Caso práctico: validación de un chatbot de políticas

"La inteligencia artificial alcanzará niveles humanos hacia 2029".
Ray Kurzweil[127]

"Creo que tendremos una IA más inteligente que cualquier ser humano probablemente a finales de 2025".
Perplexity[128]

Para ilustrar la aplicación de las técnicas de validación de un LLM descritas, en esta sección se presenta un caso práctico de validación de un *chatbot* de políticas internas de una compañía.

Definición del caso

La compañía ha desarrollado un *chatbot* basado en un LLM de código abierto para responder preguntas y proporcionar información sobre sus políticas internas. El objetivo principal de este *chatbot* es facilitar el acceso de los empleados a las políticas de la empresa.

El *chatbot* ha sido construido utilizando una infraestructura en la nube y se ha alimentado con todas las políticas de la compañía, que abarcan aproximadamente 1.000 páginas de documentación. Para mejorar su capacidad de respuesta, se han aplicado técnicas de *Retrieval-Augmented Generation* (RAG), que permiten al modelo recuperar información relevante de su base de conocimientos antes de generar una respuesta. Inicialmente, se consideró la posibilidad de aplicar *fine-tuning* al modelo, pero tras las pruebas iniciales, se concluyó que la combinación del LLM base con RAG era suficiente para obtener resultados satisfactorios.

Antes de su implementación definitiva, la compañía ha decidido llevar a cabo un proceso de validación exhaustivo para evaluar la precisión, la seguridad y la adecuación del *chatbot* en el contexto específico de su uso previsto. Este proceso de validación tiene como objetivo identificar posibles áreas de mejora y garantizar que el *chatbot* cumpla con los estándares de calidad y las expectativas de la empresa.

La validación del *chatbot* de políticas se realizará mediante una combinación de métricas cuantitativas y técnicas de evaluación humana, siguiendo el marco de validación multidimensional descrito en la sección anterior. Los resultados de este proceso servirán para tomar decisiones informadas sobre la implementación del *chatbot* y para establecer un plan de mejora continua.

Diseño del enfoque de validación

Para validar de manera integral el *chatbot* de políticas, siguiendo el marco presentado en la sección anterior, se ha diseñado un enfoque de validación a medida que abarca las dimensiones clave del ciclo de vida del modelo: datos, diseño, evaluación, implementación y uso. Este enfoque combina métricas cuantitativas y técnicas de evaluación humana, con el objetivo de obtener una visión completa del desempeño y la adecuación del *chatbot* en el contexto específico de la compañía.

A continuación, se resumen las pruebas y técnicas seleccionadas para cada dimensión:

Datos

- Métricas: se utilizará la escala Flesch-Kincaid para evaluar la legibilidad y complejidad de las políticas que alimentan al *chatbot*.
- Evaluación humana: se revisará una muestra representativa de las políticas para identificar posibles inconsistencias, errores o ambigüedades.

Diseño del modelo

- Métricas: se modificarán elementos concretos del LLM en el código de desarrollo (p. ej., la técnica de RAG y sus hiperparámetros, como el tamaño o la estrategia de "*chunking*"[129]) que pueden modificar su rendimiento ante respuestas, y se compararán los resultados contra el modelo original.

[127] Ray Kurzweil (n. 1948). Director de Ingeniería en Google, científico computacional, inventor y futurista, conocido por la invención del OCR y por sus contribuciones en IA.

[128] Elon Musk (n. 1971), CEO de X, SpaceX, Tesla. Empresario sudafricano-estadounidense, conocido por fundar o cofundar empresas como Tesla, SpaceX y PayPal, dueño de X (anteriormente Twitter), red social que tiene su propio LLM, llamado Grok.

[129] La estrategia de "chunking" se refiere al proceso de dividir el texto de entrada a un LLM en unidades más pequeñas y manejables («chunks") durante su uso o implementación.

- Evaluación humana: se realizará una revisión exhaustiva de los componentes del *chatbot*, incluyendo la configuración de RAG, los filtros de entrada y salida, la definición de *prompts* y la optimización de hiperparámetros. Además, se llevarán a cabo pruebas A/B para comparar el desempeño del *chatbot* con otros LLM disponibles en el mercado.

Evaluación del modelo

- Privacidad y seguridad
  - Métricas: se aplicarán pruebas de K-anonimato para evaluar la protección de datos personales en las respuestas del *chatbot*, y pruebas de PII (*Personal Identificable Information*) para identificar atributos sensibles en los datos, utilizando PIIfilter.
  - Evaluación humana: se realizarán pruebas de *hacking* ético para identificar posibles vulnerabilidades y se mantendrán registros detallados de las interacciones del *chatbot*.
- Precisión
  - Métricas: se utilizarán las métricas *Word Error Rate* (WER) y ROUGE para evaluar la precisión de las respuestas del *chatbot* en comparación con las políticas originales. También se emplearán *benchmarks* específicos del dominio, como un conjunto de preguntas y respuestas diseñado por expertos en políticas de la compañía.
  - Evaluación humana: se realizará una revisión caso por caso de una muestra representativa de interacciones del *chatbot* para identificar posibles errores o imprecisiones.
- Consistencia
  - Métricas: se aplicarán la similitud coseno y el índice de Jaccard para evaluar la consistencia de las respuestas del *chatbot* ante consultas similares.
  - Evaluación humana: se llevarán a cabo pruebas A/B para comparar las respuestas del *chatbot* en diferentes escenarios y se realizará una revisión caso por caso para identificar posibles inconsistencias.
- Robustez
  - Métricas: se utilizarán herramientas como TextFooler para generar texto adversario y evaluar la resiliencia del *chatbot* ante información engañosa. Además, se contabilizará el número de rechazos del *chatbot* ante *prompts* malintencionados.
  - Evaluación humana: se realizarán pruebas de hacking ético y simulacros de incidentes para evaluar la capacidad del *chatbot* para manejar situaciones adversas.
- Adaptabilidad
  - Métricas: se evaluará el rendimiento del *chatbot* ante nuevas políticas o actualizaciones mediante técnicas de *few-shot learning*. Se evaluará la respuesta del *chatbot* ante idiomas no empleados en las políticas o solicitudes de traducciones a idiomas no incluidos en el RAG (p. ej., polaco).
  - Evaluación humana: se realizarán pruebas A/B y revisiones caso por caso para evaluar la capacidad del *chatbot* para adaptarse a nuevos escenarios.
- Explicabilidad
  - Métricas: se aplicarán técnicas de explicabilidad, como SHAP, para comprender el proceso de toma de decisiones del *chatbot*. Se evaluará el módulo de interpretabilidad intrínseco del *chatbot*, mediante el cual se provee una explicación del origen de la información en la respuesta al usuario.
  - Evaluación humana: se realizará un seguimiento de la experiencia del usuario (UX) y se llevará a cabo un *focus group* para evaluar la percepción de los usuarios sobre la transparencia y la explicabilidad del *chatbot*.
- Sesgos y equidad
  - Métricas: se utilizará el conjunto de herramientas AI Fairness 360 para evaluar posibles sesgos demográficos en las respuestas del *chatbot*. También se emplearán *benchmarks* específicos, como el Bias Benchmark for QA (BBQ), para medir la equidad en el contexto de las políticas de la compañía.
  - Evaluación humana: se llevarán a cabo pruebas de *hacking* ético y un *focus group* para identificar posibles sesgos o discriminación en las respuestas del *chatbot*.
- Toxicidad
  - Métricas: se aplicarán las herramientas Perspective API y Hatebase API para evaluar la presencia de lenguaje tóxico o inapropiado en las respuestas del *chatbot*. Además, se utilizarán benchmarks específicos, como RealToxicityPrompts, para medir la toxicidad en el contexto de las políticas de la compañía.
  - Evaluación humana: se realizarán pruebas de *hacking* ético para identificar posibles casos de lenguaje ofensivo o inapropiado en las interacciones del *chatbot*.

Implementación y uso

- Escalabilidad
  - Métricas: se realizarán pruebas de estrés del sistema utilizando Apache JMeter para evaluar el rendimiento del *chatbot* bajo altas cargas de trabajo.
  - Evaluación humana: se llevarán a cabo simulacros para evaluar la capacidad del *chatbot* para manejar un aumento imprevisto en el número de usuarios o consultas.
- Eficiencia
  - Métricas: se medirán el tiempo de respuesta (Time-to-First-Byte, TTFB), el uso de recursos (GPU/CPU, memoria) y la latencia para evaluar la eficiencia del *chatbot*.
- Aceptación del usuario
  - Métricas: se establecerá una lista de verificación de los requisitos del usuario y se medirá la satisfacción del usuario utilizando indicadores como el Net Promoter Score (NPS) y el Customer Satisfaction Score (CSAT).
  - Evaluación humana: se realizará un seguimiento de la experiencia del usuario (UX) para evaluar la aceptación y satisfacción de los usuarios con el *chatbot*.

Este enfoque de validación personalizado permitirá a la compañía obtener una evaluación completa del *chatbot* de políticas, identificando áreas de mejora y garantizando su adecuación para el uso previsto. Los resultados de estas pruebas y evaluaciones servirán como base para la toma de decisiones informadas sobre la implementación y el perfeccionamiento continuo del *chatbot*.

Resultados

Tras aplicar el enfoque de validación personalizado al *chatbot* de políticas, se obtuvieron resultados prometedores que demuestran su adecuación general para el uso previsto en la compañía (Fig. 13). En la mayoría de las dimensiones evaluadas, el *chatbot* alcanzó un desempeño satisfactorio, cumpliendo con los estándares de calidad y las expectativas establecidas.

En cuanto a la calidad de los datos de entrada, se encontró que las políticas que alimentan al *chatbot* tienen, en general, un nivel de legibilidad y complejidad adecuado para su comprensión por parte de los usuarios. Además, la revisión humana no identificó inconsistencias significativas o errores en el contenido de las políticas.

El diseño del modelo también demostró ser apropiado para el caso de uso, con una configuración óptima de los componentes del *chatbot* y un rendimiento superior en comparación con otros LLM disponibles en el mercado.

En términos de evaluación del modelo, el *chatbot* obtuvo resultados positivos en la mayoría de las métricas y pruebas aplicadas. Se destacan la alta precisión de las respuestas, la consistencia en el manejo de consultas similares y la capacidad para adaptarse a nuevos escenarios. Sin embargo, se identificaron algunas áreas de mejora en aspectos como la explicabilidad, la detección de sesgos y la respuesta a preguntas muy específicas donde se requiere un mayor perfeccionamiento del modelo. En el ámbito de ciberseguridad, se requiere un análisis más detallado de las vulnerabilidades específicas de los LLM *open-source* empleados, para mitigar ese riesgo en su puesta en producción.

En cuanto a la implementación y uso, el *chatbot* demostró una buena escalabilidad y eficiencia en el manejo de altas cargas de trabajo. Además, la satisfacción de los usuarios fue alta, lo que indica una buena aceptación de la herramienta en el contexto de la compañía.

Fig. 13. Resumen de resultados de las métricas y técnicas de evaluación humana del chatbot de políticas.

Dimensión	Prueba	Resultado	Interpretación
Datos	**Flesch-Kincaid**	Legibilidad adecuada (grado 8)	Las políticas son comprensibles para la mayoría de los usuarios
	Revisión humana	Sin inconsistencias significativas	Las políticas son coherentes y no contienen errores importantes
Diseño del modelo	**Modelos challenger**	Mejoras en parámetros identificadas	Se requiere adaptar los parámetros del RAG al contexto de las políticas (p. ej., chunk size) para mejorar la captura de información en preguntas muy específicas
	Revisión de componentes	Configuración óptima	El diseño del *chatbot* es apropiado para el caso de uso
	Pruebas A/B	Rendimiento superior a otros LLM	El *chatbot* supera a otros modelos disponibles en el mercado
Evaluación del modelo	**K-anonimato**	Protección adecuada de datos personales	El *chatbot* no revela información sensible en sus respuestas
	***Hacking* ético**	Vulnerabilidades menores identificadas	Se requieren ajustes para fortalecer la seguridad del *chatbot*
	Word Error Rate (WER)	WER < 5%	Las respuestas del *chatbot* son altamente precisas
	ROUGE	ROUGE-L > 0.8	Las respuestas del *chatbot* capturan adecuadamente el contenido de las políticas
	Similitud coseno / Índice de Jaccard	Similitud > 0.9	El *chatbot* proporciona respuestas consistentes ante consultas similares
	TextFooler	Resiliencia moderada ante texto adversario	El *chatbot* es moderadamente robusto ante información engañosa
	Few-shot learning	Adaptabilidad satisfactoria	El *chatbot* puede adaptarse a nuevas políticas o actualizaciones con un entrenamiento mínimo, pero se requiere monitorizar y agregar esos nuevos documentos al RAG periódicamente
	SHAP	Explicabilidad limitada	Se requiere mejorar la capacidad del *chatbot* para explicar sus decisiones, si bien el componente de RAG se ha construido de manera que el LLM da una respuesta autoexplicativa
	AI Fairness 360 / BBQ	Sesgos demográficos menores identificados	El *chatbot* presenta algunos sesgos que deben ser mitigados
	Perspective API / RealToxicityPrompts	Toxicidad baja (< 5%)	Las respuestas del *chatbot* rara vez contienen lenguaje tóxico o inapropiado
Implementación y uso	**Apache JMeter**	Escalabilidad satisfactoria (hasta 1000 usuarios)	El *chatbot* puede manejar altas cargas de trabajo sin degradación significativa del rendimiento
	TTFB / Uso de recursos / Latencia	Eficiencia adecuada (TTFB < 1s, uso moderado)	El *chatbot* responde rápidamente y utiliza los recursos de manera eficiente
	NPS / CSAT	Satisfacción alta (NPS > 60, CSAT > 80%)	Los usuarios están altamente satisfechos con el *chatbot* y lo recomendarían a otros

Estos resultados indican que el *chatbot* de políticas está bien encaminado para su implementación en la compañía, aunque se han identificado algunas áreas específicas que requieren mejoras adicionales. La sección siguiente abordará las principales conclusiones y recomendaciones derivadas de este proceso de validación.

Principales conclusiones

El proceso de validación del *chatbot* de políticas ha demostrado que este sistema basado en LLM puede ser una herramienta valiosa para facilitar el acceso de los empleados a la información relevante de la compañía. Los resultados obtenidos en las diversas pruebas y evaluaciones indican que el *chatbot* cumple, en gran medida, con los requisitos de calidad, seguridad y eficiencia establecidos por la organización.

Entre las fortalezas identificadas, se destacan la precisión y consistencia de las respuestas del *chatbot*, su capacidad para adaptarse a nuevos escenarios y su escalabilidad para manejar altas cargas de trabajo. Además, la satisfacción de los usuarios con la herramienta es alta, lo que sugiere una buena aceptación y adopción por parte de los empleados.

Sin embargo, el proceso de validación también ha revelado algunas áreas de mejora que deben abordarse antes de la implementación definitiva del *chatbot*. En particular, se recomienda:

1. **Mejorar la explicabilidad del modelo:** es necesario desarrollar técnicas más avanzadas para que el *chatbot* pueda proporcionar explicaciones claras y comprensibles sobre su proceso de toma de decisiones. Esto aumentará la transparencia y la confianza de los usuarios en la herramienta. Si bien el componente de RAG se ha construido de manera que el LLM da una respuesta autoexplicativa y hace referencia a la política correspondiente, esta explicación no resulta del todo clara para preguntas muy específicas.

2. **Mitigar los sesgos identificados:** aunque los sesgos detectados son menores, es recomendable aplicar técnicas de *debiasing* para garantizar que las respuestas del *chatbot* sean equitativas y no discriminatorias. Se sugiere una revisión periódica de los sesgos y la implementación de medidas correctivas cuando sea necesario.

3. **Fortalecer la seguridad y privacidad:** si bien el *chatbot* cumple con los estándares básicos de protección de datos personales, se recomienda realizar pruebas adicionales y recurrentes de *hacking* ético y adoptar medidas de seguridad más robustas para prevenir posibles vulnerabilidades.

4. **Establecer un plan de monitoreo y mejora continua:** es fundamental definir un proceso de seguimiento y evaluación periódica del desempeño del *chatbot*, con el fin de identificar oportunidades de mejora y garantizar su óptimo funcionamiento a largo plazo. Este plan debe incluir la recopilación de feedback de los usuarios, la actualización regular de las políticas y su inclusión en la base de datos del *chatbot*, el monitoreo para mejorar los parámetros empleados en el RAG y su actualización, y la incorporación de nuevas técnicas y tecnologías cuando estén disponibles.

En conclusión, el *chatbot* de políticas ha demostrado tener potencial para mejorar la eficiencia y la accesibilidad de la información en la compañía. Con la implementación de las mejoras sugeridas y un enfoque de perfeccionamiento continuo, este sistema basado en LLM puede convertirse en una herramienta estratégica para el éxito de la organización. La recomendación final ha sido proceder con la implementación del *chatbot*, teniendo en cuenta las observaciones y recomendaciones derivadas de este proceso de validación.