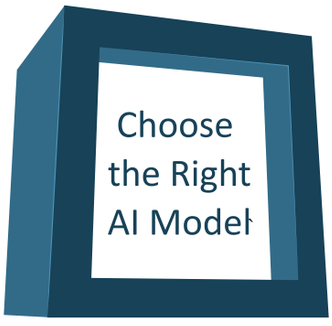
EXAFLUENCE
Data Driven Influence

Making the Leap: From GenAI Demos to Production Ready Apps

Kalaivani KG

February 2024

AI demo applications are sparking interest across enterprises in various industries. These demos are quick to develop and excellent for proof-of-concept work. However, evolving them into fully functional, production-ready applications is a different story. Let's walk through what it takes to navigate this transition successfully.



While demo apps often rely on straightforward platforms like OpenAI and simulated data, stepping up to real-world, sensitive data requires a more thoughtful choice. The decision between sticking with an external AI service or opting for a self-hosted, open-source model hinges on several key factors, such as data privacy, cost implications, and the model's fit for your specific needs. This choice is crucial and can significantly affect your timeline for launching the application.

The simple tools we love for demos, think Gradio or Streamlit, aren't quite up to snuff for the heavy lifting required in a production environment. What you need is a design that's ready to grow, separating the front-end and back-end to allow independent development and scaling. This setup ensures the app can support an increasing number of users smoothly.

A hallmark of GenAI applications is their composition of various components, like vector databases, language model APIs and prompt texts, which are subject to updates or replacements. Designing with modularity in mind allows for easy swapping of these components, facilitating updates and adapting to new tech as it emerges. This not only makes upgrades simpler but also keeps your app at the cutting edge.

Demos often show off the ideal scenario with limited examples. Moving to production, however, means preparing for a broader range of situations, including those edge cases, to ensure consistent and reliable results. A solid testing strategy across diverse scenarios is essential, coupled with an efficient framework for repetitive testing and outcome logging. For critical applications, incorporating a Human-in-the-Loop approach can add an essential layer of accuracy and reliability.

Running an AI application can vary widely in cost, influenced by user numbers and data volume. Early design decisions should anticipate scalability and the potential for cost increases. Effective cost management involves not just the AI model usage but also the infrastructure needed for supporting a growing user base and data processing demands. Strategies like optimizing query efficiency and leveraging caching can help keep costs in check.

As your application becomes accessible to a wider audience, safeguarding against risks becomes paramount. This includes defending against jailbreaks, prompt injection and other security threats. Robust security measures and access controls are essential for protecting the app and its data, ensuring a safe and reliable user experience.

The transition from a demo to a production-grade app is fundamentally about engaging with real users. Gathering and incorporating their feedback is key to understanding how they interact with your app, identifying issues, and evolving the app in line with user needs. This iterative process is vital for ensuring your application remains relevant and valuable to your audience.

Keeping tabs on your app's performance and quality post-launch is crucial. Continuous monitoring allows for timely adjustments, ensuring the app remains effective and responsive to both user needs and changes in the underlying AI model.

Post-launch, your app will need updates, bug fixes, and enhancements. Embracing continuous deployment facilitates the seamless rollout of these changes, ensuring your app remains up-to-date with minimal downtime.

Transitioning from a demo to a production-ready app involves a complex process that requires meticulous planning and attention to detail. By concentrating on these critical aspects, you can ensure your app is not only functional but also scalable, cost-effective, and secure.

If you would like to learn more, write to us at marketing@exafluence.com

For interesting videos about our solutions subscribe to our YouTube channel-
https://tinyurl.com/YouTubeExf

For regular updates on our solutions follow us on LinkedIn
https://tinyurl.com/LinkedInExf