

Inglés → Español ▾



Buscar



Escribir

Inscribirse

Iniciar sesión



Creación de aplicaciones RAG listas para producción: una guía para arquitectos de soluciones de IA



Nilay Parikh · [Seguir](#)

Publicado en Avances de la IA · 8 minutos de lectura · 1 de febrero de 2024



1



**BUILDING PRODUCTION-READY
RAG APPLICATIONS: AN AI
SOLUTION ARCHITECT'S GUIDE**

LLM Guardrails
<https://nilayparikh.com/llms/>

ErgoSum / X Labs
<ergosum.in> | nilayparikh.com

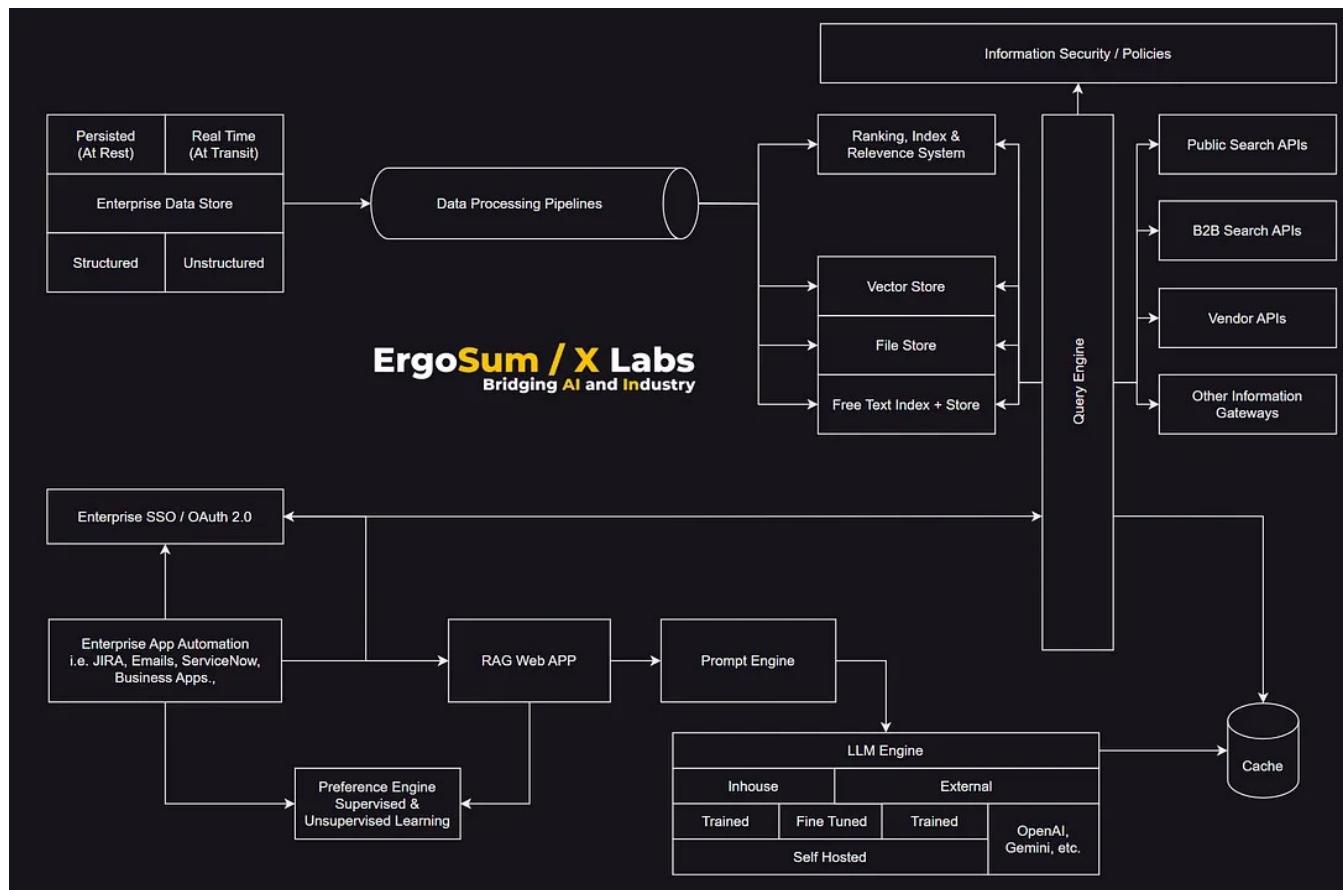
**FOLLOW
SUBSCRIBE**

niparikh
nilayparikh
@ergosumxlab

**LLM
RAG
PROMPT
ENGINEERING**

La recuperación de generación aumentada (RAG) es una técnica avanzada de inteligencia artificial que combina la recuperación de información con la generación de texto, lo que permite a los modelos de IA recuperar información relevante de una fuente de conocimiento e incorporarla al texto generado. RAG ha surgido como un punto de inflexión en la generación de contenido, revolucionando la forma en que generamos e interactuamos con el texto. En esta guía completa, profundizaremos en el mundo de la

generación aumentada de recuperación (RAG) y brindaremos una descripción detallada de cómo crear aplicaciones RAG listas para producción.



Plano de RAG de ErgoSum / X Labs

Si tiene poco tiempo y desea abarcar el concepto del sistema RAG, entonces

está en el lugar correcto. Este artículo está redactado para que la audiencia obtenga una vista única de los LLM RAG. En este artículo, aprenderá qué significa RAG, cómo funciona, cuáles son sus beneficios y desafíos, y cómo utilizarlo en su propio contexto. Si usted es gerente de proyectos, arquitecto de soluciones, líder de equipo o propietario de procesos, este artículo le resultará útil e informativo. ¡Entonces empecemos!

Una solicitud

Si disfrutó este artículo o lo encontró útil, muestre su agradecimiento aplaudiendo, compartiéndolo o dale me gusta. Sus comentarios nos animan a seguir compartiendo nuestra perspectiva neutral.

Introducción

Las aplicaciones RAG son cada vez más relevantes en diversos campos, incluido el procesamiento del lenguaje natural, la creación de contenidos y la respuesta a preguntas. RAG proporciona una solución para generar texto que no sólo es fluido sino también preciso y rico en información. Al combinar modelos de recuperación con modelos generativos, RAG garantiza que el texto que produce esté bien informado y bien escrito.

Construir sistemas RAG robustos puede ser un desafío debido a la

complejidad de los modelos subyacentes y la necesidad de integrarlos con los sistemas existentes. Algunos de los desafíos incluyen:

- **Calidad de los datos** : los modelos RAG requieren datos de alta calidad para generar texto preciso e informativo.
- **Entrenamiento de modelos** : entrenar modelos RAG requiere grandes cantidades de datos y recursos computacionales.
- **Integración con sistemas existentes** : la integración de modelos RAG con sistemas existentes puede resultar un desafío debido a las diferencias en los formatos de datos y las API.

Comprender los paradigmas RAG

La recuperación de generación aumentada (RAG) es una técnica poderosa que combina la recuperación de información con la generación de texto. Hay dos paradigmas principales para construir modelos RAG:

- Aumento de recuperación
- Sintonía FINA.

Aumento de recuperación

El aumento de recuperación implica agregar contexto al mensaje mediante canalizaciones de datos. El modelo recupera información relevante de una fuente de conocimiento y la utiliza para aumentar la indicación. Este enfoque es útil cuando la fuente de conocimiento es grande y diversa, y el modelo necesita recuperar información específica para generar un texto preciso e informativo.

Sintonía FINA

El ajuste implica incorporar conocimiento en los pesos de la red del modelo de lenguaje. Este enfoque es útil cuando la fuente de conocimiento es pequeña y específica, y el modelo necesita aprender a generar texto que sea coherente con la fuente de conocimiento. El ajuste se puede realizar de dos maneras: destilación de conocimientos y formación previa.

En la destilación de conocimiento, se utiliza un modelo previamente entrenado para generar texto y el resultado se utiliza para entrenar un modelo más pequeño. Este enfoque es útil cuando el modelo previamente entrenado es demasiado grande para usarlo en producción.

En el entrenamiento previo, el modelo se entrena en un corpus de texto grande y los pesos se ajustan en un corpus de texto más pequeño. Este

enfoque es útil cuando la fuente de conocimiento es pequeña y específica, y el modelo necesita aprender a generar texto que sea coherente con la fuente de conocimiento.

Componentes de la pila RAG

La pila RAG es una colección de herramientas y tecnologías que se utilizan para crear aplicaciones RAG listas para producción. La pila consta de varios componentes, que incluyen:

Ingestión de datos

La ingestión de datos es el proceso de recopilación y preparación de datos para su uso en aplicaciones RAG. Esto implica identificar fuentes de datos relevantes, extraer datos de esas fuentes y limpiar y formatear los datos para su uso en modelos RAG.

Consulta de datos

La consulta de datos es el proceso de recuperar datos de una fuente de conocimiento y utilizarlos para generar texto. Hay dos tipos principales de consulta de datos utilizados en las aplicaciones RAG: recuperación y síntesis.

La recuperación implica recuperar información relevante de una fuente de conocimiento y utilizarla para aumentar la indicación. Este enfoque es útil

cuando la fuente de conocimiento es grande y diversa, y el modelo necesita recuperar información específica para generar un texto preciso e informativo.

La síntesis implica generar nueva información a partir de una fuente de conocimiento y utilizarla para aumentar la indicación. Este enfoque es útil cuando la fuente de conocimiento es pequeña y específica, y el modelo necesita aprender a generar texto que sea coherente con la fuente de conocimiento.

Desafíos con Naive RAG

Si bien la recuperación de generación aumentada (RAG) ha revolucionado la forma en que generamos e interactuamos con el texto, no está exenta de limitaciones. Los sistemas RAG ingenuos pueden sufrir varios problemas que afectan la calidad del texto generado. Algunos de los desafíos incluyen:

- **Mala recuperación** : los modelos RAG se basan en modelos de recuperación para recuperar información relevante de una fuente de conocimiento. Si el modelo de recuperación no está bien diseñado, puede recuperar información irrelevante o incorrecta, lo que generará un texto de mala calidad.

- **Baja precisión** : los modelos RAG pueden generar texto objetivamente incorrecto o incompleto. Esto puede suceder si el modelo no está entrenado con datos de alta calidad o si los datos no son representativos del dominio de destino.
- **Eliminación de datos no conformes**: los datos y otros registros pueden eliminarse sin tener en cuenta las posibles consecuencias, como requisitos legales, reglamentarios o comerciales².
- **Alucinación** : los modelos RAG pueden generar texto que no está respaldado por la fuente de conocimiento. Esto puede suceder si el modelo no está bien diseñado o si la fuente de conocimiento es incompleta o inexacta.
- **Información desactualizada** : los modelos RAG pueden generar texto basado en información desactualizada. Esto puede suceder si la fuente de conocimiento no está actualizada o si el modelo no está diseñado para manejar información temporal.

Además de estos desafíos, los modelos RAG también pueden sufrir desafíos relacionados con el modelo de lenguaje (LM), como:

- **Alucinación** : los modelos basados en LM pueden generar texto que no es

compatible con el mensaje de entrada. Esto puede suceder si el modelo no está bien diseñado o si la solicitud de entrada es ambigua o incompleta.

- **Irrelevancia** : los modelos basados en LM pueden generar texto que es irrelevante para el mensaje de entrada. Esto puede suceder si el modelo no está bien diseñado o si el mensaje de entrada es demasiado general o demasiado específico.
- **Toxicidad** : los modelos basados en LM pueden generar texto ofensivo o dañino. Esto puede suceder si el modelo no está bien diseñado o si los datos de entrenamiento contienen contenido sesgado o tóxico.
- **Sesgo** : los modelos basados en LM pueden generar texto sesgado hacia ciertos grupos o perspectivas. Esto puede suceder si los datos de entrenamiento contienen contenido sesgado o si el modelo no está diseñado para manejar diversas perspectivas.

Para abordar estos desafíos, los investigadores están explorando nuevas técnicas como el entrenamiento adversario, el aprendizaje multitarea y el aprendizaje por transferencia. Estas técnicas pueden ayudar a mejorar la calidad de los modelos RAG y hacerlos más robustos y fiables.

Evaluación y Optimización de Sistemas RAG

La recuperación de generación aumentada (RAG) es una técnica poderosa para generar texto que sea fluido e informativo. Sin embargo, crear aplicaciones RAG listas para producción puede resultar un desafío debido a la complejidad de los modelos subyacentes y la necesidad de integrarlos con los sistemas existentes. A continuación se presentan algunas estrategias para evaluar y optimizar los sistemas RAG:

Evaluación en Benchmarking de Sistemas RAG

La evaluación es fundamental para comparar los sistemas RAG y garantizar que cumplan con los criterios de rendimiento deseados. Algunas de las métricas utilizadas para evaluar los sistemas RAG incluyen:

- **Precisión** : la proporción de información recuperada que es relevante para la consulta.
- **Recuperación** : la proporción de información relevante que recupera el sistema.
- **Puntuación F1** : La media armónica de precisión y recuperación.
- **Puntuación BLEU** : métrica utilizada para evaluar la calidad del texto

generado por máquina comparándolo con uno o más textos de referencia.

Optimización de los sistemas RAG

La optimización de los sistemas RAG implica el uso de una combinación de estrategias básicas, métodos de recuperación avanzados y arquitecturas basadas en agentes. Algunas de las técnicas utilizadas para optimizar los sistemas RAG incluyen:

- **Optimización de datos** : almacenamiento de información adicional, optimización de la representación incrustada y ajuste del tamaño de los fragmentos.
- **Métodos de recuperación avanzados** : recuperación de pequeño a grande, incorporación de referencias y filtrado de metadatos.
- **Aprovechamiento de modelos de lenguaje** : uso de modelos de lenguaje para razonamiento y resultados estructurados.

Estrategias de ajuste

El ajuste fino es una técnica que se utiliza para mejorar el rendimiento de los modelos RAG ajustando los pesos del modelo en función de datos de entrenamiento adicionales. Algunas de las estrategias utilizadas para ajustar

los modelos RAG incluyen:

- **Destilación de conocimiento** : usar un modelo previamente entrenado para generar texto y usar el resultado para entrenar un modelo más pequeño.
- **Entrenamiento previo** : entrene el modelo en un corpus de texto grande y ajuste los pesos en un corpus de texto más pequeño.
- **Transferir aprendizaje** : utilizar un modelo previamente entrenado para inicializar los pesos del modelo RAG y ajustar los pesos en la tarea objetivo.

Conclusión

En este artículo, exploramos el mundo de la generación aumentada de recuperación (RAG) y brindamos una descripción detallada de cómo crear aplicaciones RAG listas para producción. Discutimos los desafíos que se enfrentan en la construcción de sistemas RAG robustos y cómo superarlos. También proporcionamos una guía paso a paso para crear aplicaciones RAG listas para producción, incluida la preparación de datos, la capacitación de modelos y la integración con sistemas existentes.

También discutimos las limitaciones de los sistemas Naive RAG y las estrategias para mejorar el rendimiento de RAG en todo el proceso.

Exploramos la importancia de la evaluación en la evaluación comparativa de los sistemas RAG y las técnicas para optimizar los sistemas RAG, incluidas estrategias básicas, métodos de recuperación avanzados y arquitecturas basadas en agentes. También discutimos estrategias de ajuste para mejorar los modelos de integración y las capacidades de LM.

Alentamos a los desarrolladores a explorar y experimentar con técnicas RAG y a contribuir a la investigación y los avances continuos en la tecnología RAG. Con las herramientas y técnicas adecuadas, podemos crear potentes aplicaciones RAG que generen texto fluido e informativo.

Gracias por leer esta guía completa para crear aplicaciones RAG listas para producción. Espero que lo hayas encontrado informativo y útil. Si tienes alguna pregunta o si hay algo más en lo que pueda ayudarte, no dudes en preguntar.

Sí Si has disfrutado esta pieza, permíteme extenderme una cálida invitación a profundizar más en nuestra colección de historias.

¡“Descubrimiento observado” es todo lo que necesita! Explorando el futuro de la IA en operaciones

La exploración continua de MoE, RLHF, la complejidad de los modelos avanzados y los datos sintéticos resalta el compromiso de ampliar los límites de la IA. Al fomentar el desarrollo humano, el monitoreo centralizado y las prácticas responsables de datos, el futuro de los sistemas inteligentes parece brillante y lleno de posibilidades. Este vistazo a los avances en curso sirve como testimonio del continuo viaje de innovación en el mundo de la IA.

Lea aquí: <https://medium.com/@nilayparikh/observed-discovery-is-all-you-need-exploring-the-future-of-ai-in-ops-4d0f64bb4fa4>

Optimización de preferencias directas para modelos de lenguaje grandes: una mirada a su potencial

DPO está a la vanguardia en la configuración del futuro de los LLM, capacitando a los usuarios para que se conviertan en cocreadores en su viaje lingüístico. A medida que avanza la investigación y el desarrollo, DPO puede

desbloquear todo el potencial de los LLM, fomentando un futuro en el que la IA realmente hable nuestro idioma y comprenda nuestras necesidades únicas.

Lea aquí: <https://medium.com/ai-advances/direct-preference-optimization-for-large-language-models-a-look-at-its-potential-6f980fb8b0c9>

Acerca de ErgoSum / X Labs

Ergosum / X Labs es una consultora que ofrece servicios de diseño, arquitectura e investigación en el ámbito de la inteligencia artificial. Desde su creación en 2011, ha estado realizando investigaciones de campo sobre diversos temas, como modelos de lenguaje grandes, IA generativa, descubrimiento de contenido inteligente, AIOps y análisis de series temporales.

La firma cree en brindar una perspectiva abierta, basada en la investigación, basada en datos, democrática e imparcial como asesor confidencial de sus clientes.

sobre el autor

Siga el viaje en [el sitio web](#) , [el blog personal](#) , [LinkedIn](#) , [YouTube](#) , [Ergosum/X Labs](#) y [Medium](#) para mantenerse conectado y ser parte de la conversación en curso.

Recuperación aumentada

Inteligencia artificial

Llm

Llmops

Ai generativa



Escrito por Nilay Parikh

90 seguidores · Escritor para Avances de la IA

Seguir

Ingeniero especializado en MLOps, DevSecOps y Azure. Producción de ML/AI y experiencia en plataformas financieras.

[Ayuda](#) [Estado](#) [Acerca de](#) [Carreras](#) [Blog](#) [Privacidad](#) [Términos](#) [Texto a voz](#) [equipos](#)