REVISTA CIENTÍFICA

Vol. 21, No.2 (2025) Monográfico ISSN -E 1683-8947

bibliotecas anales de investigación

BIBLIOTECA NACIONAL DE CUBA JOSÉ MARTÍ

INTELIGENCIA ARTIFICIAL GENERATIVA: ANÁLISIS CONCEPTUAL Y PRÁCTICO DEL MODELO GENERACIÓN AUMENTADA DE RECUPERACIÓN

GENERATIVE ARTIFICIAL INTELLIGENCE: CONCEPTUAL AND PRACTICAL ANALYSIS OF THE RETRIEVAL-AUGMENTED GENERATION MODEL

Angel Freddy Godoy Viera
Universidade Federal de Santa Catarina
a.godoy@ufsc.br
https://orcid.org/0000-0001-6657-4734

José Antonio Moreiro-González
Universidad Carlos III de Madrid, España
joseantonio.moreiro@uc3m.es
https://orcid.org/0000-0002-8827-158X

Recibido: 1 de julio de 2025
Revisado: 7 de agosto de 2025
Aprobado: 19 de septiembre de 2025

Cómo citar: Godoy Viera, A. F. y Moreiro González, J. A. (2025). Inteligencia Artificial Generativa: análisis conceptual y práctico del Modelo Generación Aumentada de Recuperación. *Bibliotecas. Anales de Investigacion;21*(2) Monográfico, 1-14

RESUMEN

Objetivo: Para revisar sistemáticamente el estado de las aplicaciones del modelo de generación aumentada de recuperación (RAG) para la generación de texto se pretende como objetivo principal, hacer un levantamiento diagnóstico, con los propósitos secundarios de identificar las formas de difusión preferidas por los autores, revelar los temas abordados, las entidades financiadoras, además de establecer las redes de coautoría. **Diseño/Metodología/Enfoque:** Por su metodología es una investigación exploratoria-descriptiva y bibliográfica. **Resultados/Discusión:** Los resultados muestran que los autores prefieren publicar sus trabajos en congresos, la institución que financia más investigaciones es la *National Natural Science Foundation of China*, mientras que en Estados Unidos se produce el mayor número de publicaciones, predominan los autores afiliados a universidades y los temas de investigación más atendidos son los sistemas de grandes modelos de lenguaje, la respuesta a preguntas, la generación de texto y la elaboración de grafos de conocimiento. **Conclusiones:** En el campo científico de la información, el modelo RAG, se relaciona con la recuperación de información y con el procesamiento del lenguaje natural. Supera las limitaciones de grandes modelos de lenguaje, al agregarles información específica y actualizada de fuentes externas. Tiene grandes

potencialidades de aplicación en las investigaciones de las humanidades digitales, por aprovechar el conocimiento aprendido en los parámetros de los grandes modelos de lenguaje pre-entrenados con informaciones de repositorios específicos que podrían ser bases de datos de documentos históricos, o de bibliotecas que pueden ser utilizados para crear nuevos servicios. **Originalidad/Valor:** Ofrece un primer panorama sobre modelo RAG, permitiendo que el profesional de la información entienda esa importante tecnología de inteligencia artificial generativa, además de conocer sus principales financiadores de las investigaciones, las redes de autoría entre las organizaciones y sus principales aplicaciones en la generación textual.
PALABRAS CLAVE: generación aumentada de recuperación, grandes modelos de lenguaje, generación de texto, inteligencia artificial generativa, revisión sistemática.

ABSTRACT

Objective: To systematically review the status of applications of the retrieval-augmented generation model (RAG) for text generation, the main objective is to make a diagnostic survey, with the secondary purposes of identifying the forms of dissemination preferred by the authors, revealing the topics addressed, the funding entities, and establishing co-authorship networks. **Design/Methodology/Approach:** Due to its methodology, it is an exploratory-descriptive and bibliographical research. **Results/Discussion:** The results show that authors prefer to publish their work in conferences, the institution that funds the most research is the National Natural Science Foundation of China, while in the United States the largest number of publications are produced, authors affiliated with universities predominate, and the most attended research topics are large language model systems, question answering, text generation, and the development of knowledge graphs. **Conclusions:** In the scientific field of information, the RAG model is related to the information retrieval and natural language processing. It overcomes the limitations of large language models by adding specific and updated information from external sources. It has great potential for application in digital humanities research, by taking advantage of the knowledge learned in the parameters of pre-trained large language models with information from specific repositories that could be databases of historical documents, or libraries that can be used to create new services. **Originality/Value:** It offers a first overview of the RAG model, allowing information professionals to understand this important generative artificial intelligence technology, in addition to knowing its main research funders, the authorship networks between organizations and its main applications in text generation.
KEYWORDS: retrieval-augmented generation, large language models, text generation, generative artificial intelligence, systematic review.

INTRODUCCIÓN

Se hace necesario atender al comportamiento de las aplicaciones de inteligencia artificial (IA) debido al gran interés que están suscitando en ámbitos tan diversos como la industria, las empresas, los gobiernos o las universidades, pues contribuyen a automatizar diversos tipos de procesos y servicios. Después del primer impacto causado por la divulgación impactante de los sistemas de IA, principalmente de los que utilizan grandes modelos de lenguaje preentrenados (*pre-trained large language models* - *PLLM*), es acertado hacer un seguimiento de la investigación dedicada en empresas e instituciones académicas a mejorar los sistemas de IA para impulsar y mejorar los *chatbots* que interactúan con los usuarios y de ofrecer respuestas, desde un comportamiento cercano al humano, a los diversos tipos de preguntas que se les formulen. Al aumentar el uso de las aplicaciones de IA se evidencian algunos problemas que se recogen en los artículos académicos y en los informes técnicos empresariales. En ellos se identifican, entre otros, los fallos a la hora de responder en forma satisfactoria a *prompts* de dominios específicos, la falta de actualidad en muchas respuestas o la generación de respuestas alucinantes y sin fundamento como si fuesen verdaderas (Mao et al., 2020; Shuster et al., 2021; Zhao et al., 2024).

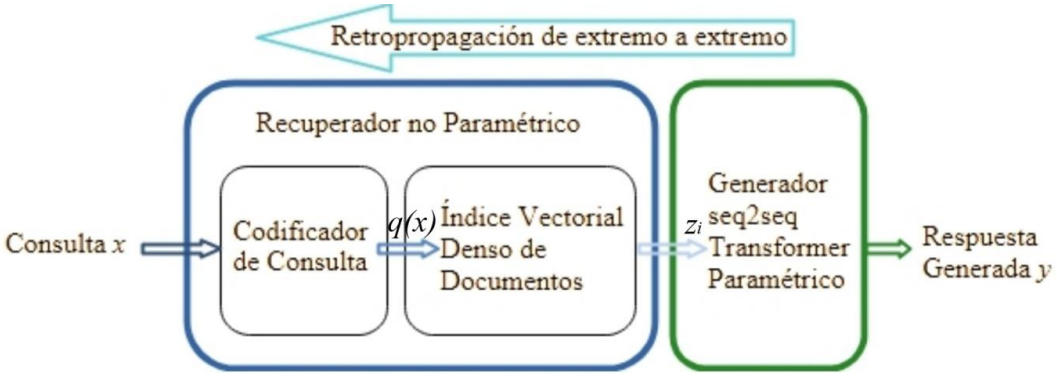
En este sentido, Yu, W. (2022) menciona el alto coste que supone entrenar estos sistemas de IA mediante grandes modelos de lenguaje preentrenados (PLLM). Lo que deja en manos de muy pocas empresas la capacidad de ofrecer esos modelos de manera gratuita a los usuarios y organizaciones, suponiendo un serio problema para quienes tienen recursos más limitados y no pueden afrontar el pago. Por ese motivo surge la

necesidad de buscar la forma de resolver o minimizar los problemas que resultan de utilizar modelos neurales de lenguaje preentrenados. En ese contexto surge la propuesta de un nuevo modelo llamado Generación aumentada de recuperación (*Retrieval-augmented Generation)*, RAG por su sigla en inglés. El objetivo principal es valorar un levantamiento bibliográfico sobre RAG cuando se aplica a la generación de textos. Como objetivos específicos se busca identificar y contrastar los títulos y las formas de comunicación preferidas por los autores, junto a los temas abordados, sus aplicaciones y financiadores; así como establecer las redes de coautoría tanto por autores como por organizaciones de pertenencia.

CONSIDERACIONES TEÓRICO-PRÁCTICAS AL MODELO RAG

El término generación aumentada de recuperación (RAG) fue introducido por Patrick Lewis et al. (2020) en cuanto planteamiento de ajuste fino y de propósito general que proporciona una memoria no paramétrica a un modelo generativo de memoria paramétrica preentrenada. Proponen además la arquitectura del modelo RAG que se compone de un recuperador de pasajes densos (*Dense Passage Retriever* – DPR) preentrenado, la memoria no paramétrica, así como de un Codificador de consulta y un índice vectorial denso de documentos Wikipedia. Por su parte, la memoria paramétrica integra un modelo generador *seq2seq transformer* preentrenado con afinación de extremo a extremo. Resaltan que para una consulta determinada (x) el modelo RAG utiliza el *Maximum Inner Product Search* (MIPS) para encontrar los k documentos del tope del rango z_i. Para realizar la predicción final y se trata z como si fuese una variable latente cuyos documentos están marginalizados por la predicción del generador seq2seq.

Figura 1. *Arquitectura RAG propuesta por Lewis et al. (2020)*



Fuente:, *adaptada de Lewis et al. (2020)*

El artículo de Lewis et al. (2020), al presentar un sistema de IA generativa que usa el contexto para proporcionar respuestas acordes con las consultas, tuvo continuidad inmediata mediante el desarrollo de modelos que utilizan RAG en las aplicaciones más variadas. Así, Zhao et al. (2024) consideran que incorpora el tratamiento de recuperación de la información a fin de mejorar la generación de contenido con objetos relevantes de repositorios disponibles, por lo que aumenta la precisión y robustez del proceso de generación. Destacan que el componente recuperador del RAG puede aprovecharse, primero, como recuperador poco denso que utiliza técnicas de recuperación de información, como los índices invertidos, para mejorar la eficiencia; luego para representar las consultas y las palabras claves mediante vectores *embedding* (*dense embedding vectors*) y elaborar índices de aceleración de búsqueda del vecino más cercano (*approximate nearest neighbor - ANN)*; por fin como otros tipos de recuperador: los que utilizan directamente la distancia de edición entre los textos en lenguaje natural, los grafos de conocimiento en que las entidades se vinculan con sus relaciones y la recuperación por reconocimiento de entidades nominadas (*Named entity recognition – NER*). Mientras que según Yu, W. (2022) es un método que une modelos de lenguaje preentrenados con técnicas tradicionales de recuperación de la información alcanzando un buen desempeño en muchas de las tareas intensivas del Procesamiento del lenguaje natural (PLN). Al considerar las ventajas del modelo RAG (Lewis et al., 2020) encontraron que en las pruebas del modelo se prefieren las respuestas de un sistema de respuesta de preguntas (QA por su sigla en inglés) con evaluadores humanos ante las respuestas de un generador de respuesta puramente paramétrico como BART. Ya que al utilizar el modelo RAG se generan

respuestas basadas en hechos contrastados con datos muy recientes, lo que origina menos alucinaciones, ofrece mayor control y menor desfase, a la vez que ayuda a interpretar mejor los textos. Se comprueba además que el índice de recuperación utilizado en el modelo RAG puede ser cambiado en caliente para actualizar el modelo sin necesidad de realizar un nuevo entrenamiento. Otra ventaja está en poder utilizar sistemas AI con modelo RAG para combatir los contenidos engañosos y la generación automática de *spam/phishing* (Lewis et al., 2020). RAG emerge como un paradigma para abordar los problemas de la generación de contenido por inteligencia artificial (AIGC, por su sigla en inglés) a fin de actualizar el conocimiento aprendido, de evitar la fuga de datos y de gestionar el alto coste del entrenamiento y de la inferencia (Zhao et al., 2024). Para Yu, W. (2022) el modelo RAG, comparado con los grandes modelos de lenguaje preentrenados, presenta notables beneficios a la hora de adquirir el conocimiento de forma explícita y no acumularlo de forma implícita en los parámetros del modelo entrenado; o el de generar salidas desde algunas de las referencias recuperadas facilitando el proceso de generación.

Algunas de las desventajas encontradas en el modelo RAG se deben a que incorpora fuentes de información que, en ocasiones, no se atienen a hechos reales o que no son del todo carentes de sesgos. Además, como cualquier modelo de lenguaje, puede usarse para generar contenido engañoso, falso o abusivo en las noticias o en las redes sociales, para automatizar la producción de contenido *spam/phishing* o utilizarse para automatizar algunas tareas realizadas actualmente por humanos (Lewis et al., 2020). El uso del modelo RAG causa otros problemas que van desde el ruido en los resultados de las recuperaciones o la sobrecarga adicional por el procesamiento de las consultas, pasando por el aumento del tiempo de respuesta y la brecha entre el recuperador y el generador por no tener alineados sus espacios latentes, hasta el incremento de la complejidad del propio sistema al incorporar el proceso de recuperación y del contexto aportado por los documentos. Estos inconvenientes pueden ralentizar el proceso de generación (Zhao et al., 2024).

METODOLOGÍA

De acuerdo con el enfoque seguido en su realización se trata de una investigación cualitativa y cuantitativa. Mientras que, desde el punto de vista de sus objetivos es una investigación exploratoria y descriptiva. Finalmente, a partir de los procedimientos técnicos seguidos es una investigación bibliográfica y documental, con utilización de técnicas de minería de texto. Se realizaron búsquedas sobre RAG en las bases de datos *Web of Science Core Collection*, Scopus e IEEE Xplore, en julio de 2024. La estrategia de búsqueda utilizada se detalla en la Tabla I.

Tabla I. *Estrategia de búsqueda utilizada para la realización del levantamiento bibliográfico*

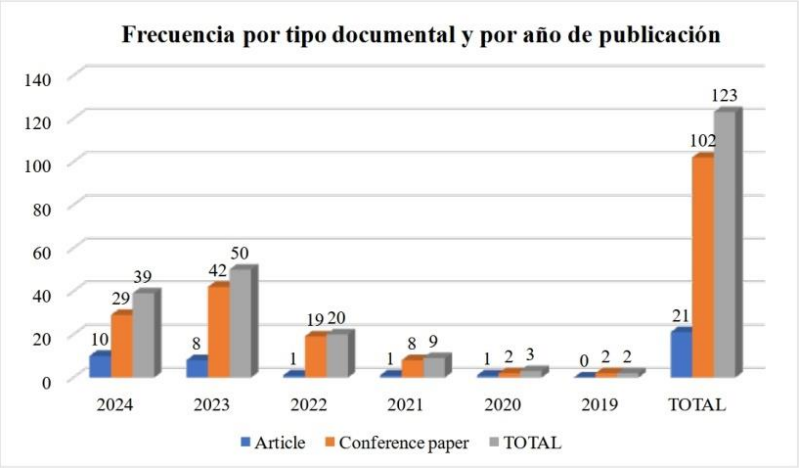
Base de Datos	**Estrategia de búsqueda**		**Nro. Registros**
	Expresión de Busca	**Recursos de Refinamiento Utilizados**	
Scopus	retrieval AND augmented AND generation	Buscar los términos en los campos título, resumen y palabras clave. Años de publicación: 2019 a 2024. Idioma inglés.	364
Web of Science	Retrieval augmented generation (Topic)	Busca por temas: busca los términos de la consulta en el título, resumen, palabras clave del autor y la *keyword plus*. Rango de fecha de publicación: 2019 a 2024. Idioma inglés.	183
IEEE Xplore	("Document Title":retrieval augmented generation) OR ("Abstract": retrieval augmented generation) OR ("Author Keywords": retrieval augmented generation) AND ("Index Terms": retrieval augmented generation)	Buscar los términos en los campos: Título, resumen, palabras clave del autor y palabras claves del índice. Filtro aplicado: años de publicación de 2019 a 2024.	73
Total			620

Los metadatos de los registros recuperados en las bases de datos se exportaron a archivos CSV. Luego se excluyeron los registros duplicados y se procedió con la lectura de los registros para decidir la pertinencia con el objeto y los objetivos seguidos. Los registros se seleccionaron desde estos criterios: las publicaciones debían versar sobre RAG, tenían que proponer o evaluar un modelo y no podían ser revisiones de literatura. Tras consultarse detenidamente se seleccionaron 123 documentos que cumplían esos criterios y se procedió a descargar el texto completo en formato PDF y se extrajo el texto con la herramienta de línea de comandos PDFBox versión 2.0.10. En el preprocesamiento los textos se pasaron a fuentes minúsculas, formato que se mantuvo en la presentación de los resultados y su consiguiente discusión; se eliminaron las palabras vacías y se generaron luego las palabras compuestas bigramas y trigramas. Para efectuar el procesamiento se utilizó la biblioteca de software de aprendizaje de máquina Weka (Bouckaert et al., 2015) combinada con el software VOSviewer (Van Eck y Waltman, 2023) para crear y visualizar mapas a partir de los datos obtenidos en el análisis, con el que se elaboraron figuras, gráficos, grafos de coautoría por autor y por organización, y se hicieron los agrupamientos.

RESULTADOS

En primer lugar, los resultados describen las características generales de la muestra bibliográfica estudiada a través del comportamiento de la tipología de su fuente documental, año de publicación y frecuencia por año. En la figura 2, se comprueba que los autores prefieren publicar en actas de congresos. También que RAG, pese a ser un modelo relativamente nuevo, ha conocido una explosión de interés en la comunidad académica y empresarial, como muestra el crecimiento de las publicaciones sobre el tema que tuvieron un salto superior al 150% del año 2022 al 2023. Ese interés sobre RAG continuaba en alza al cerrarse el levantamiento de datos para esta investigación en julio de 2024, cuando se contabilizaban ya 39 publicaciones en total.

Figura 2. *Frecuencia por tipo documental y por año de publicación*



La distribución de frecuencia de los documentos seleccionados por base de datos es de 105 en Scopus, 12 en *Web of Science* y 6 en *IEEExplorer*. De los cuales el 29,27% son de acceso abierto con diferentes vías de acceso. El 42,03% tienen todos los tipos de vías de acceso abierto, el 23,19% la vía de acceso abierto dorado, el 21,74% la vía de acceso abierto verde, el 7,25% la vía de acceso abierto bronce, el 4,35% la vía de acceso híbrido dorado y el 1,45% la vía de acceso abierto híbrido. Debe de tenerse en cuenta que algunos documentos contaban con más de un tipo de vía de acceso abierto. Se identifican 83 títulos de fuentes de información en las que se difunden los documentos. Un 6,02% de los títulos distribuyen de 5 a 9 artículos, el 14,46% de 2 a 3 publicaciones, mientras que el 79,52% editan sólo un artículo. En la Tabla II se presentan los títulos de las fuentes de información con 6 publicaciones o más. Se destacan por la cantidad de publicaciones en encuentros específicos: *Conference on Empirical Methods in Natural Language Processing, Proceedings (EMNLP)* y *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).*

***Tabla II**. Títulos de las fuentes de información con 6 o más publicaciones*

Título de la Fuente de Información	**Tipo Reg.**	**Evento**	***f***
Conference on Empirical Methods in Natural Language Processing, Proceedings	ISBN	*EMNLP 2023*	8
		EMNLP 2021	3
Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	ISBN	*ECIR 2024*	4
		ECIR 2023	1
		BDA 2023	2
		ADMA 2023	1
		NLPCC 2023	1
Findings of the Association for Computational Linguistics	ISBN	*EMNLP 2023*	6
		EMNLP 2022	3
Proceedings of the AAAI Conference on Artificial Intelligence	ISSN	*AAAI 2024*	5
		AAAI 2022	1
		AAAI 2019	1
Proceedings of the Annual Meeting of the Association for Computational Linguistics	ISBN/ISSN	*ACL 2023*	4
		ACL 2022	1
		ACL 2019	1

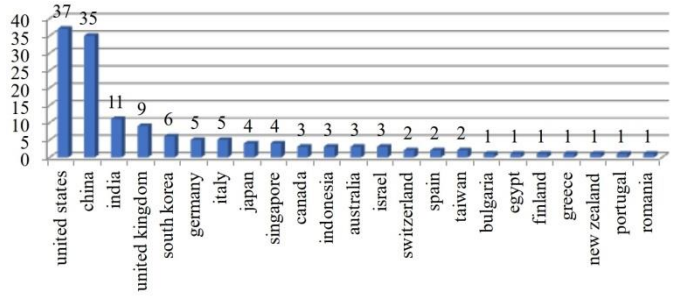
El 42,28% de los artículos declara las organizaciones que financian su investigación relacionada con RAG. La Tabla III lista esas organizaciones con una frecuencia superior a 2. La institución que más investigaciones financia es la *National Natural Science Foundation of China (NSFC)* que dobla a la segunda en el rango la *National Science Foundation (NSF)* de los Estados Unidos. Hay que poner de relieve la cercanía a esta de una empresa tecnológica, la *International Business Machines Corporation (IBM)*. Además, se muestra en el rango que el 50% de las organizaciones financiadoras son organismos del gobierno relacionados con la ciencia y la tecnología, 25% son agencias relacionadas con la defensa y 25% son empresas de tecnología.

***Tabla III.** Organizaciones financiadoras declaradas en los artículos con **f** > 2*

Organización financiadora	***f***
National Natural Science Foundation of China (NSFC), China	18
National Science Foundation (NSF), Estados Unidos	9
International Business Machines Corporation (IBM)	6
Office of Naval Research (ONR), Estados Unidos	4
Google	4
Defense Advanced Research Projects Agency (DARPA), Estados Unidos	3
Institute of Information and Communications Technology, Planning and Evaluation (IITP), Corea del Sur	3
Ministry of Science, ICT and Future Planning (MSIP), Corea del Sur	3

La Figura 3 presenta la frecuencia de las publicaciones del corpus analizado por país. Los países que más destacan son Estados Unidos, seguidos muy de cerca por China. Otros países asiáticos que sobresalen son India y Corea del Sur. Por su parte, y por orden decreciente de publicación, Reino Unido, Alemania e Italia se distinguen entre los países europeos.

Figura 3. *Frecuencia de las publicaciones por país*

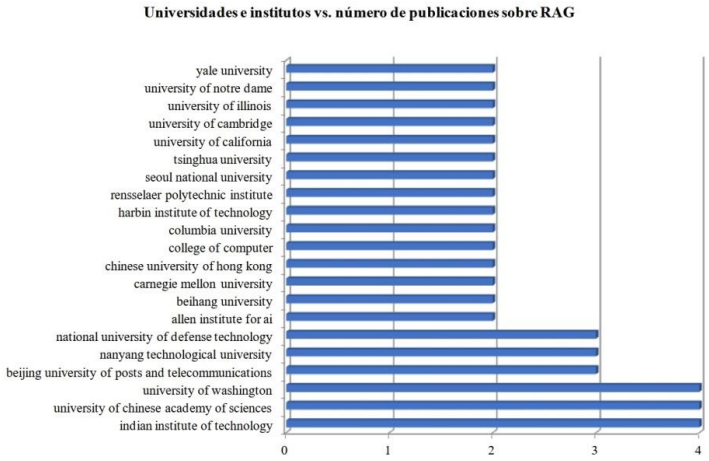


En las publicaciones seleccionadas se identifican 541 autores. De los cuales, un solo autor firma 4 publicaciones; otro autor lo hace en 3 publicaciones; mientras que 40 autores producen 2 publicaciones y 499 autores aparecen sólo una vez. Para el análisis de coautoría se consideran los autores como unidad de análisis, en tanto que el clúster se forma con 3 o más autores, se utiliza como peso para la visualización de los autores el número de documentos publicados. De esta manera se obtienen 84 clústers. En la red de coautoría se distinguen: Zhiliang Tian que realiza 4 publicaciones con 20 autores distintos en los años 2023, 2022 y 2019 y Hannaneh Hajishirzi que publica 3 trabajos con 13 autores en los años 2023 y 2022. De media cada autor publica con 5 coautores, siendo el máximo de 20 coautores para un artículo y desviación estándar de 3,10.

Utilizando el número de citas contabilizados por *Scopus* de los artículos, con 588 citas en dos publicaciones (Lewis, 2020; Min, 2023) aparecen como autores más citados Mike Lewis y Wen-tau Yih. La mayoría de las 585 citas obtenidas se refieren al primer artículo que se publicó sobre RAG en el que ambos eran coautores con Patrick Lewis (2020), como autor principal, y junto a otros 9 investigadores. Les siguen con 78 citas Yu Wu et al. (2019), con 70 citas Jianfeng Gao en dos artículos (Gui et al., 2022; Mao et al. 2021) y con 60 citas Kurt Shuster (Adolphs et al., 2022; Komeili et al., 2022). La media de citas recibidas es de 18, con una desviación estándar de 86,54.

Al analizar las redes de coautoría de las publicaciones sobre RAG, utilizando como unidad de análisis las organizaciones, se identifican 225 organizaciones participantes. De las cuales 130 son universidades o institutos universitarios, 70 empresas, 21 Institutos de I+D+i no universitarios e incluso 2 academias de ciencias, una china y otra búlgara. La figura 4 atiende al número de publicaciones sobre RAG realizadas por universidades e institutos universitarios que hayan producido 2 documentos o más. Entre los que sobresalen *Indian institute of technology*, *University of chinese academy of sciences* y *University of Washington.*

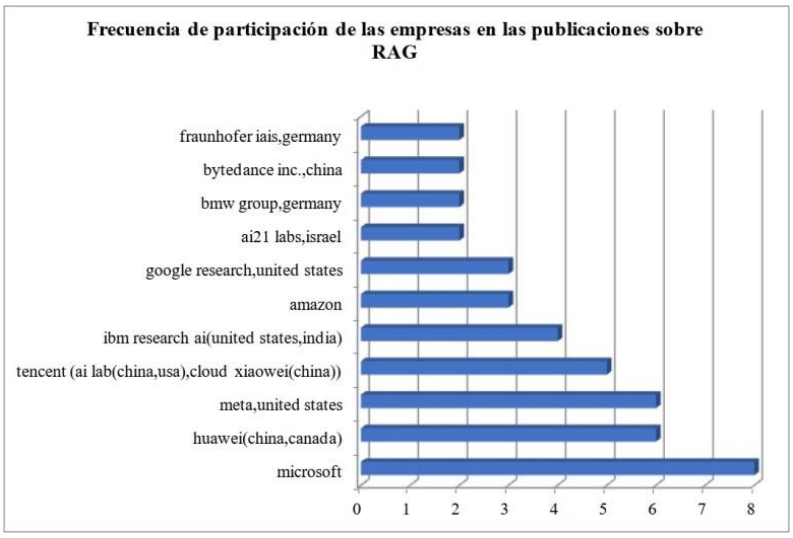
Figura 4. *Universidades e institutos vs. número de publicaciones sobre RAG*



En la Figura 5 se refleja la frecuencia de publicaciones de los autores afiliados a empresas que han producido 2 publicaciones o más sobre RAG. Se destacan por su participación en las publicaciones diversas divisiones

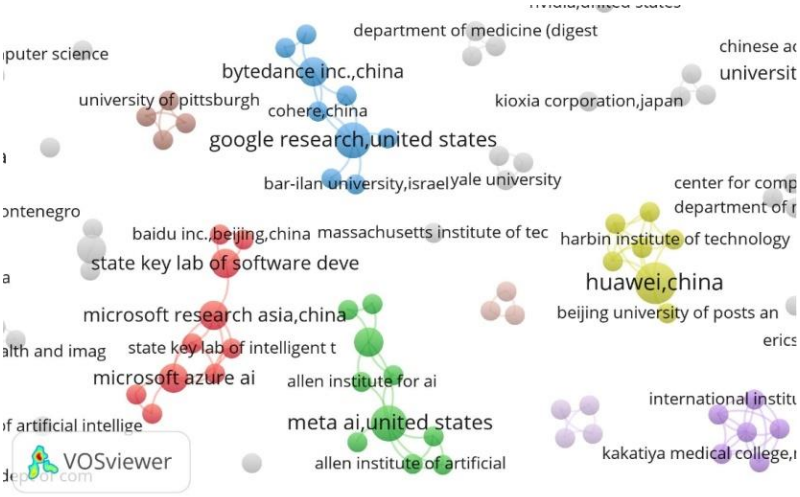
de la empresa tecnológica Microsoft (*semantic machines, azure ai, research, research China, research India*), así como varios departamentos de Meta (*meta, meta ai, facebook ai research*) y de *Huawei (China, Canadá)*. También se constata la presencia en ese rango de otras empresas que actúan en Estados Unidos, China, Alemania, India e Israel.

Figura 5. *Frecuencia de participación de las empresas en las publicaciones sobre RAG*



Al analizar las relaciones de coautoría entre las empresas que han participado en mayor número de publicaciones, se evidencia la conexión existente entre las empresas tecnológicas y las universidades. Así se refleja en la Figura 6 que muestra la red de coautoría entre las diversas organizaciones que más publican sobre RAG. Se identifica la colaboración entre Microsoft *semantic machine* a la hora de participar con las Universidades de Washington y de *Massachusetts Amherst*. *Microsoft*, a través de su división *Research Asia*, colabora con la *Tsinghua University* y la *Beihang University* de China, que a su vez publica con Tencen y Baidu inc. de China. *Microsoft Research* coopera además con la *University of Illinois*, con la *Bharathidasan University* y el *Indian Institute of Science.* La empresa Huawei contribuye con la *Beijing University of Posts and Telecommunications*, la *Chinese University of Hong Kong* y la *Harbin Institute of Technology* de China. Por su parte, Meta AI lo hace con la Universidad de *Washington*, el *Indian Institute of Technology*, la *Carnegie Mellon University* y la *University of Massachusetts Amherst.* Mientras que *Tencent AI lab* colabora con *Hong Kong University of Science and Technology, Minjiang University, Xiamen University, University of Cambridge, The Chinese University of Hong Kong* y la canadiense *University of Alberta*. En tanto que *Google Research* prefiere a la *University of Pennsylvania*, la *Bar-Ilan University* de Israel y la *University of Massachusetts Amherst*. Finalmente, Amazon acompaña a la *University of Cambridge*, la *University of Notre Dame* de los Estados Unidos y el *Indian Institute of Technology.*

Figura 6. *Red de coautoría entre las diversas organizaciones que más publican sobre RAG*



Los bigramas y trigramas de los *tokens* más frecuentes en los documentos de esta investigación se presentan en la tabla IV. En los trigramas se identifican temas como *large language models*, *open web index*, *retrieval augmented generation*, *code comment generation*, *event argument extraction* y *generative artificial intelligence.* Al considerar los bigramas se destacan además asuntos como *question answering, text generation, knowledge graph, dialogue generation* y *search engine.*

Tabla IV. *Bigramas y trigramas de los tokens más frecuentes por documentos*

Token 2-grama	**Doc. *f***	**Token 3-grama**	**Doc. *f***
large language	84	large language models	48
language models	73	open web index	43
natural language	70	retrieval augmented generation	33
question answering	69	code comment generation	26
generative ai	64	source code summarization	19
text generation	62	lay language generation	17
training data	61	human evaluation results	17
external knowledge	58	language processing nlp	16
rag model	57	bart rag dpr	15
knowledge graph	50	artificial intelligence ai	14
augmented generation	48	external knowledge base	13
retrieval augmented	47	rag score test	13
knowledge base	47	event argument extraction	12
dialogue generation	47	natural language inference	11
slot filling	45	pulitzer prize fiction	11
ground truth	42	natural questions triviaqa	11
language modeling	41	code generation summarization	11
models llms	38	generative artificial intelligence	10
search engine	37	computer speech language	10

La palabra clave de *retrieval augmented generation* aparece con otras 52 palabras clave del corpus seleccionado. Sobresalen los términos *semantic search* que consta en 33 documentos, *text generation* en 62 documentos, *search engines* en 37 documentos, *speech processing* que figura en 16 documentos, *semantic embedding* en 14, *social media* en 6, *scientific literature* en 2 y *sentiment analysis* en otros 2. En la Tabla V se presenta el análisis de la aplicación, modelo o técnica y los autores de los artículos del corpus analizado con mayor número de citas.

Tabla V. *Aplicación, modelos o técnicas utilizadas y los autores con mayor número de citas*

Aplicación	**Modelo/Técnicas Utilizadas**	**Autores**
Generación de respuesta en dominio abierto.	Modelo RAG, Modelo seq2seq y índice de vector denso.	(Lewis et al., 2020)
Generación de respuesta en dominio abierto.	Modelo de edición con conocimiento de contexto, *Framework* codificador-decodificador.	(Wu et al., 2019)
Generación de respuesta en dominio abierto.	Generación de texto y Generación de contextos para consultas.	(Mao et al., 2021)
Generación de diálogos	Agente conversacional, Grandes modelos de lenguaje y Generación de consulta basada en contexto.	(Komeili et al., 2022)
Generación de diálogos con conocimiento de estilo.	Recuperación de la información, Generador de respuesta con estilo, Objetivo de aprendizaje con conocimiento de estilo (*Learning objectives*).	(Su et al., 2021)
Generación de respuesta.	Modelo generativo de respuesta aumentado por memoria, *Clustering* y Extracción de características de	(Tian et al., 2020)

	cluster.	
Generación de texto en lenguaje natural a partir de grafos de conocimiento.	Grafos de conocimiento, Procesamiento de lenguaje natural y Generación de datos para texto.	(Agarwal et al., 2021)
Generación de consultas en lenguaje natural.	Modelo generativo *Graph-augmented Sequence to Attention* (*G-S2A*), Arquitectura codificador-decodificador, *Graph Convolutional Network* (*GCN*) en sentencia, *Graph Convolutional Network* en palabras clave, Red neuronal recurrente jerárquica (HRNN) en el codificador y Decodificador *attentional Transformer*.	(Han et al., 2019)
Generación de grafos de conocimiento.	*Zero-shot Slot Filling*, Extensión de un recuperador denso de pasajes y Modelo RAG.	(Glass et al., 2021)
Generación de para-frases. *(Paraphrase Generation)*	*Retrieval Augmented Prompt Tuning* (*RAPT*), *Novelty Conditioned Retrieval Augmented Prompt Tuning* (*NC-RAPT*)	(Chowdhury et al., 2022)
Generación guiada por evidencias.	Generador basado en evidencias, *Multi-task learning framework y Task-agnostic method*.	(Asai et al., 2022)
Transformer aumentado con conocimiento para visión y lenguaje.	*Knowledge Augmented Transformer* (*KAT*) Arquitectura codificador-decodificador con conocimiento explícito e implícito.	(Gui et al., 2022)
Generación de diálogos orientados a tareas basados en diagramas de flujo.	*Deep learning* RAG FLONET	(Raghu et al., 2021)
Open Domain Question Answering (*ODQA*)	*RAG end2end*, Recuperador y Generador para tareas de Respuesta a preguntas (*Question Answering - QA*)	(Siriwardhana et al., 2023)
Detección de prominencia para comprender narrativa de textos largos.	Modelo de lenguaje aumentado por conocimiento y memoria, Detección de prominencia no supervisada, Función cardinal de Barthes, Teoría de la sorpresa, Modelos de lenguaje *transformer*, RAG y Detección de destaques utilizando resumen alineados con el capítulo.	(Wilmot y Keller, 2021)
Generación de texto a partir de fuentes de conocimiento heterogéneos.	Método de generación neural grafo para texto y tabla para texto, Modelo de Recuperador denso DPR, Decodificador T5, *Multi-virtual hops retrieval* (*MVHL*), RNN, Codificador de grafos multirelacional.	(Yu, W., 2022)

DISCUSIÓN

En general los autores prefieren realizar sus publicaciones sobre RAG en actas de congresos, debido a que las reuniones científicas permiten, por una parte, el contacto personal entre los investigadores cuando se exponen las comunicaciones y, por otra, su publicación suele ofrecer las ventajas de la creciente rapidez y alta repercusión cuando cuentan con factor de impacto. Lo que se alinea con lo expuesto por Urbano (2000) respecto a que los investigadores de TIC prefieren publicar en actas de congresos cuando están revisadas. Esta tendencia de los investigadores sobre RAG se desmarca de lo que sucede en Inteligencia artificial, en su conjunto, cuyos autores prefieren publicar en revistas científicas y sólo después en congresos (Maslej et al., 2024). El congreso que publica más sobre RAG es la *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pues alcanza el 19,51% de las publicaciones de nuestro análisis. Se refleja así no sólo el gran atractivo que tiene para los investigadores de PLN, si no que los métodos empleados en RAG cuentan con un fuerte componente empírico en su experimentación, aplicación y evaluación.

La mayor disponibilidad de documentos de acceso abierto en las bases de datos podría explicarse por su rápida divulgación, mayor visibilidad y aumento de las citas. Tal como confirma el trabajo de Rodríguez-Pomeda et al. (2024) en torno a las motivaciones de los académicos españoles para publicar en acceso abierto.

Otro factor puede deberse a las normativas emanadas en las instituciones financiadoras, principalmente del sector público, que obligan a emplear el acceso abierto para comunicar los resultados de las investigaciones. En la distribución de los artículos por fuentes de información se constata el efecto Mateo "al que tiene se le dará aún más, y al que no tiene nada, incluso lo que tenga se le quitará", pues cinco publicaciones concentran 42 artículos, mientras que las 78 publicaciones restantes difunden 81 artículos. Si se considera en cuantos artículos participa un autor no se constata lo que estipula la Ley de Lotka respecto al inverso del cuadrado, ya que se aproxima más al inverso del cubo.

En el análisis de las organizaciones que financian mayor número de investigaciones sobre RAG se comprueba que pertenecen a los países con mayor inversión privada sobre inteligencia artificial. Los datos de inversión privada en inteligencia artificial relativos al año 2023 se recogen en el informe anual de la Universidad de Stanford (Maslej et al., 2024) y muestran que en Estados Unidos hubo una inversión de 67.220.000.000 $; en China de 7.760.000.000 $; mientras que en la Unión Europea y el Reino Unido se invirtieron 11.000.000.000 $, de los cuales 3.780.000.000 $ lo fueron en el Reino Unido y 1.910.000.000 $ en Alemania. Si se considera la inversión de las empresas privadas, en 2023 Microsoft destinó 10.000.000.000 a ChatGPT, a la vez que anuncia su integración al Office 365. Amazon y Google hicieron un acuerdo para dedicar cuatro mil millones y dos mil millones de dólares respectivamente a Anthropic, empresa concurrente de OpenAI. Esas inversiones se reflejan directamente en el predominio de Estados Unidos y de China en las publicaciones sobre RAG, ante el volumen inferior de publicaciones de los países europeos por separado, no así en conjunto que se sitúan en la tercera posición después de China. Se debe tener en cuenta que para elaborar este rango se considera lo declarado en los artículos, pues la financiación de los salarios de los investigadores, que perciben el salario de su organización, y la infraestructura tecnológica utilizada para investigar, no se declara (Sánchez-García-de las Bayonas, 2007).

En la red de coautoría entre las organizaciones se constata que algunas universidades tienen una conexión con empresas que son concurrentes en el mercado. La Universidad de Washington, por ejemplo, publica con Microsoft y con Meta AI; el *Indian Institute of Technology* lo hace con Meta AI y con Amazon y la *Beihang University* de China lo hace con *Microsoft Semantic machine*, Tencen y Baidu de China. Ese tipo de colaboración entre las grandes empresas tecnológicas y las universidades son muy necesarias en las investigaciones sobre RAG, debido al alto coste de la infraestructura tecnológica necesaria para entrenar los grandes modelos de lenguaje. Se estima que el precio del entrenamiento del GPT-4 de la OpenAI se sitúa en 78 millones de dólares (Maslej et al., 2024). La colaboración también es necesaria para hacer viable la puesta en marcha de nuevos prototipos RAG. Este informe muestra que, en el 2023, la colaboración entre las empresas y el mundo académico generó 21 modelos de aprendizaje de máquina; por su parte y en solitario, la industria produjo 51 modelos y el sector académico 15 modelos.

Al comparar los términos identificados en los bigramas y trigramas de la tabla V con las aplicaciones RAG de los autores más citados, se aprecia la progresión de la generación de respuestas en campos de conocimiento abiertos, así como de diálogos con o sin reconocimiento de estilos. Otras aplicaciones destacables de RAG son la generación de grafos de conocimiento y la generación de texto en lenguaje natural mediante la utilización de esos grafos. Asimismo, la generación de consultas en lenguaje natural. Además, en el conjunto de artículos analizados se observa un creciente interés en la realización de generación guiada por evidencias y la generación de texto a partir de fuentes de conocimiento heterogéneas. Sin olvidarnos del desarrollo de métricas de evaluación para la RAG.

CONCLUSIONES

RAG es un modelo de sistema de inteligencia artificial generativa que recupera datos desde bases de conocimientos externas de las que toma información muy precisa y actualizada para fundamentar grandes modelos lingüísticos. En el campo científico de la información se relaciona con la recuperación y con el procesamiento del lenguaje natural. Esta técnica gana adeptos rápidamente, debido, por una parte, a que interviene a la hora de recuperar, completar, actualizar e interpretar información relevante. Además, porque ocasiona respuestas a preguntas complejas. Su intensa actividad afecta tanto a la comunidad académica como

a la empresarial. Su notorio ascenso se deriva en gran parte de una experiencia eficiente que consigue superar las limitaciones de los modelos anteriores de grandes modelos de lenguaje, al agregarles información específica y actualizada proveniente de fuentes externas. Así se infiere del incremento observado en la disponibilidad de publicaciones de acceso abierto sobre RAG en las bases de datos que facilitan el acceso a los textos completos sin necesidad de acudir a otros repositorios. Lo que también da indicios de una mayor concienciación de la comunidad académica para publicar en acceso abierto por cuanto sirve para llegar a un público mayor de forma más rápida, a la vez que aumenta el impacto de sus trabajos.

Hay que tener en cuenta que se trataba de ofrecer un primer panorama sobre RAG. Lo que se ha conseguido a partir de la identificación de sus principales investigadores, las vías de publicación utilizadas y las organizaciones que financian sus proyectos. En las aplicaciones RAG se comprueba el predominio de la generación de contenido que, en el caso concreto de este análisis, se centra en las aplicaciones relacionadas con la generación textual. Ya sea de respuestas en campos de conocimiento abiertos, de diálogos con y sin reconocimiento de estilos, de etiquetas y resúmenes documentales, de códigos de programación, de respuestas en lenguaje natural a partir de grafos de conocimientos, de consultas en lenguaje natural, de frases y paráfrasis para explicar o traducir textos, de textos desde la evidencia, de diálogos orientados a tareas sobre todo con *chatBots* e incluso de texto a partir de fuentes de conocimiento heterogéneas. Aun cuando se ha podido comprobar que la utilización del modelo RAG se amplía con prisa hacia objetos no textuales como los audios, las imágenes o los videos, por no hablar de estructuras de proteínas o de otros que se integran en campos sin fin, por más que estén fuera de la esfera de esta investigación.

El panorama de la investigación sobre RAG se delinea desde la presentación de las publicaciones de manera preferente en congresos, los autores que más publicaron lo hicieron con muchos coautores. Al mismo tiempo, es determinante la estrecha colaboración entre las grandes empresas tecnológicas y las universidades o institutos en los que se integran los investigadores. Las instituciones que financian mayor número de investigaciones sobre RAG se corresponden con fundaciones gubernamentales de ciencias naturales, organizaciones de investigación de defensa y grandes empresas tecnológicas. Los países con mayor inversión privada en inteligencia artificial son los que más publicaron. Finalmente, el modelo RAG tiene grandes potencialidades de aplicación en las investigaciones de las humanidades digitales, por permitir que se aproveche el conocimiento aprendido en los parámetros de los grandes modelos de lenguaje preentrenados con informaciones de repositorios específicos que podrían ser bases de datos de documentos históricos, o de bibliotecas que pueden ser utilizados para crear nuevos servicios para los usuarios aprovechando la inteligencia artificial generativa.

REFERENCIAS BIBLIOGRÁFICAS

Agarwal, O., Ge, H., Shakeri, S., y Al-Rfou, R. (2021). Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. En Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T. y Zhou, Y. (Eds.), Proceedings *of the* 2021 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3554-3565). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.278

Asai, A., Gardner ,M., y Hajishirzi, H. (2022). Evidentiality-guided generation for knowledge-intensive NLP tasks. En Carpuat, M., de Marneffe, M.-C. y Meza Ruiz, I. V. (Eds.), *Proceedings of the* 2022 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2226-2243). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.162

Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., y Scuse, D. (2015). *WEKA Manual for Version 3-7-13*. The University of Waikato. https://master.dl.sourceforge.net/project/weka/documentation/3.7.x/WekaManual-3-7-13.pdf

Cai. D., Wang, Y., Liu, L., y Shi, S. (2022). Recent advances in retrieval-augmented text generation. En Amigo, E., Castells, P., Gonzalo, J., Carteree, B., Culpepper, J.S. y Kazai, G. (Eds.), *SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3417-3419). Association for Computational Linguistics. https://doi.org/10.1145/3477495.3532682

Chowdhury, J.R., Zhuang, Y., y Wang, S., (2022). Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning. En Sycara, K., Honavar, V. y Spaan, M. (Eds.), *Proceedings of the 36th AAAI Conference on Artificial Intelligence* (pp. 10535-10544). Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v36i10.21297

Glass, M., Rossiello, G., Chowdhury, M.F.M., y Gliozzo A. (2021). Robust retrieval augmented generation for zero-shot slot filling. En Moens, M.-F., Huang, X., Specia, L., y Yih, S.W. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1939-1949). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.148

Gui, L., Wang, B., Huang, Q., Hauptmann, A., Bisk, Y., y Gao, J. (2022). KAT: a knowledge augmented transformer for vision-and-language. En Carpuat,M., de Marneffe, M.-C. y Meza Ruiz, I.V. (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 956–968). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.70

Han, F.X., Guo, W., Niu, D., He, Y., Lai. K., y Xu, Y. (2019). Inferring search queries from web documents via a graph-augmented sequence to attention network. En Liu, L. y White, R. (Eds.), *Proceedings of the WWW '19: The World Wide Web Conference* (pp. 2792-2798). International World Wide Web Conference Committee. https://doi.org/10.1145/3308558.3313746

Komeili, M., Shuster, K., y Weston, J. (2022). Internet-augmented dialogue generation. En Muresan, S., Nakov, P., y Villavicencio, A. (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1,* (pp. 8460-8478*)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.579

Lewis, P., Perez, E., Piktus, A., Petroni ,F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T. y otros. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. En Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 9459–9474). Curran Associates Inc. https://dl.acm.org/doi/pdf/10.5555/3495724.3496517

Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., y Chen, W. (2021). Generation-augmented retrieval for open-domain question answering. En Zong, C., Xia, F., Li, W. y Navigli, R. (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 4089-4100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.316

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J. y otros. (2024). *Artificial Intelligence Index Report 2024*. Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf

Raghu, D., Agarwal, S., Joshi, S., y Mausam. 2021. End-to-End learning of flowchart grounded task-oriented dialogs. En Moens, M.-F., Huang, X., Specia, L., y Yih, S.W. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 4348-4366). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.357.pdf

Rodríguez-Pomeda, J., Perez-Encinas, A., De La Torre, E. M. (2024). Motivaciones de los académicos españoles para publicar en revistas de acceso abierto: un análisis sociodemográfico. *Revista Española de Documentación Científica*, 47(3), e393. https://doi.org/ 10.3989/redc.2024.3.1555

Sánchez-García-de las Bayonas, S. (2007). Repercusión de la publicación científica electrónica de acceso abierto en los presupuestos y en el acceso a la información científica en las bibliotecas universitarias

españolas. *Revista Española De Documentación Científica*, 30(3), 323–342. https://doi.org/10.3989/redc.2007.v30.i3.388

Shuster, K., Poff, S., Chen, M., Kiela, D., y Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. En Moens, M.-F., Huang, X., Specia, L., y Yih, S. W.-T. (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784-3803). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., y Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11, 1-17. https://doi.org/10.1162/tacl_a_00530

Su, Y., Wang, Y., Cai, D., Baker, S., Korhonen, A., y Collier, N. (2021). PROTOTYPE-TO-STYLE: Dialogue generation with style-aware editing on retrieval memory. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 29, 2152-2161. https://doi.org/10.1109/TASLP.2021.3087948

Tian, Z., Bi, W., Li, X., y Zhang, N.L. (2020). Learning to abstract for memory-augmented conversational response generation. En Korhonen, A., Traum, D. y Màrquez, L. (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3816-3825). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1371

Urbano Salido, C. (2000). Tipología documental citada en tesis doctorales de informática: bases empíricas para la gestión equilibrada de colecciones. *Textos universitaris de biblioteconomia i documentació*, (5). https://bid.ub.edu/05urban2.htm

Van Eck, N.J. y Waltman, L. (2023). *VOSviewer Manual version 1.6.20*. Universiteit Leiden. http://www.vosviewer.com/

Wilmot, D., y Keller, F. (2021). Memory and knowledge augmented language models for inferring salience in long-form stories. En Moens, M.-F., Huang, X., Specia, L. y Yih, S. W. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 851-865).* Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.65

Wu, Y., Wei, F., Huang, S., Wang, Y., Li, Z., y Zhou, M. (2019). Response generation by context-aware prototype editing. In Van Hentenryck, P. y Zhou, Z.-H (Eds.), *Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 7281-7288. AAAI Press. https://doi.org/10.1609/aaai.v33i01.33017281

Yu, W. (2022). Retrieval-augmented Generation across Heterogeneous Knowledge. In Ippolito, D., Li, L.H., Pacheco, M.L., Chen, D. y Xue, N. (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop* (pp. 52–58). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2022.naacl-srw.7

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., y Cui, B. (2024). Retrieval-Augmented Generation for AI-Generated Content: A Survey. ArXiv, abs/2402.19473. https://doi.org/10.48550/arXiv.2402.19473