Implementación de IA generativa en la retención y transferencia del conocimiento experto

Por ***Simón Marco, Alma Sinojmeri, José Marco Murabito, Marcela Zaira Mucci, Martín Rendichi, María Fernanda Santilli*** y ***Gerardo Emanuel Bagnati*** (Y-TEC).

La inteligencia artificial generativa puede convertirse en una aliada clave para preservar y transferir el conocimiento experto en la industria energética. Este trabajo explora su aplicación concreta en la temática de daño de formación, comparando el desempeño de dos versiones de ChatGPT. El objetivo: potenciar la toma de decisiones técnicas y mejorar la formación de nuevos profesionales.

Este trabajo fue seleccionado en las 3º Jornadas de Revolución Digital para Petróleo y Gas.

En este trabajo se ha presentado una comparación entre dos modelos de generación de lenguaje natural basados en ChatGPT (en sus versiones 3.5 Turbo y 4.0), aplicados a la tarea de responder consultas sobre Daño de Formación. El daño de formación es un fenómeno que afecta negativamente a la producción de pozos convencionales y *Tight*. Existen múltiples mecanismos físicos, químicos, térmicos o incluso biológicos que pueden dar origen a un daño de formación e impactar en diferentes magnitudes individualmente, o suceder varios en simultaneo, por lo que suele ser una temática compleja de abordar. Algunos de estos mecanismos ocurren de manera inevitable, pero otros pueden deberse a decisiones de diseño/planificación o errores durante las operaciones de perforación, terminación, *workover* y/o estimulación.

Sin dudas, la ocurrencia de muchas de las pérdidas de producción asociadas a daño de formación está relacionada con la experiencia y el expertise del profesional a cargo de tomar decisiones. El acceso rápido al conocimiento experto potenciará al profesional de menor expertise ayudándolo a prevenir la generación de daños o diseñar tratamientos de mitigación más efectivos. El objetivo principal de este trabajo es crear una herramienta digital que permita principalmente:

- Retener de manera permanente el conocimiento experto y volcarlo en una herramienta digital que facilite el acceso al mismo.
- Acompañar personal de menor expertise en la toma de decisiones.
- Reforzar el modelo de formación activa *On The Job* (Modelo 70-20-10).

Como solución tecnológica se optó por el uso de inteligencia artificial generativa, permitiendo abaratar costos y disminuir tiempos de desarrollo e implementación. Específicamente se trabajó con la arquitectura RAG (Retriever-Augmented Generative) que incluye el modelo de OpenAI y permite la ingesta de información adicional, sin descuidar aspectos relacionados a ciberseguridad, gobierno y legales.

Se han evaluado diferentes aspectos relacionados con la calidad, la fiabilidad, la velocidad y la relevancia de las respuestas generadas por ambos modelos, y los resultados han mostrado que el modelo ChatGPT 4.0 supera al modelo ChatGPT 3.5 Turbo en la mayoría de los aspectos evaluados, excepto en la velocidad de respuesta, donde el modelo más antiguo presenta una mayor fluidez.

Introducción

El daño de formación es uno de los problemas más significativos que afectan la productividad de los pozos de petróleo y gas, tanto en reservorios convencionales como en Tight. Este fenómeno altera las propiedades naturales de la roca circundante al pozo, incluyendo la permeabilidad y la capilaridad, lo que a su vez disminuye la eficiencia del flujo de fluidos desde y hacia el reservorio. Múltiples mecanismos físicos, químicos, térmicos e incluso biológicos pueden desencadenar el daño de formación, actuando de forma individual o combinada según las características específicas del yacimiento, la composición de la roca y los procedimientos operativos implementados.

El diagnóstico, la prevención y la mitigación del daño de formación requieren de un análisis integral y multidisciplinario, que involucre el conocimiento de la geología, la petrofísica, la ingeniería de yacimientos, la ingeniería de producción, la química y la mecánica de rocas. Sin embargo, este análisis muchas veces se ve dificultado por la escasez, la incertidumbre o la inaccesibilidad de los datos necesarios para caracterizar el sistema pozo-reservorio. Además, no existe una metodología estándar o universal para abordar el problema del daño de formación, sino que cada caso debe ser estudiado de forma particular, considerando las condiciones específicas y las hipótesis formuladas. Esto implica que la experiencia y el criterio del profesional a cargo son factores

clave para tomar decisiones acertadas y optimizar la producción de los pozos.

No obstante, el sector petrolero y gasífero enfrenta el desafío de la rotación acelerada del personal, debido al recambio generacional, la migración a otras compañías o la rotación interna a otras áreas. Esto implica la pérdida del conocimiento experto y la dificultad para capacitar y acompañar al personal de menor experiencia. Asimismo, la formación académica en las carreras afines al sector suele ser insuficiente o poco profunda en lo que respecta al daño de formación y la estimulación matricial, lo que genera una brecha entre la teoría y la práctica. Por otro lado, los objetivos orientados a reducción de costos y tiempos, así como otros objetivos que no necesariamente priorizan la producción final del pozo, pueden afectar negativamente la calidad y la cantidad de información generada para el análisis del daño de formación, impactando en la certeza del diagnóstico necesario para la correcta prevención o mitigación del daño.

En este contexto, surge la necesidad de desarrollar una herramienta digital que permita retener y transferir el conocimiento experto sobre el daño de formación, así como asistir y capacitar al personal de menor experiencia en la toma de decisiones. Para ello, se propone el uso de inteligencia artificial generativa (IAG), una rama de la inteligencia artificial que se dedica a generar contenidos de forma automática, imitando el estilo y el lenguaje humano. La IAG puede aprovechar el gran volumen de información disponible en diferentes formatos y fuentes para crear contenidos originales, coherentes y relevantes para una determinada tarea o consulta. La IAG tiene aplicaciones potenciales en diversos campos, como la educación, el periodismo, la publicidad, el entretenimiento o la medicina (Jarrahi et al., 2023) (Dam et al., 2024).

En este trabajo, se utiliza la arquitectura RAG (Singh et al., 2024), que combina el modelo de OpenAI GPT-3.5 Turbo y GPT-4.0 con un mecanismo de recuperación y atención de documentos, para generar respuestas a preguntas relacionadas con el daño de formación. El modelo se entrena con datos provenientes de diversas fuentes, como libros, artículos, informes, presentaciones, manuales, protocolos, etc., que abarcan distintos aspectos del daño de formación y la estimulación matricial. El modelo se implementa en una plataforma web, que permite al usuario interactuar con el sistema mediante un chatbot, que simula una conversación con un experto en la materia. El sistema es capaz de responder a consultas específicas, proporcionar información general, sugerir recomendaciones, mostrar ejemplos, generar tablas comparativas o de resumen, entre otras funcionalidades.

Para garantizar la calidad y fiabilidad del contenido que nutre las respuestas de la IAG, se ha conformado un equipo de expertos en la temática de daño de formación. Este grupo tiene acceso a una base documental o repositorio, donde puede actualizar o eliminar los documentos que no sean pertinentes, actuales o exactos. De esta forma, la herramienta dispone de información validada y revisada por los especialistas, lo que facilita su uso a los usuarios de la plataforma. Con este procedimiento, se pretende reducir las brechas generacionales existentes en los métodos tradicionales de transferencia de conocimiento, modernizar los medios de acceso a la información y optimizar el tiempo de los profesionales delegando a la IA la búsqueda de respuestas entre miles de documentos.

Metodología

La solución propuesta se basa en el uso de ChatGPT (GPT-3.5 Turbo y GPT-4.0) de la compañía OpenAI. La elección de este modelo se basó en la rapidez con que se podía desplegar los componentes en la nube, fácil configuración y soporte técnico y funcional. Además, se buscaba transparencia y manejo de una IA responsable. También es importante considerar el tiempo que la industria dispone para desarrollar una Prueba de Concepto y su posterior puesta a disposición de los usuarios. Contar con un modelo entrenado y validado ofrece ventajas (tiempos, costos y menor esfuerzo) sobre otros que necesitan de un "tunning" previo, es decir, personalización y optimización de un LLM para que responda de manera efectiva a las necesidades y expectativas del usuario final.

Se implementó una arquitectura del tipo RAG (detallada anteriormente) que desplegó en la nube los siguientes recursos:

- Repositorio de documentación: almacenamiento privado, que permitió incorporar información curada de la compañía. Este repositorio contiene dos carpetas, en una de ellas se encuentran los documentos que la herramienta permite reproducir y mostrar al pulsar una cita de la respuesta, y en la otra carpeta corresponde a documentos que por confidencialidad o copyright no tiene permitido reproducir. En estos casos, el usuario deberá acceder al documento por sus medios, respetando los permisos de acceso heredados de su perfil en la compañía. Los documentos se almacenan en un primer repositorio y luego pasan a las diferentes carpetas del almacenamiento privado.
- Base de Datos no relacional: almacenamiento, trazabilidad de preguntas y repuestas por usuario, tiempos de respuestas, datos de auditoría y otros de carácter analítico.
- Recursos para el manejo de índices: los documentos almacenados se separan en partes o bloques que luego son indexados y de esta manera fácilmente accesibles por los demás recursos.
- Componente OpenAI: despliegue del modelo dentro de la arquitectura cloud.
- Prompt: instrucciones que permiten configurar estilos y formas de respuestas, así como también bloquear palabras y/o temáticas que contenga la consulta y rechazarla en el chat.
- FrontEnd: para el manejo de la interacción entre usuario y el chat, como historial de últimas conversaciones, posibilidad de calificar respuestas con escala de 5 estrellas, citas con link a cada página de los documentos utilizados para elaborar esa respuesta, panel de visualización de dichos documentos dentro del propio módulo de chat, y preguntas sugeridas para continuar la conversación.

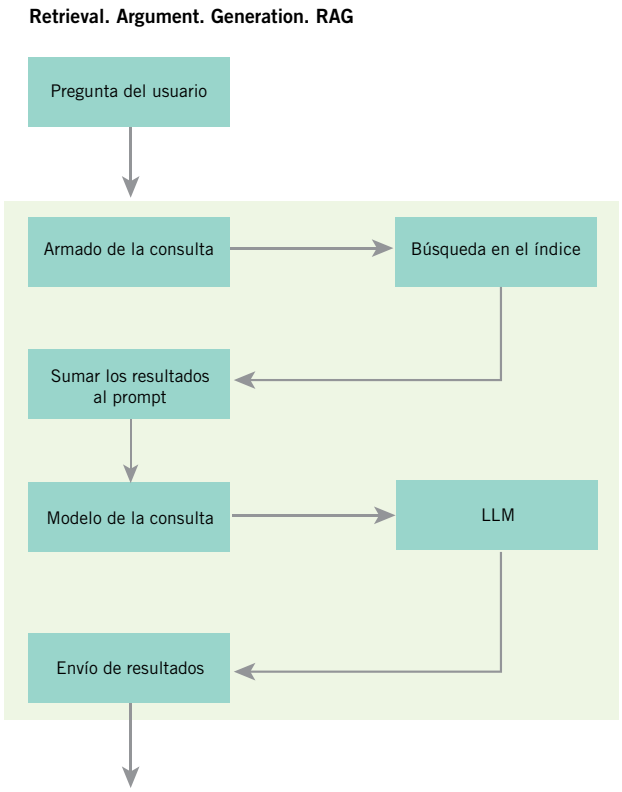

Figura 1. **Circuito que sintetiza la "conversación" que el usuario tiene sobre la información corporativa de la organización.**

- Arquitectura Zero Trust: componentes que aseguraron un alto nivel de seguridad y aislamiento de la red debido a que se requería un adecuado entorno de producción, donde las estrictas medidas de seguridad son primordiales (Rose et al., 2020). Este modelo de arquitectura asegura que cada solicitud esté autenticada y autorizada, lo que reduce significativamente el riesgo de amenazas internas. Asimismo, garantiza una comunicación segura, aislada y eficiente entre los componentes, siguiendo un modelo de confianza cero y de capa cerrada de redes.

En la Fig. 1, se representa el circuito que se sigue desde el momento que el usuario realiza una consulta, busca en el índice la información correspondiente en el repositorio de información y ese resultado se agrega a la consulta del prompt que fue configurado previamente. Con esa información se llama al servicio de OpenAI que devuelve la respuesta al chat en el FrontEnd.

Cada respuesta provista en el chat puede ser evaluada por los usuarios y calificada en cuanto a su nivel de certeza, duda o precisión. Estas respuestas, en conjunto con sus calificaciones, se almacenan en una base de datos interna para su posterior evaluación por parte de un experto en la materia. A su vez, la trazabilidad posterior en cuanto a la repetición de una consulta es un indicador interno para detectar puntos de debilidad en el "know-how" general y contemplar capacitaciones específicas relacionadas con ese tema.

Gobierno de Datos y Ciberseguridad

Debido al manejo sensible de la información y al cumplimiento estricto de las normas establecidas a nivel gobierno corporativo, es que se tuvieron en cuenta los siguientes puntos que se explican a continuación:

- Políticas de uso: se estableció un marco normativo que garantizó el acceso a datos curados, gestión del costo y la calidad de la operación de la información. Fue crucial la definición del uso coherente, integrado y controlado de la información de la organización.
- Marco legal: cada unidad de negocio se responsabilizó por definir, catalogar y asegurar la calidad de datos bajo su dominio, tanto propio como de terceros de acuerdo con las legislaciones vigentes.
- Monitoreo y métricas: Se evaluó el rendimiento de los modelos de LLM mediante métricas objetivas como la precisión, repetitividad, diversidad, ocurrencia de alucinaciones (respuesta confiada que no está justificada por los datos con los que ha sido entrenada) y relevancia de las respuestas generadas, así como métricas subjetivas como la satisfacción y confianza de los usuarios.
- Ataques del tipo inyección: Se evaluaron casos donde se pudieran insertar entradas maliciosas para alterar el comportamiento del modelo, así como también la extracción de información sensible.

Resultados

En este trabajo se realizó una evaluación comparativa entre dos modelos de generación de lenguaje natural basados en GPT-3: ChatGPT 3.5 Turbo y ChatGPT 4.0. El objetivo fue medir el desempeño de ambos modelos en la tarea de responder preguntas sobre información corporativa de la organización, utilizando como fuente de datos un repositorio de documentos oficiales. Para ello, se consideraron los siguientes aspectos:

- Velocidad de respuesta: el tiempo que tarda el modelo en generar una respuesta a una consulta del usuario.
- Calidad y completitud de la respuesta: el grado de precisión, coherencia, relevancia y exhaustividad de la respuesta generada por el modelo, en relación a la consulta y a la información disponible en los documentos.
- Ocurrencia de alucinaciones: la frecuencia con la que el modelo genera información falsa, inventada o no respaldada por los documentos.
- Confusión de temas: la frecuencia con la que el modelo mezcla o cambia de tema en la respuesta, sin atender a la consulta o al contexto de la conversación.
- Interpretación de la información proveniente de los documentos: la capacidad del modelo para procesar, comprender y extraer información relevante de los documentos.
- Cambios de respuesta ante la repregunta: la consistencia del modelo al responder a la misma consulta

o a una consulta similar, en diferentes ocasiones o con diferentes formulaciones.

- Preguntas fuera de los temas permitidos: la forma en que el modelo maneja las consultas que no están relacionadas con la información corporativa de la organización, o que violan las políticas de uso o el marco legal establecido.
- Búsqueda de respuestas en tablas: la capacidad del modelo para localizar, interpretar y sintetizar información proveniente de tablas presentes en los documentos.
- Generación de tablas: la capacidad del modelo para crear tablas a partir de la información disponible en los documentos, o a partir de la combinación de diferentes fuentes de datos.
- Reconocimiento y representación de ecuaciones químicas: se refiere a la capacidad del modelo para identificar, reproducir y presentar adecuadamente las ecuaciones químicas contenidas en los textos.
- Manejo de símbolos especiales (como subíndices, superíndices y el alfabeto griego): hace referencia a la habilidad del modelo para identificar, reproducir y presentar de manera correcta los caracteres especiales frecuentemente usados en documentos, en particular dentro de fórmulas químicas y ecuaciones matemáticas.
- Reproducción correcta de ecuaciones matemáticas: describe la capacidad del modelo para detectar, duplicar y mostrar con exactitud las ecuaciones matemáticas incluidas en los documentos.

Los resultados de la evaluación mostraron que ChatGPT 4.0 superó ampliamente a ChatGPT 3.5 Turbo en todos los aspectos, excepto en la velocidad de respuesta. ChatGPT 4.0 demostró una mayor calidad y completitud de las respuestas, una menor ocurrencia de alucinaciones, una mejor interpretación de la información proveniente de los documentos, una mayor consistencia ante la repregunta, una mejor gestión de las preguntas fuera de los temas permitidos, una mayor habilidad para buscar respuestas en tablas y generar tablas, y una mejor reproducción de ecuaciones de reacción química y caracteres especiales. La capacidad de interpretar y reproducir ecuaciones matemáticas sigue siendo un desafío no superado por ambos modelos. Cabe aclarar que esta situación se ha observado en el contexto de modelos aplicados dentro de la arquitectura RAG, operando sobre documentación indexada (mayoritariamente documentos en formato .pdf), lo cual podría variar en cuanto a respuestas generadas en versiones libres que operan fuera de ambientes corporativos.

Un aspecto diferencial de ChatGPT 3.5 Turbo fue su velocidad, lo que generaba una mayor fluidez en la interacción con el usuario, pero a costa de una menor calidad y fiabilidad en las respuestas. Cabe señalar que, en determinados momentos, el modelo ChatGPT 4.0 ha presentado tiempos de espera superiores al minuto, lo cual puede afectar negativamente a la motivación del usuario para continuar el diálogo.

La versión avanzada de ChatGPT 4.0 se destaca por ser un modelo más desarrollado y complejo, reflejado en las optimizaciones en la arquitectura, el incremento en su tamaño, las mejoras metodológicas durante el entrenamiento, la calidad del conjunto de datos utilizado y el refinamiento posterior. Estas mejoras le permiten al modelo captar mejor las relaciones semánticas y sintácticas entre las palabras, las frases y los párrafos, así como generar textos más naturales, fluidos y coherentes. Además, ChatGPT 4.0 tiene una mayor capacidad para manejar información estructurada, como tablas y fórmulas, y para adaptarse al dominio y al contexto de la conversación.

Algunos de los aspectos evaluados se pudieron solucionar o mejorar con modificaciones en el prompt, es decir, en la forma en que se formula la consulta al modelo. Por ejemplo, la velocidad de respuesta se puedo aumentar levemente limitando el número de tokens o palabras que el modelo puede generar, o indicando al modelo que termine la respuesta cuando encuentre un punto final. La calidad y completitud de la respuesta se puedo mejorar proporcionando al modelo más información sobre el contexto, el propósito y el formato de la respuesta esperada, o utilizando palabras clave o frases específicas que orienten al modelo hacia la información relevante. Los cambios de respuesta ante la repregunta se pudieron minimizar utilizando un prompt que sea consistente y claro, o que le pida al modelo que confirme o corrija su respuesta anterior. Las preguntas fuera de los temas permitidos se pueden manejar incorporando en el prompt una lista de temas permitidos y no permitidos, indicándole al modelo que no puede responder a ese tipo de consultas, o redirigiendo al usuario a otro canal de atención.

Otros aspectos evaluados requieren de mejoras más profundas en el modelo, que no se pueden solucionar solo con modificaciones en el prompt. Por ejemplo, la interpretación de la información proveniente de los documentos se puede mejorar con un mayor preprocesamiento y normalización de los datos, o con una mayor integración entre el modelo de generación de lenguaje natural y el modelo de comprensión de documentos. Otro ejemplo son las dificultades que se a veces presentan en la búsqueda e interpretación de contenido tabulado. Los documentos que se utilizan como fuente de información están indexados y divididos en partes más pequeñas, llamadas chunks, que facilitan su procesamiento por parte del modelo. Sin embargo, esta fragmentación puede provocar que algunas tablas se corten entre dos o más chunks, lo que impide al modelo acceder a toda la información que contiene la tabla y, por lo tanto, afecta la precisión y la completitud de la respuesta.

Los avances en la interpretación y representación de ecuaciones químicas, caracteres especiales (como subíndices, superíndices y el alfabeto griego) y ecuaciones matemáticas forman parte del desarrollo continuo y evolutivo del modelo. Según nuestra experiencia, ChatGPT 3.5 gestionó adecuadamente los subíndices y superíndices; ChatGPT 4.0 avanzó incluyendo la capacidad de manejar ecuaciones químicas y el alfabeto griego; y ninguno de ellos pudo procesar ecuaciones matemáticas

con eficacia, aunque modelos más recientes y perfeccionados están empezando a tener esta capacidad.

En cuanto al motivo de la presencia de alucinaciones en los modelos que se utilizaron, no existe un consenso definido al respecto ya que las redes neuronales que se utilizan tienen procesos en los que podemos saber su entrada y salida, pero poco se publica en forma específica sobre su funcionamiento interno. Algunos investigadores expresan que el origen de las alucinaciones se produce por material inexacto que se encuentra en el conjunto de datos con el que fueron entrenados o bien por las inferencias que hacen los modelos de situaciones específicas que no se encuentran en su material de entrenamiento. Con lo cual, cuando el chat alucina es porque está buscando información o análisis que no está presente en el texto de datos con el que lo han alimentado y que para rellenar los "espacios en blanco" lo hace con palabras que suenen bien, aunque resulten inexactas (Zhang et al., 2023) (Rawte et al., 2023). Por esta razón, el modelo ChatGPT 4.0 presenta una menor ocurrencia de alucinaciones y mejor performance en los demás aspectos detallados que el modelo ChatGPT 3.5 Turbo, ya que se trata de un modelo mejorado, con un nivel de entrenamiento superior y se ha utilizado un conjunto de datos más amplio y diverso (Koubaa, 2023).

Estos resultados demuestran la buena implementación y despliegue de ambos modelos dentro de la herramienta, que actualmente se encuentra en fase de testeo con una célula de usuarios (n 30), donde se está evaluando el grado de satisfacción, la ocurrencia de errores y alucinaciones en las respuestas, dificultades de acceso a la herramienta, aspectos relacionados con la UX y el FrontEnd y otras métricas. Mientras siga activa la fase de testeo, la herramienta cuenta con un switch que permite a los usuarios elegir qué modelo usar (ChatGPT 3.5 Turbo o ChatGPT 4.0) y comparar las respuestas.

Este testeo nos permitirá recoger más feedback y mejorar aún más el rendimiento y la calidad de los modelos, así como detectar posibles áreas de mejora en la herramienta y en la experiencia de usuario. Por ahora, el testeo muestra un alto grado de satisfacción general.

Conclusiones

En este trabajo se ha presentado una comparación entre dos modelos de generación de lenguaje natural basados en ChatGPT, aplicados a la tarea de responder consultas sobre información corporativa de una organización. Se han evaluado diferentes aspectos relacionados con la calidad, la fiabilidad, la velocidad y la relevancia de las respuestas generadas por ambos modelos, utilizando como fuente de información un conjunto de documentos indexados por una plataforma basada en Retriever-Augmented Generative (RAG). Los resultados han mostrado que el modelo ChatGPT 4.0 supera al modelo ChatGPT 3.5 Turbo en la mayoría de los aspectos evaluados, excepto en la velocidad de respuesta, donde el modelo más antiguo presenta una mayor fluidez.

Estos resultados demuestran el potencial de la inteligencia artificial generativa para facilitar la retención y la transferencia del conocimiento experto dentro de una organización, así como para mejorar la experiencia y la satisfacción de los usuarios que requieren acceder a esa información. Uno de los beneficios clave de integrar ChatGPT con una plataforma basada en RAG es que evita la necesidad de desarrollar y entrenar un modelo propio, un proceso que sería intensivo en recursos humanos, tiempo de procesamiento y costos computacionales considerables. Además, al utilizar RAG, se puede generar respuestas más detalladas y pertinentes que con métodos de generación tradicionales, ya que el modelo puede aprovechar el contenido de los documentos para construir sus respuestas. Sin embargo, el uso de RAG también implica algunos desafíos, como la necesidad de contar con un repositorio de datos bien curado y supervisado para agregar información nueva, además de los costes asociados con el mantenimiento de dicha infraestructura y los tiempos de respuesta incrementados

Agradecimientos

Los autores queremos expresar nuestro agradecimiento a Y-TEC por brindarnos la oportunidad de innovar en esta disciplina, así como a los equipos de Upstream, Tecnologías Digitales, Legales, Compliance y otros de YPF por el apoyo y la confianza que nos han otorgado y por permitirnos publicar este trabajo. También agradecemos a Microsoft y Pi Consulting por sus valiosas colaboraciones y orientaciones durante el desarrollo de este proyecto.

Referencias

Dam S. K., Hong C. S., Qiao Y., Zhang C. (2024). arXiv:2406.16937v1 [cs.CL] - "A Complete Survey on LLM-based AI Chatbots"

Jarrahi M. H., Askay D., Eshraghi A., Simth P. (2023). Business Horizon Journal 66, pag 87-99 - "Artificial intelligence and knowledge management: A partnership between human and AI"

Koubaa A. (2023). Preprints 2023030422 - "GPT-4 vs. GPT-3.5: A Concise Showdown"

Rawte V., Sheth A., Das A. (2023) arXiv:2309.05922v1 [cs.AI] - "A Survey of Hallucination in "Large" Foundation Models"

Rose S., Borchert O., Mitchell S., Connelly S. (2020). NIST Special Publication 800-207 - "Zero Trust Architecture"

Singh K. P., Kumar P. (2024). IJETA Journal ISSN:2393-9516 - "BIRAG: Basic Introduction to Retrieval Augmented Generation"

Zhang Y., Li Y., Cui L., Cai D., Liu L., Fu T., Huamg X., Zhao E., Zhang Y., Chen Y., Wang L., Luu A. T. Bi W. Shi F., Shi S. (2023) arXiv:2309.01219v2 [cs. CL] - "Siren's song in the AI ocean: a survey on hallucination in large language models"