



Construindo soluções de GenAI usando Open Source: impacto, desafios e custos

**Gustavo Pinto*

O uso de técnicas de Geração de Inteligência Artificial (GenAI) tem se expandido rapidamente em diversas indústrias, impactando áreas como geração de imagens, chatbots, sistemas de recomendação e geração de música.

[De acordo com uma pesquisa da Gartner com 1.400 líderes executivos](#), 45% informaram que sua organização está em fase piloto ou experimentando IA generativa, e outros 10% disseram que já implementaram soluções em produção. [Em um relatório semelhante da JPMorgan](#), foi apresentado que um dos principais benefícios do uso de ferramentas de GenAI provavelmente será a aceleração da produtividade do trabalho. Embora estimativas do impacto potencial variem bastante, segundo os especialistas, é esperado um aumento de 1,5% a 3,0% ao ano globalmente na próxima década.

A GenAI tem transformado segmentos como o setor de serviços ao cliente, onde serviços inteligentes otimizam interações e aprimoram a experiência dos consumidores. De acordo com 61% dos 65.000 desenvolvedores profissionais que responderam ao [StackOverflow survey de 2024](#), o uso de ferramentas de GenAI já faz parte do seu processo de desenvolvimento (enquanto que outros 13% embora ainda não o utilizem, pretendem fazê-lo em breve). Áreas como marketing, educação e saúde também devem ser impactadas de maneira semelhante.

Grande parte das soluções de GenAI no mercado atual são baseadas em modelos proprietários, como GPT da OpenAI e Gemini do Google. Esses modelos, aliados a pipelines orquestrados por ferramentas como Databricks e SageMaker, facilitam o desenvolvimento e implementação de soluções avançadas de IA. No entanto, uma das grandes limitações desse modelo fechado é a imprevisibilidade dos custos. Organizações que pretendem adotar GenAI enfrentam desafios em estimar quanto irão gastar em serviços de nuvem, infraestrutura e recursos computacionais, dado que os modelos privados cobram pelo uso de dados, processamento e armazenamento. Isso pode representar um entrave na tomada de decisão e na escalabilidade de projetos de GenAI.

Recentemente, o avanço de ferramentas e modelos de código aberto tem criado oportunidades para a construção de aplicações GenAI sem a dependência exclusiva de soluções proprietárias. Ferramentas como Hugging Face Transformers, LangChain e frameworks como Kubeflow e MLFlow permitem a criação de pipelines de aprendizado de máquina utilizando componentes abertos, com maior controle e flexibilidade sobre a customização e o desenvolvimento do fluxo de trabalho. Esses avanços são um marco importante, permitindo que empresas e desenvolvedores experimentem com modelos



GenAI e construam soluções que podem ser adaptadas para diferentes necessidades, com custos mais previsíveis e controlados.

Além das ferramentas, o surgimento de modelos de linguagem abertos, como o LLaMA da Meta, acelerou ainda mais o ecossistema GenAI. Diferentemente dos modelos fechados, como GPT ou Gemini, esses modelos podem ser treinados e ajustados diretamente pelas empresas, oferecendo maior transparência e capacidade de personalização para atender a requisitos específicos de negócio. O uso de modelos abertos facilita a adoção de práticas de privacidade mais rigorosas, uma vez que os dados não precisam ser enviados a provedores externos, e promove a inovação colaborativa, com desenvolvedores de todo o mundo contribuindo para a evolução contínua desses modelos. A combinação de ferramentas e modelos abertos abre novas possibilidades para criar soluções GenAI mais acessíveis e flexíveis, adaptadas a diferentes setores e demandas.

Ainda assim, o desenvolvimento de aplicações GenAI usando ferramentas e modelos abertos enfrenta uma série de desafios que devem ser superados. Por exemplo:

- A construção de casos de uso eficazes depende de um entendimento profundo das necessidades do negócio e da capacidade de integrar as soluções de IA nos processos existentes.
- A fragmentação de ferramentas no ecossistema de GenAI aberto representa um obstáculo significativo. Muitas dessas ferramentas não são integradas de forma nativa, dificultando a criação de pipelines coesos e eficientes. (Não é incomum que cientistas de dados gastam mais tempo resolvendo problemas com ferramentas do que desenvolvendo e treinando modelos.)
- A segurança e conformidade também são preocupações essenciais, especialmente em setores como o público e o de saúde, onde o manuseio de dados sensíveis exige camadas adicionais de proteção.
- A falta de padrões arquiteturais robustos para construção de projetos de GenAI, a escassez de talentos qualificados e os custos imprevisíveis tornam o cenário ainda mais complexo.

Se por um lado o uso de ferramentas e modelos abertos para criação de projetos de GenAI tendem a ser mais eficientes em termos de custos, personalização, além do suporte da comunidade e a possibilidade de contribuição ativa para o desenvolvimento das ferramentas, por outro lado ainda há desafios como preocupações com falta de suporte técnico formal, curva de aprendizado mais íngreme, documentação limitada e demora na correção de bugs — desafios estes bem conhecidos do ecossistema de software aberto.

O uso de GenAI apoiado por projetos de software livre oferece um caminho promissor para o desenvolvimento de soluções inovadoras e personalizáveis. Contudo, as organizações devem estar cientes dos desafios inerentes, como a fragmentação de ferramentas e as questões de segurança e conformidade. A previsão de custos e a falta de padrões estabelecidos são questões centrais que precisam ser resolvidas para que o potencial da GenAI, tanto em soluções abertas quanto proprietárias, seja plenamente realizado.



**Gustavo Pinto é Research Engineer na Zup, empresa de tecnologia parte do grupo Itaú Unibanco. Gustavo é doutor em Ciência da Computação pela UFPE, em 2015. É autor de mais de 100 artigos científicos na área da engenharia de software. Hoje atua como research engineer na Zup Innovation, utilizando técnicas de GenAI para apoiar o processo de modernização de código legado.*