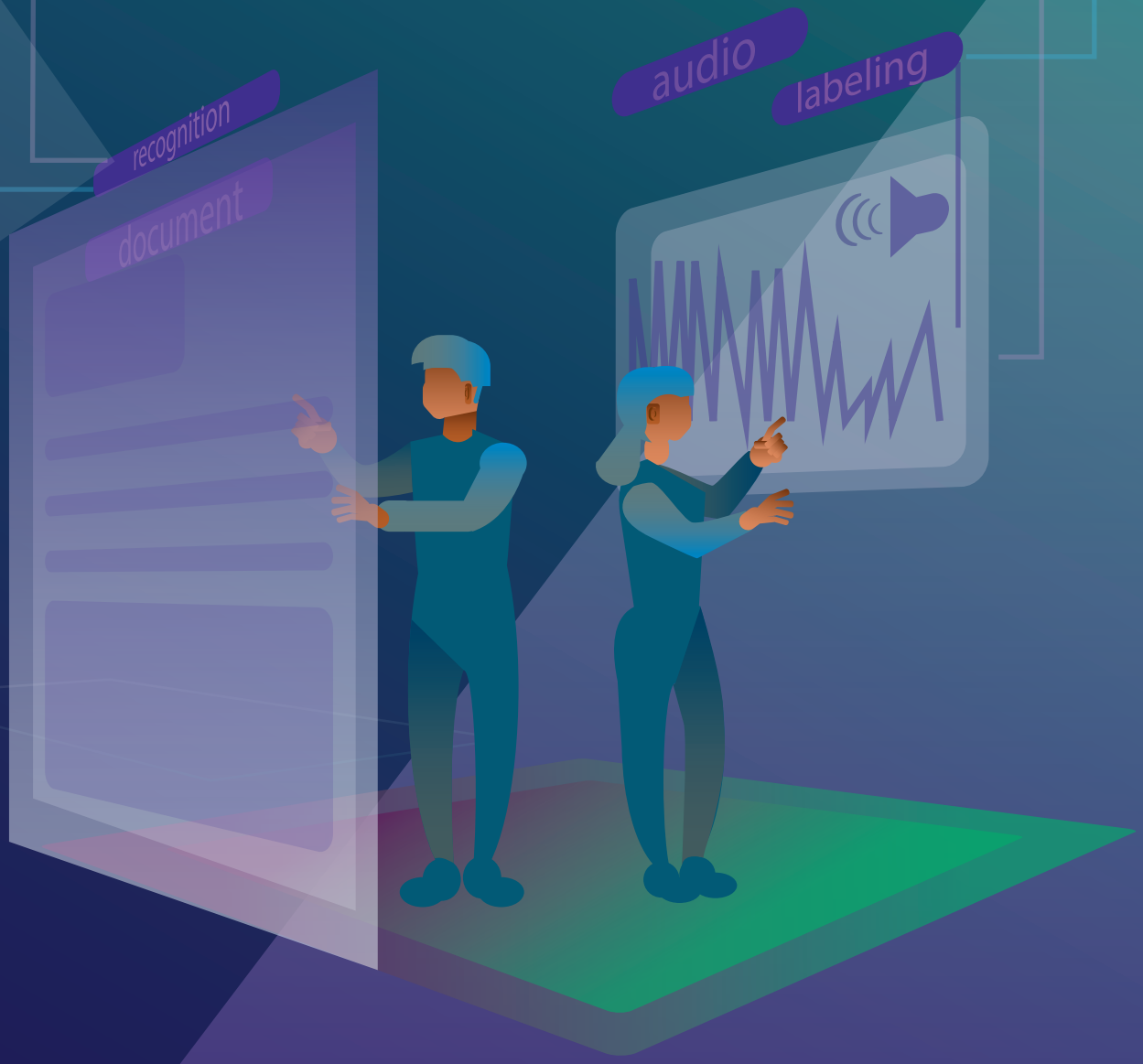
datasaur.ai

NOVEMBER 2023
COMPREHENSIVE GUIDE

PRODUCTIZING LARGE LANGUAGE MODELS



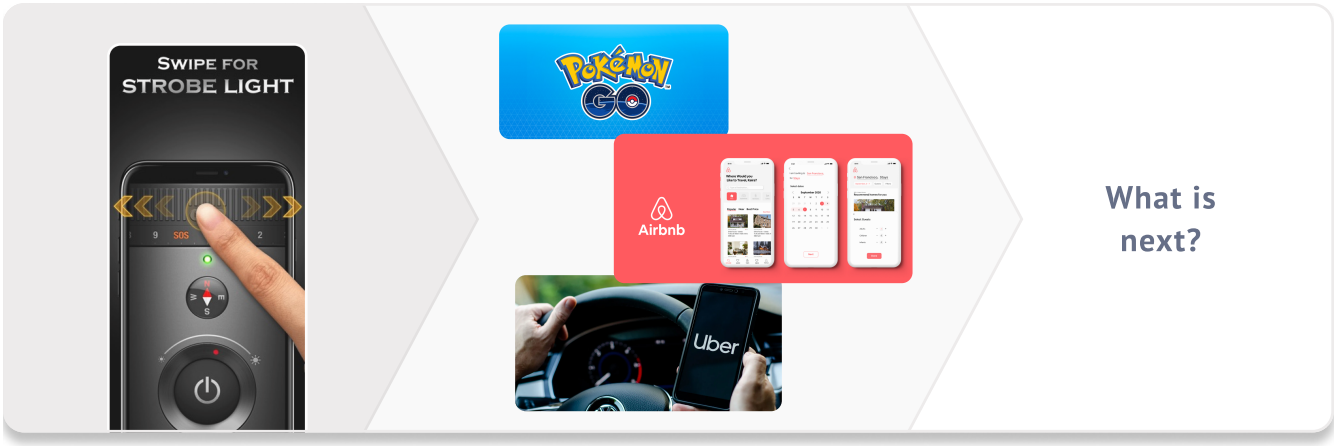
Introduction

Datasaur was founded four years ago to democratize access to the rapidly evolving field of Natural Language Processing. Its NLP platform is currently employed by over 100 companies and universities worldwide, in diverse sectors such as healthcare, legal, fintech, and e-commerce.

Datasaur supports companies such as Google, Zoom, Netflix, Qualtrics and Spotify and has been on the frontlines of Large Language Model (LLM) adoption and product development consultation. LLMs have generated a large amount of hype, but the industry still seems uncertain on how to concretely apply the technology in business use cases. This guide offers insights into how companies are thinking through their final product deliverables.

Looking to the Past to Predict the Future (of LLMs)

ChatGPT has become a viral hit with consumers and companies are now exploring how to best adopt this technology to their own real-world problems. In order to predict the trajectory of this technology's adoption, it's important to look back in time to the launch of the App Store in 2008. Everyone was excited about this new platform, where developers could leverage the potential of a PC in everyone's pocket and reach a wide audience. Yet despite the powerful capabilities of the iPhone, the first few years of the App Store were inundated with trivial products (flashlight apps, paper toss). It wasn't until years later that we saw mobile-native applications like Shazam, Uber, and AirBnB emerge. Similarly, right now we are seeing the low-hanging fruit that can be launched within months. Over the next couple of years, we are going to see more LLM-first applications. We need time for mass adoption of the new platform and time for innovation to occur.



Graphic: Application development from 2008 - now

In the 1980s, we had to purchase spell check software on a floppy disk, install this on our computers, write our document on a word processor, copy and paste into the spell check software, then copy and paste the results back into our word processor. Sound familiar? Companies everywhere are clamoring to build their own ChatGPT-like chatbots, but ChatGPT is as primitive as this interface will ever be. Moving forward, this feature will be baked right into the applications we use daily such as Gmail, Word, etc. Just as we find it odd to use an application without a spell check built in, so too will it feel weird to interact with an application that doesn't understand us speaking to it in ordinary language, or to not help predict what we're going to type next.

Graphic: The early days of spell checkers looked very similar to the early days of ChatGPT

Key Factors

- Defining success criteria with business stakeholders
- Balancing for precision vs. coverage
- OpenAI-based, other commercial vendor, or custom model (built on open source)
- Data privacy/InfoSec requirements

Planning Your LLM Journey

We've consulted with dozens of companies in helping build LLMs. Before you dive in, there are several critical factors to consider:

What is the final product deliverable?

It is imperative to work backwards from what you want to deliver to your users. One of the most common reasons AI solutions fail is mismatched expectations between what the stakeholders expect and what the technology can deliver. Surveys show that while 80% of business leaders said they use generative AI regularly, only 20% of frontline employees said the same. Take some time to meet with all your key stakeholders to answer the questions below:

- Will the LLM responses be directed straight to the end user (ex: a search results page) or will there be a human-in-the-loop to first verify the response (ex: suggesting copy for your website)?
- What defines success for the model? Answer correctness? User satisfaction? Cost savings? How will this be measured?
- What are the specific use cases you need the model to handle? Create a list of 20-30 prompts that are representative of the use cases, and decide how many need to be correct in order to consider this phase of the project a success.
    - Expert note: many AI projects fail because an executive comes in at the last moment, tries a new use case and declares you can't launch until it's resolved. This is why it's important to have everyone agree ahead of time on the metrics required for launch.
- Does the model need to be precise and accurate, or does it just need to help brainstorm ideas?
- Does it need to be deterministic (i.e. give the same response every time you ask it the same question)?

Understanding these questions will help set the right direction for the project and help navigate the questions below.

Technical questions

Foundation model

The first question you'll have to answer is which model will serve as your starting point. OpenAI, Cohere, Dolly, and LLaMa Claude 2, LlaMa-2, Falcon, and Vicuna are all viable options. OpenAI (and its Azure-based counterparts) are going to be the easiest and simplest options, but some use cases won't allow sending data to third parties. There are new open source models launching all the time, so this guide won't make any recommendations here. Each option comes with tradeoffs in quality and cost, so don't just pick the model with the most parameters! The most advanced users are building their technology stack in a foundation model-agnostic way, ensuring that they can quickly switch to and experiment with new models as they emerge.

Datasaur's aim is to maintain compatibility with all models. Users should be able to annotate datasets on Datasaur's platform and easily switch between one model and another to experiment with what provides the best results.

Deployment options

If you choose an open source model, how and where will you deploy your model? AWS, Azure and GCP are all racing to offer a wide range of services, alongside a new batch of startups. Understanding if you can leverage a third party service, if you need to host this on your servers, or if you even need this hosted on-prem on bare metal servers will impact the budget and timeline of your project. Datasaur supports all deployment options and major cloud providers.

Data Privacy and Compliance

This is the most overlooked topic when it comes to deploying LLMs. Enterprise software has long been required to meet a wide variety of compliance requirements, depending on the sensitivity of the data and the governing region. LLMs are brand new and regulation in the field is constantly evolving. In addition to understanding local laws, you also have to take access controls into account. We have seen what happens when users are given time to reverse engineer ChatGPT. What happens when an employee starts poking around looking for the size of the business contract with Coca Cola, or asking the LLM questions that the employee should not otherwise be able to access?

Datasaur is SOC-2 Type 2 and HIPAA compliant, and offers deployment options to customer VPCs or on-premise servers.

Conclusion

2023 will go down as a turning point as we move from the Web Era into the Age of AI. It's a thrilling time to be working with NLP and building new products. Taking a moment to understand lessons from the past and plan for your goals will pay many dividends as you dive into this field. Schedule a 45-minute consultation with an LLM expert here.