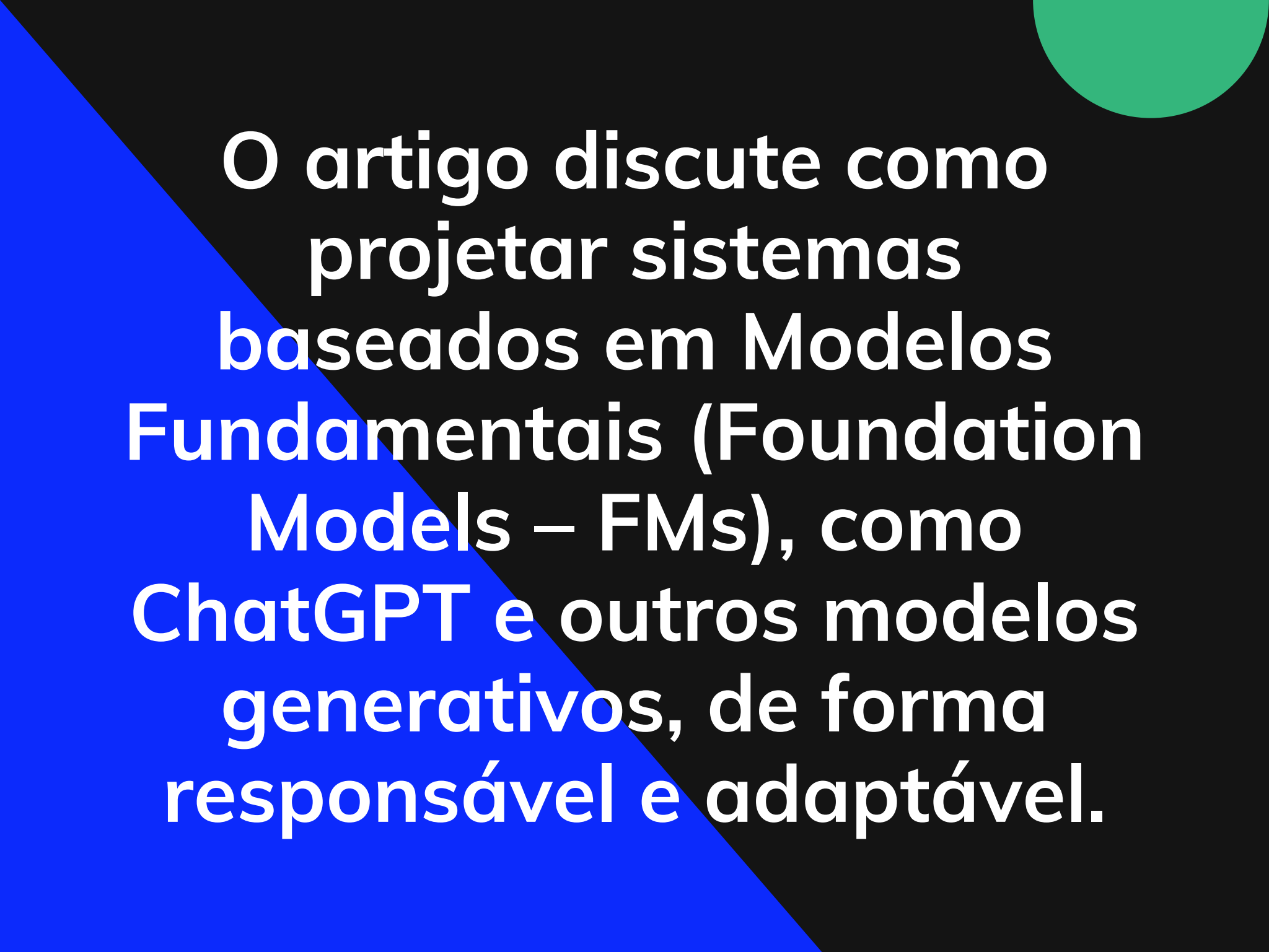




# Toward Responsible AI in the Era of Generative AI

Thiago Nascente Borges  
Vinícius Silva Benevides  
Moisés Ferreira Protázio Bastos  
Ana Liz Bomfim Gomes  
Mateus Silva de Souza



O artigo discute como  
projetar sistemas  
baseados em Modelos  
Fundamentais (Foundation  
Models – FMs), como  
ChatGPT e outros modelos  
generativos, de forma  
responsável e adaptável.



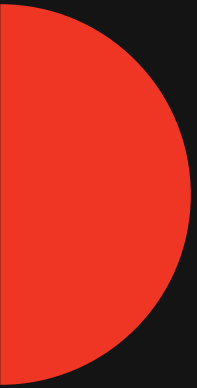
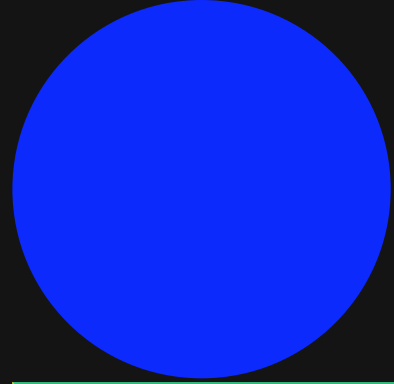
1 - 0  
problema



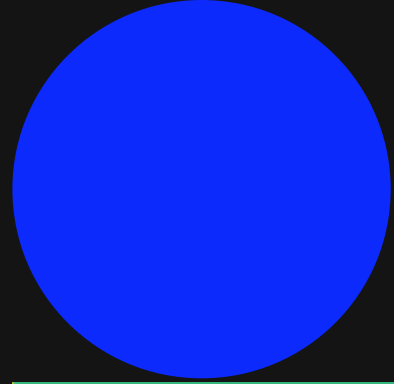
Os FMs são enormes modelos de IA treinados em dados gerais que podem ser adaptados para muitas tarefas.

Eles estão se tornando o bloco central dos sistemas de IA do futuro.

Mas isso gera desafios:



**Responsabilidade: quem responde por erros — o dono do sistema, o provedor do FM ou terceiros?**



**Confiabilidade: como garantir que saídas sejam corretas e seguras?**

**Riscos de mau uso: FMs podem ser usados para fins prejudiciais.**



**Evolução da arquitetura: FMs tendem a absorver funções de outros componentes, mudando a estrutura dos sistemas.**



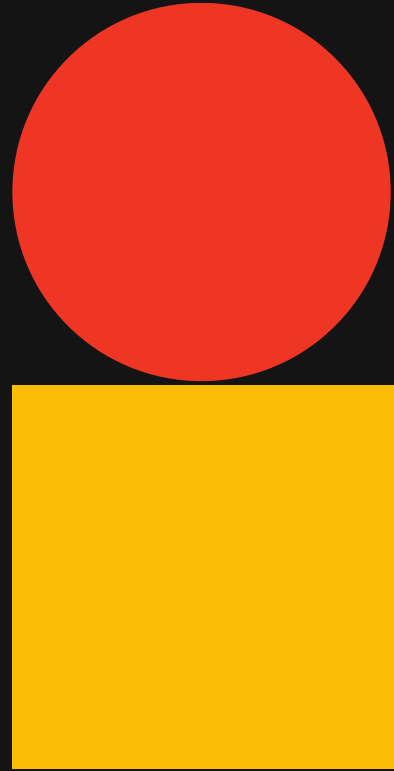
# 2 - Evolução da Arquitetura de Sistemas de IA



**Agora: muitos modelos  
menores de IA mais  
componentes tradicionais.**

**Em 5 anos: 1 FM como conector,  
coordenando outros componentes.**

**Durante 10 anos houveram  
duas possibilidades:**






1) Cadeia de FMs (vários modelos especializados se comunicando).

2) Um FM gigante e único (monolítico), capaz de fazer quase tudo sozinho.

Isso exige adaptabilidade (capacidade de se ajustar em tempo de execução) e modificabilidade (facilidade de alterar/atualizar).







# 3 - Decisões de Design Importantes

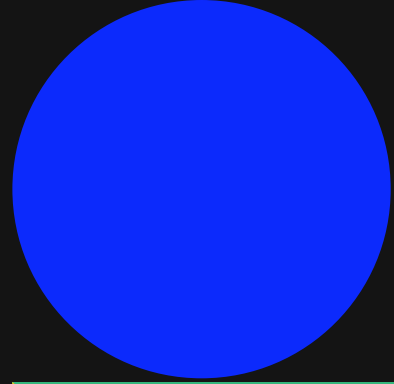


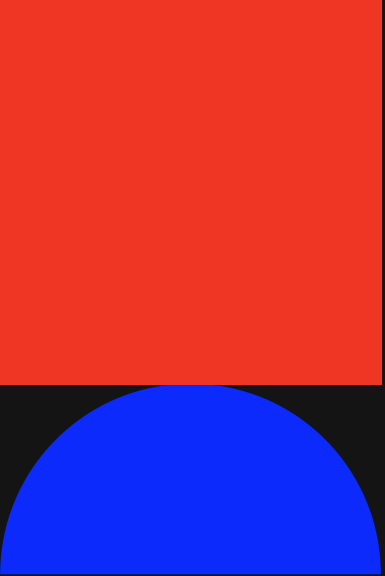
Os arquitetos de software precisam tomar várias decisões, como:

Qual tipo de FM usar: externo, ajustado (fine-tuned) ou soberano (treinado pela própria empresa).

Cadeia de FMs vs. FM único gigante?

Quais as responsabilidades do FM e o que ficará em componentes externos.






Respostas automáticas vs. uso de verificadores (humanos ou IA).

Interação passiva (usuário pergunta) vs. interação proativa (sistema prevê intenções do usuário).

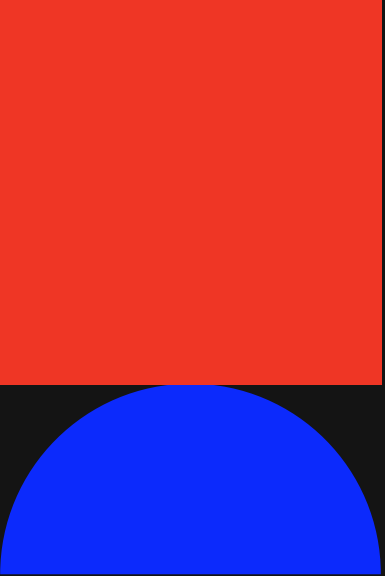
Um único agente vs. equipe de agentes colaborativos.

Pensar em voz alta (mostrar o raciocínio) vs. pensar silenciosamente (apenas resultado final).





# 4 - Arquitetura de Referência Proposta



O artigo apresenta uma arquitetura em três camadas:

System Layer → onde ficam os FMs, agentes, contexto multimodal, geração de prompts, verificadores.

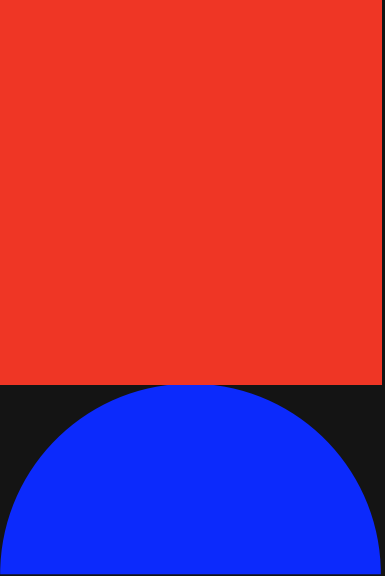
Operation Layer → garante IA responsável, com:

Black box recorder (registrar entradas/saídas para auditoria).

Guardrails (limites contra usos indevidos).

Avaliação contínua de riscos.





Supply Chain Layer → como os FMs e componentes são desenvolvidos ou adquiridos, incluindo:

Fine-tuning com feedback humano (RLHF).

Registro de procedência e métricas de IA responsável.





# 5- Avaliação com um exemplo

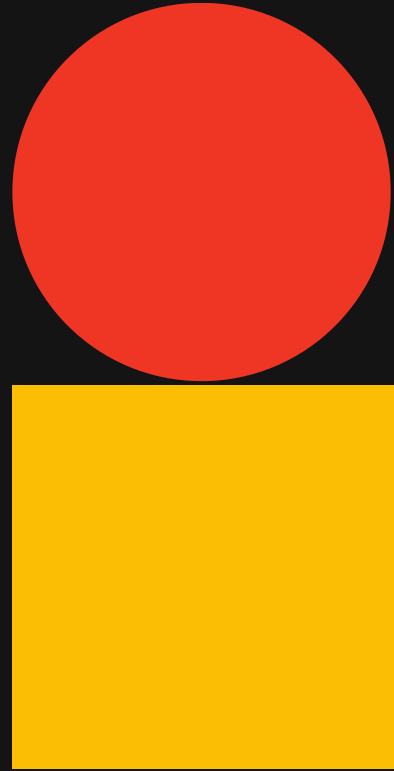
Eles aplicaram a arquitetura em um chatbot de Responsible AI (RAI) baseado no GPT-4. Esse chatbot:

Usa RAG (busca em base de dados local).

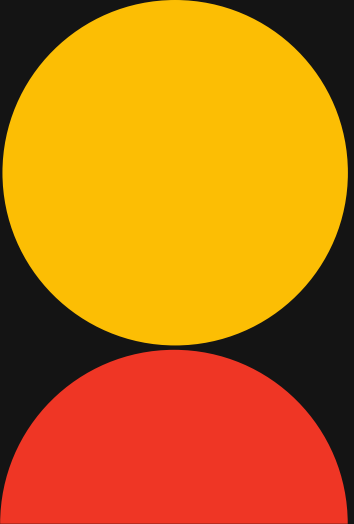
Tem verificadores humanos.

Registra conversas em uma black box.

Rejeita perguntas fora do escopo (responsabilidade).





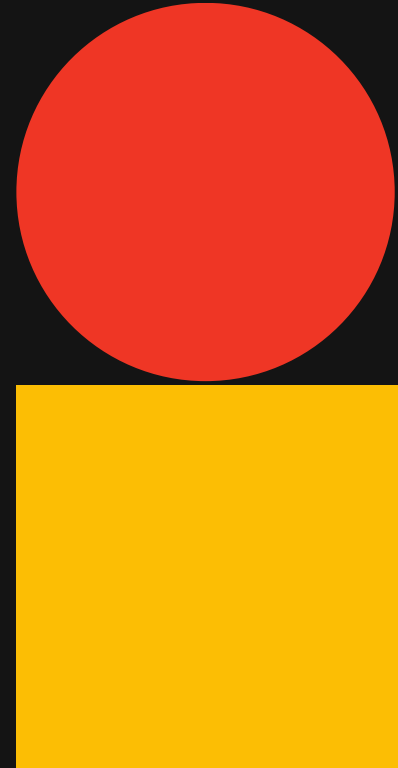


# Conclusão

O artigo propõe uma arquitetura de referência com padrões de design para garantir que sistemas com FMs sejam responsáveis, confiáveis e adaptáveis.

Os pontos-chave são: adaptabilidade, modificabilidade, e responsabilidade clara entre stakeholders.

No futuro, eles planejam criar um catálogo de padrões para facilitar o design de sistemas com FMs.



# OBRIGADO

Fonte:

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10553223>

