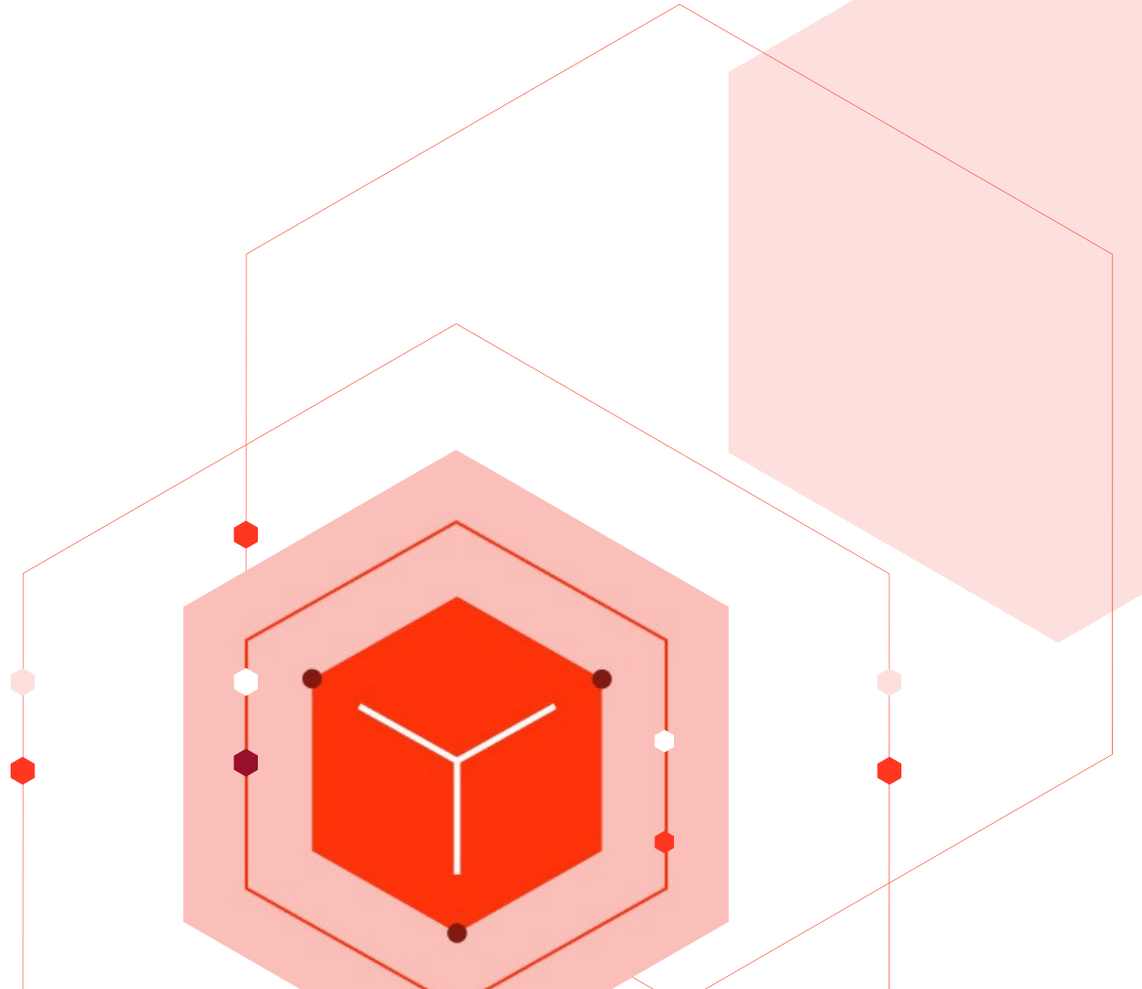
The other avenue for language models is to go to the open source community, where there has been similarly explosive growth over the past few years. Communities like Hugging Face gather hundreds of thousands of models from contributors that can help solve tons of specific use cases such as text generation, summarization and classification. The open source community has been quickly catching up to the performance of the proprietary models, but ultimately still hasn't matched the performance of something like GPT-4.

It does currently take a little bit more work to grab an open-source model and start using it, but progress is moving very quickly to make them more accessible to users. On AWS or Azure, for example, significant Improvements are made to open-source frameworks like MLflow to make it very easy for someone with a bit of Python experience to pull any Hugging Face transformer model and use it as a Python object. Oftentimes, you can find an open-source model that solves your specific problem that is **orders of magnitude** smaller than ChatGPT, allowing you to bring the model into your environment and host it yourself. This means that you can keep the data in your control for privacy and governance concerns, as well as manage your costs.

Another huge upside to using open-source models is the ability to fine-tune them to your own data. Since you're not dealing with a black box of a proprietary service, there are techniques that let you take open source models and train them to your specific data, greatly improving their performance on your specific domain. We believe the future of language models is going to move in this direction as more and more organizations will want full control and understanding of their LLMs.

Conclusion and general guidelines

Ultimately, every organization is going to have unique challenges to overcome, and there isn't a one-size-fits-all approach when it comes to LLMs. As the world becomes more data driven, everything, including LLMs, will be reliant on having a strong foundation of data. LLMs are incredible tools, but they have to be used and implemented on top of this strong data foundation. SCALExM.ai brings both that strong data foundation as well as the integrated tools to let you use and fine-tune LLMs in your domain.



SECTION 4

So What Do I Do Next If I Want to Start Using LLMs?

That depends where you are on your journey! Fortunately, we have a few paths for you.

If you want to go a little deeper into LLMs but aren't quite ready to do it yourself, you can watch one of SCALExM.ai' most talented developers and speakers go over these concepts in more detail during the on-demand talk "How to Build Your Own Large Language Model Like Dolly."

If you're ready to dive a little deeper and expand your education and understanding of LLM foundations, we'd recommend checking out our course on LLMs. You'll learn how to develop production-ready LLM applications and dive into the theory behind foundation models.

If your hands are already shaking with excitement and you already have some working knowledge of Python, we'll provide some great examples with sample code that can get you up and running with LLMs right away!

Getting started with NLP using Hugging Face transformers pipelines

Fine-Tuning Large Language Models with Hugging Face and DeepSpeed

Introducing AI Functions: Integrating Large Language Models with SCALExM.ai SQL

SCALExM.ai

YOUR TRUSTED PARTNER IN AI JOURNEY

About SCALExM.ai

At SCALExM.ai, we're more than just a consulting firm. We're your partners in innovation, dedicated to helping small to medium-sized businesses thrive in a rapidly changing world. You can reach us at sales@scalexm.ai or visit us our SCALEXM website

CONTACT US

Contact us for a personalized workshop:
scalexm.ai/workshop/