Caso prático: validação de um chatbot de políticas

"A inteligência artificial atingirá níveis humanos até 2029".
Ray Kurzweil[132]

"Acho que teremos uma IA mais inteligente do que qualquer ser humano provavelmente até o final de 2025".
Perplexity[133]

Para ilustrar a aplicação das técnicas de validação do LLM descritas acima, esta seção apresenta um estudo de caso da validação de um chatbot de políticas internas de uma empresa.

Definição de caso

A empresa desenvolveu um chatbot baseado em um LLM de código aberto para responder a perguntas e fornecer informações sobre suas políticas internas. O principal objetivo desse chatbot é facilitar o acesso dos funcionários às políticas da empresa.

O chatbot foi criado usando uma infraestrutura de nuvem e foi alimentado com todas as políticas da empresa, que compreendem aproximadamente 1.000 páginas de documentação. Para melhorar sua capacidade de resposta, foram aplicadas técnicas de Retrieval-Augmented Generation (RAG), permitindo que o modelo recupere informações relevantes de sua base de conhecimento antes de gerar uma resposta. Inicialmente, foi considerado o fine-tuning do modelo, mas, após os testes iniciais, concluiu-se que a combinação do LLM básico com o RAG era suficiente para obter resultados satisfatórios.

Antes de sua implementação final, a empresa decidiu conduzir um processo de validação completo para avaliar a precisão, a segurança e a adequação do chatbot no contexto específico de seu uso pretendido. Esse processo de validação tem como objetivo identificar possíveis áreas de melhoria e garantir que o chatbot atenda aos padrões de qualidade e às expectativas da empresa.

A validação do chatbot de políticas será conduzida usando uma combinação de métricas quantitativas e técnicas de avaliação humana, seguindo a estrutura de validação multidimensional descrita na seção anterior. Os resultados desse processo serão usados para tomar decisões informadas sobre a implementação do chatbot e para estabelecer um plano de melhoria contínua.

Desenho da abordagem de validação

Para validar de forma abrangente o chatbot de políticas, seguindo a estrutura de validação apresentada na seção anterior, foi desenhada uma abordagem de validação personalizada que abrange as principais dimensões do ciclo de vida do modelo: dados, design, avaliação, implementação e uso. Essa abordagem combina métricas quantitativas e técnicas de avaliação humana, com o objetivo de obter uma visão completa do desempenho e da adequação do chatbot no contexto específico da empresa.

Os testes e as técnicas selecionados para cada dimensão estão resumidos abaixo:

Dados

- Métricas: a escala Flesch-Kincaid será usada para avaliar a legibilidade e a complexidade das políticas que alimentam o chatbot.
- Avaliação humana: uma amostra representativa das políticas será analisada para identificar possíveis inconsistências, erros ou ambiguidades.

Desenho do modelo

- Métricas: elementos específicos do LLM serão modificados no código de desenvolvimento (por exemplo, a técnica RAG e seus hiperparâmetros, como o tamanho ou a estratégia de "chunking"[134]) que podem modificar seu desempenho de resposta, e os resultados serão comparados com o modelo original.

[132] Ray Kurzweil (nascido em 1948). Diretor de engenharia do Google, cientista da computação, inventor e futurista, conhecido pela invenção do OCR e por suas contribuições à IA.

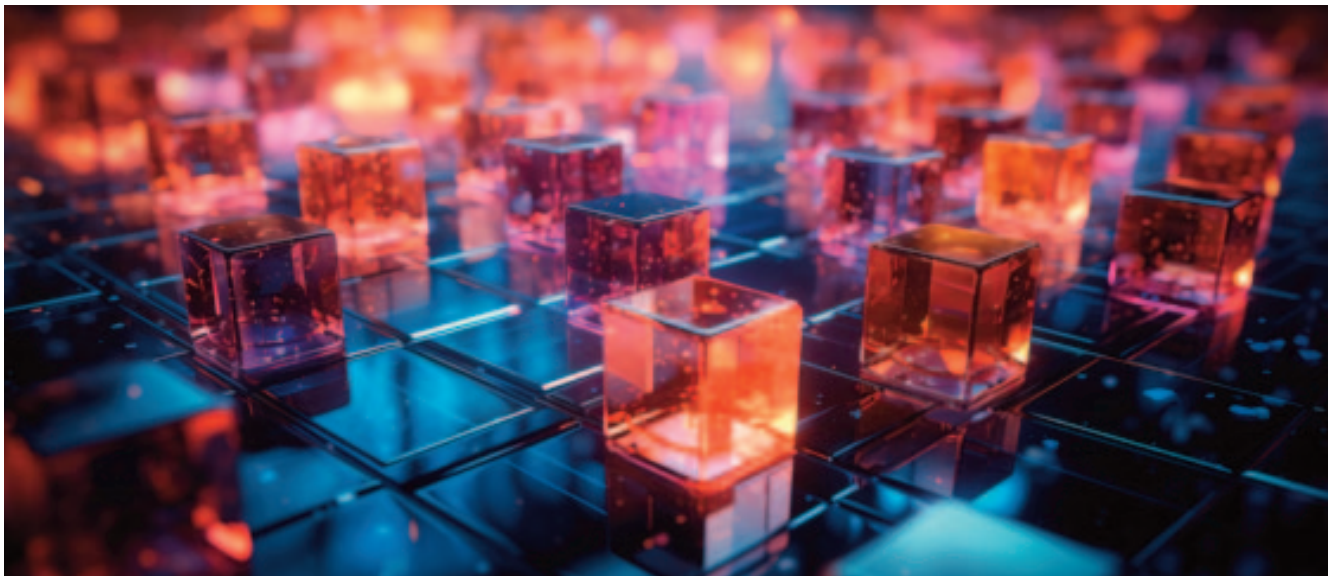
[133] Elon Musk (nascido em 1971), CEO da X, SpaceX e Tesla. Empresário sul-africano, conhecido por fundar ou cofundar empresas como Tesla, SpaceX e PayPal, proprietário do X (antigo Twitter), uma rede social que tem seu próprio LLM, chamado Grok.

[134] Chunking refere-se ao processo de dividir o texto de entrada do LLM em unidades menores e mais gerenciáveis ("chunks") durante o uso ou a implementação.

- Avaliação humana: será realizada uma revisão completa dos componentes do chatbot, incluindo a configuração do RAG, filtros de entrada e saída, definição de prompts e otimização de hiperparâmetros. Além disso, serão realizados testes A/B para comparar o desempenho do chatbot com outros LLMs disponíveis no mercado.

Avaliação do modelo

- Privacidade e segurança
  - Métricas: os testes de anonimização K serão aplicados para avaliar a proteção de dados pessoais nas respostas do chatbot, e os testes de PII (Personal Identificable Information) serão aplicados para identificar atributos confidenciais nos dados, usando o PIIfilter.
  - Avaliação humana: serão realizados testes de hacking ético para identificar possíveis vulnerabilidades e serão mantidos registros detalhados das interações do chatbot.
- Precisão
  - Métricas: as métricas Word Error Rate (WER) e ROUGE serão usadas para avaliar a precisão das respostas do chatbot em comparação com as políticas originais. Também serão usados benchmarks específicos do domínio, como um conjunto de perguntas e respostas criadas pelos especialistas em políticas da empresa.
  - Avaliação humana: será realizada uma revisão caso a caso de uma amostra representativa das interações do chatbot para identificar possíveis erros ou imprecisões.
- Consistência
  - Métricas: a similaridade de cosseno e o índice de Jaccard serão aplicados para avaliar a consistência das respostas do chatbot a consultas semelhantes.
  - Avaliação humana: testes A/B serão realizados para comparar as respostas do chatbot em diferentes cenários e uma revisão caso a caso será conduzida para identificar possíveis inconsistências.
- Robustez
  - Métricas: ferramentas como o TextFooler serão usadas para gerar textos adversários e avaliar a resistência do chatbot a informações enganosas. Além disso, será contado o número de rejeições de prompts maliciosos pelo chatbot.
  - Avaliação humana: testes de hacking ético e incidentes simulados serão realizados para avaliar a capacidade do chatbot de lidar com situações adversas.
- Adaptabilidade
  - Métricas: o desempenho do chatbot será avaliado em relação a novas políticas ou atualizações usando técnicas de few-shot learning. Será avaliada a resposta do chatbot a idiomas não usados nas políticas ou solicitações de tradução para idiomas não incluídos no RAG (por exemplo, polonês).
  - Avaliação humana: testes A/B e revisões caso a caso serão realizados para avaliar a capacidade do chatbot de se adaptar a novos cenários.
- Explicabilidade
  - Métricas: técnicas de explicabilidade, como o SHAP, serão aplicadas para entender o processo de tomada de decisão do chatbot. O módulo de interpretabilidade intrínseca do chatbot, que fornece uma explicação sobre a origem das informações na resposta ao usuário, será avaliado.
  - Avaliação humana: o monitoramento da experiência do usuário (UX) e um focus group serão conduzidos para avaliar a percepção dos usuários sobre a transparência e a capacidade de explicação do chatbot.
- Vieses e imparcialidade
  - Métricas: o kit de ferramentas AI Fairness 360 será usado para avaliar possíveis vieses demográficos nas respostas do chatbot. Referências específicas, como a Bias Benchmark for QA (BBQ), também serão usadas para medir a imparcialidade no contexto das políticas da empresa.
  - Avaliação humana: testes éticos de hacking e um focus group serão conduzidos para identificar possíveis vieses ou discriminação nas respostas do chatbot.
- Toxicidade
  - Métricas: as ferramentas da API Perspective e da API Hatebase serão aplicadas para avaliar a presença de linguagem tóxica ou inadequada nas respostas do chatbot. Além disso, benchmarks específicos, como o RealToxicityPrompts, serão usados para medir a toxicidade no contexto das políticas da empresa.
  - Avaliação humana: serão realizados testes de hacking ético para identificar possíveis instâncias de linguagem ofensiva ou inadequada nas interações do chatbot.



Implementação e uso

- Escalabilidade
  - Métricas: o teste de estresse do sistema será realizado usando o Apache JMeter para avaliar o desempenho do chatbot sob altas cargas de trabalho.
  - Avaliação humana: serão realizadas simulações para avaliar a capacidade do chatbot de lidar com um aumento inesperado no número de usuários ou consultas.
- Eficiência
  - Métricas: o tempo de resposta (Time-to-First-Byte, TTFB), o uso de recursos (GPU/CPU, memória) e a latência serão medidos para avaliar a eficiência do chatbot.
- Aceitação do usuário
  - Métricas: uma lista de verificação dos requisitos do usuário será estabelecida e a satisfação do usuário será medida usando indicadores como o Net Promoter Score (NPS) e o Customer Satisfaction Score (CSAT).
  - Avaliação humana: o rastreamento da experiência do usuário (UX) será realizado para avaliar a aceitação e a satisfação do usuário com o chatbot.

Essa abordagem de validação personalizada permitirá que a empresa obtenha uma avaliação abrangente do chatbot de políticas, identificando áreas de melhoria e garantindo sua adequação ao uso pretendido. Os resultados desses testes e avaliações servirão como base para decisões informadas sobre a implementação e o refinamento contínuo do chatbot.

Resultados

Depois de aplicar a abordagem de validação personalizada ao chatbot de políticas, foram obtidos resultados promissores, demonstrando sua adequação geral ao uso pretendido pela empresa (Fig. XX). Na maioria das dimensões avaliadas, o chatbot obteve um desempenho satisfatório, atendendo aos padrões de qualidade e às expectativas estabelecidas.

Em termos de qualidade dos dados de entrada, verificou-se que as políticas que alimentam o chatbot têm, em geral, um nível adequado de legibilidade e complexidade para os usuários entenderem. Além disso, a revisão humana não identificou inconsistências ou erros significativos no conteúdo das políticas.

O design do modelo também se mostrou apropriado para o caso de uso, com uma configuração ideal dos componentes do chatbot e desempenho superior em comparação com outros LLMs disponíveis no mercado.

Em termos de avaliação do modelo, o chatbot obteve resultados positivos na maioria das métricas e testes aplicados. Destacam-se a alta precisão das respostas, a consistência no tratamento de consultas semelhantes e a capacidade de adaptação a novos cenários. No entanto, foram identificadas algumas áreas de melhoria em aspectos como explicabilidade, detecção de vieses e resposta a perguntas muito específicas, em que é necessário um refinamento adicional do modelo. Na área de segurança cibernética, é necessária uma análise mais detalhada das vulnerabilidades específicas dos LLMs de código aberto usados para mitigar esse risco na produção.

Em termos de implementação e uso, o chatbot demonstrou boa escalabilidade e eficiência no tratamento de altas cargas de trabalho. Além disso, a satisfação do usuário foi alta, indicando uma boa aceitação da ferramenta no contexto da empresa.

Fig. 13. Resumo dos resultados de métricas e técnicas para avaliação humana do chatbot de políticas.

Dimensão	Teste	Resultado	Interpretação
Dados	**Flesch-Kincaid**	Legibilidade adequada (nota 8)	As políticas são compreensíveis para a maioria dos usuários
	Revisão humana	Não h inconsistências significativas	As políticas são coerentes e não contêm erros graves
Desenho do modelo	**Modelos challenger**	Melhorias de parâmetros identificadas	É necessário adaptar os parâmetros do RAG ao contexto da política (ou seja, tamanho do bloco) para melhorar a captura de informações sobre perguntas muito específicas.
	Revisão dos componentes	Configuração ideal	O desenho do chatbot é apropriado para o caso de uso.
	Testes A/B	Desempenho superior ao de outros LLMs	O chatbot supera o desempenho de outros modelos disponíveis no mercado.
Avaliação do modelo	**K-anonimato**	Proteção adequada de dados pessoais	O chatbot não revela informações confidenciais em suas respostas
	***Hacking* ético**	Vulnerabilidades menores identificadas	Ajustes necessários para fortalecer a segurança do chatbot
	Word Error Rate (WER)	WER < 5%	As respostas do chatbot são altamente precisas
	ROUGE	ROUGE-L > 0.8	As respostas do chatbot capturam adequadamente o conteúdo da política
	Similaridade de cosseno / índice Jaccard	Similaridade > 0.9	O chatbot fornece respostas consistentes para consultas semelhantes
	TextFooler	Resiliência moderada a textos contraditórios	O chatbot é moderadamente resistente a informações enganosas
	Aprendizagem com poucos disparos	Adaptabilidade satisfatória	O chatbot pode se adaptar a novas políticas ou atualizações com o mínimo de treinamento, mas é necessário monitorar e adicionar esses novos documentos ao RAG periodicamente.
	SHAP	Explicabilidade limitada	A capacidade do chatbot de explicar suas decisões precisa ser aprimorada , embora o componente RAG tenha sido desenvolvido de forma que o LLM dê uma resposta autoexplicativa.
	AI Fairness 360 / BBQ	Pequenos vieses demográficos identificados	O chatbot tem alguns vieses que precisam ser atenuados
	API do Perspective / API do Hatebase	Baixa toxicidade (< 5%)	As respostas do chatbot raramente contêm linguagem tóxica ou inadequada.
Implementação e uso	**Apache JMeter**	Escalabilidade satisfatória (até 1.000 usuários)	O chatbot pode lidar com altas cargas de trabalho sem degradação significativa do desempenho
	TTFB / Uso de recursos / Latência	Eficiência adequada (TTFB < 1s, uso moderado)	O chatbot responde rapidamente e usa os recursos de forma eficiente
	NPS / CSAT	Alta satisfação (NPS > 60, CSAT > 80%)	Os usuários estão muito satisfeitos com o chatbot e o recomendariam a outras pessoas.

Esses resultados indicam que o chatbot de políticas está no caminho certo para ser implementado na empresa, embora tenham sido identificadas algumas áreas específicas para melhorias adicionais. A seção a seguir abordará as principais conclusões e recomendações derivadas desse processo de validação.

Principais conclusões

O processo de validação do chatbot de políticas mostrou que esse sistema baseado em LLM pode ser uma ferramenta valiosa para facilitar o acesso dos funcionários às informações relevantes da empresa. Os resultados obtidos nos vários testes e avaliações indicam que o chatbot atende amplamente aos requisitos de qualidade, segurança e eficiência definidos pela organização.

Entre os pontos fortes identificados, destacam-se a precisão e a consistência das respostas do chatbot, sua capacidade de se adaptar a novos cenários e sua escalabilidade para lidar com altas cargas de trabalho. Além disso, a satisfação do usuário com a ferramenta é alta, o que sugere uma boa aceitação e adoção pelos funcionários.

No entanto, o processo de validação também revelou algumas áreas de melhoria que precisam ser abordadas antes da implementação final do chatbot. Em particular, recomenda-se:

1. **Melhorar a explicabilidade do modelo:** técnicas mais avançadas precisam ser desenvolvidas para que o chatbot possa fornecer explicações claras e compreensíveis sobre seu processo de tomada de decisão. Isso aumentará a transparência e a confiança dos usuários na ferramenta. Embora o componente RAG tenha sido desenvolvido de forma que o LLM dê uma resposta autoexplicativa e faça referência à política correspondente, essa explicação não é totalmente clara para perguntas muito específicas.
2. **Mitigar os vieses identificados:** embora os vieses detectados sejam pequenos, é aconselhável aplicar técnicas de redução de vieses para garantir que as respostas do chatbot sejam justas e não discriminatórias. Sugere-se a revisão periódica dos vieses e a implementação de medidas corretivas quando necessário.
3. **Reforçar a segurança e a privacidade:** embora o chatbot esteja em conformidade com os padrões básicos de proteção de dados pessoais, recomenda-se realizar testes adicionais e recorrentes de hacking ético e adotar medidas de segurança mais robustas para evitar possíveis vulnerabilidades.
4. **Estabelecer um plano de monitoramento e melhoria contínua:** é essencial definir um processo de monitoramento e avaliação regulares do desempenho do chatbot para identificar oportunidades de melhoria e garantir seu funcionamento ideal a longo prazo. Esse plano deve incluir a coleta de feedback dos usuários, a atualização regular das políticas e a inclusão delas no banco de dados do chatbot, o monitoramento para melhorar os parâmetros usados no RAG e atualizá-los, além da incorporação de novas técnicas e tecnologias à medida que elas se tornarem disponíveis.

Em conclusão, o chatbot de políticas demonstrou potencial para melhorar a eficiência e a acessibilidade das informações na empresa. Com a implementação das melhorias sugeridas e o foco na melhoria contínua, esse sistema baseado em LLM pode se tornar uma ferramenta estratégica para o sucesso da organização. A recomendação final foi a de prosseguir com a implementação do chatbot, levando em conta as observações e recomendações derivadas desse processo de validação.