https://doi.org/10.69639/arandu.v12i3.1543

Modelos de inteligencia artificial generativa y grandes modelos de lenguaje(LLM): Análisis sistemático basado en metodología PRISMA de evolución, rendimiento y aplicaciones (2020-2025)

Generative artificial intelligence models and large language models(LLM): Systematic analysis based on PRISMA methodology of evolution, performance and applications (2020-2025)

Rodrigo Aryan Hernández García
rodrhernandez@uv.mx
https://orcid.org/0000-0002-2299-5366
Universidad Veracruzana
México

Cesar Augusto Mejia Gracia
cemejia@uv.mx
https://orcid.org/0000-0001-8874-0473
Universidad Veracruzana
México

Vicente Josue Aguilera Rueda
vaguilera@uv.mx
https://orcid.org/0000-0002-1952-7860
Universidad Veracruzana
México

Carlos Francisco Domínguez Domínguez
cardominguez@uv.mx
https://orcid.org/0000-0001-7113-4063
Universidad Veracruzana
México

Luis Enrique Corona Morales
lucorona@uv.mx
https://orcid.org/0009-0003-7643-1739
Universidad Veracruzana
México

Artículo recibido: 18 agosto 2025 - Aceptado para publicación: 28 septiembre 2025
Conflictos de intereses: Ninguno que declarar.

RESUMEN

El presente trabajo presenta un análisis sistemático basado en metodología PRISMA que examina 89 estudios primarios y 20 fuentes de alta calidad sobre Inteligencia Artificial Generativa (IAGen) y los grandes modelos del lenguaje (LLM, Large Language Models), publicados entre 2020-2025. La revisión identifica más de 50 LLMs existentes y documenta aplicaciones en 10 dominios críticos, estableciendo un marco de evaluación destinado a orientar

futuras investigaciones. Los hallazgos revelan que GPT-4 (88.7% MMLU), Claude 3.5 Sonnet, y los modelos emergentes de razonamiento como DeepSeek-R1 lideran el rendimiento actual, siendo aplicados principalmente en atención sanitaria (94.4% de estudios), educación, desarrollo de software, e investigación científica. Los resultados demuestran mejoras significativas de eficiencia: 40% reducción en tiempo de documentación clínica, 35% incremento en productividad de desarrollo de software, y 58% reducción en tiempo de revisiones sistemáticas. El sector evoluciona aceleradamente hacia capacidades de razonamiento especializado, con modelos como DeepSeek-R1 alcanzando 97.3% en MATH-500, multimodalidad nativa que permite procesamiento integrado de texto, imagen, audio y video, y democratización mediante modelos de código abierto competitivos como LLaMA-4. Los desafíos persistentes incluyen alucinaciones (15-25% de resultados), sesgo cultural documentado, y necesidad urgente de marcos éticos estandarizados para aplicaciones críticas que requieren garantías de seguridad.

Palabras clave: grandes modelos de lenguaje, inteligencia artificial generativa, transformers, evaluación de modelos, análisis sistemático

ABSTRACT

This study presents a systematic analysis based on PRISMA methodology examining 89 primary studies and 20 high-quality sources on Generative Artificial Intelligence (GAI) and Large Language Models (LLMs) published between 2020-2025. The review identifies over 50 existing LLMs and documents applications across 10 critical domains, establishing an evaluation framework intended to guide future research. Findings reveal that GPT-4 (88.7% MMLU), Claude 3.5 Sonnet, and emerging reasoning models such as DeepSeek-R1 lead current performance, being primarily applied in healthcare (94.4% of studies), education, software development, and scientific research. Results demonstrate significant efficiency improvements: 40% reduction in clinical documentation time, 35% increase in software development productivity, and 58% reduction in systematic review time. The sector is rapidly evolving toward specialized reasoning capabilities, with models like DeepSeek-R1 achieving 97.3% on MATH-500, native multimodality enabling integrated processing of text, image, audio, and video, and democratization through competitive open-source models like LLaMA-4. Persistent challenges include hallucinations (15-25% of outputs), documented cultural bias, and urgent need for standardized ethical frameworks for critical applications requiring safety guarantees.

Keywords: large language models, generative artificial intelligence, transformers, model evaluation, systematic analysis

INTRODUCCIÓN

La aparición de ChatGPT en el mes de noviembre del 2022 marcó un punto de inflexión tecnológico cuya magnitud trasciende a las revoluciones digitales previas, alcanzando 100 millones de usuarios en apenas 60 días y catalizando una transformación acelerada en múltiples sectores económicos y académicos a escala global. Mientras que Internet (7 años) y los smartphones (5 años) requirieron períodos significativos para lograr adopción masiva, los modelos de IA generativas han demostrado velocidades de penetración sin precedentes, con un crecimiento del mercado global proyectado de entre $2.6 billones y $4.4 billones USD anuales a la economía global, comparado con el PIB total del Reino Unido de $3.1 billones en 2021 (McKinsey Global Institute, 2023).

Este fenómeno representa más que una innovación tecnológica, está redefiniendo las fronteras entre capacidades humanas y automatización inteligente, exigiendo marcos de análisis rigurosos que permitan comprender las implicaciones científicas, económicas y sociales. La progresión técnica desde el modelo Transformer original (Vaswani et al., 2017) hasta los contemporáneos revela una secuencia de avances cualitativos que desafían proyecciones lineales de mejora tecnológica, y apuntan a lo exponencial.

A pesar del impacto y la documentación al respecto de la IAGen, la literatura científica carece de síntesis comprehensivas que integren desarrollos técnicos, métricas de rendimiento estandarizadas, y análisis de impacto sectorial mediante metodologías de revisión sistemática rigurosas. Esta fragmentación del conocimiento limita la capacidad de investigadores, responsables de políticas, y profesionales para tomar decisiones fundamentadas empíricamente sobre adopción, regulación, e inversión en IA generativa.

Este análisis sistemático basado en metodología PRISMA tiene como objetivo principal mapear exhaustivamente el ecosistema actual de los distintos modelos de IA generativa, establecer benchmarks de rendimiento comparativo fundamentados en evidencia empírica, e identificar tendencias críticas destinadas a orientar investigación futura y políticas de implementación responsable.

Fundamentos

Modelo Transformer

El modelo Transformer, establecido por Vaswani et al. (2017), constituye el fundamento teórico que sustenta prácticamente a todas las Inteligencias Artificiales Generativas contemporáneas. La innovación central reside en el mecanismo de auto-atención que permite análisis simultáneo de secuencias completas, superando limitaciones fundamentales de modelos recurrentes previos.

El mecanismo de atención escalado (Scaled Dot-Product Attention) se define matemáticamente como:

Attention(Q, K, V) = softmax(QK^T / √d_k)V

- Q es el vector de consulta
- K es el vector clave
- V es el vector de valores
- T es la operación de transposición
- Softmax es la función softmax que se utiliza en redes neuronales.

Transformer posibilita la captura simultánea de relaciones tanto locales como globales dentro de las secuencias, facilitando así la comprensión de dependencias a largo alcance que anteriormente resultaban inmanejables en las primeras redes neuronales.

Evolución hacia multimodalidad nativa

Las primeras aplicaciones de IAGen solo permitían trabajar datos en forma de texto, y generaba salidas en el mismo, sin embargo, los LLMS que son multimodales nativos como GPT-4V y Gemini procesan texto, imágenes, audio, y video a través de un marco de trabajo unificado desde "las capas de entrada"(inputs layer). Esta integración posibilita capacidades emergentes como razonamiento visual y comprensión temporal de video, mostrando 34% mayor precisión en tareas de razonamiento transmodal comparado con aproximaciones basadas en solo texto.

MATERIALES Y MÉTODOS

La búsqueda sistemática se ejecutó en seis bases de datos prioritarias: Scopus, Web of Science/Clarivate, Springer, IEEE Xplore, ACM Digital Library, y Nature Portfolio, complementadas por bases de datos regionales SciELO, Redalyc, y Latindex destinadas a capturar perspectivas en español y contextos latinoamericanos.

Términos de búsqueda: Se utilizaron combinaciones booleanas de ("large language models" OR "LLM" OR "generative artificial intelligence" OR "ChatGPT" OR "GPT-4" OR "BERT") AND ("evaluation" OR "performance" OR "implementation" OR "application") con filtros temporales 2020-2025 y restricciones de calidad académica.

Criterios de inclusión

Artículos de investigación primaria publicados en revistas indexadas Q1 o Q2, con DOI válido, metodología clara, y enfoque directo en modelos LLM o IA generativa. Se incluyeron estudios empíricos, revisiones sistemáticas, meta-análisis, y reportes técnicos de instituciones reconocidas, se consideró únicamente literatura en inglés y español.

Criterios de exclusión

Opiniones editoriales sin datos originales, publicaciones en revistas predatorias, estudios sin información metodológica suficiente, duplicados, y artículos no relacionados directamente con LLM o IA generativa, estudios en un idioma distinto al inglés y al español.

Proceso de selección

El proceso resultó en la identificación inicial de 3,921 registros únicos tras eliminación de duplicados. Después del cribado (screening) por título y resumen, se evaluaron 309 artículos de texto completo, resultando 109 estudios incluidos en la síntesis cualitativa y 89 estudios en la síntesis cuantitativa.

Limitaciones

Primero la rápida evolución del campo puede haber resultado en sesgo de publicación hacia modelos más recientes. Segundo, la heterogeneidad metodológica entre estudios limitó las posibilidades de metaanálisis cuantitativo.

RESULTADOS Y DISCUSIÓN

A continuación, se presentan los resultados de la investigación, donde se pueden apreciar la cantidad de artículos por LLM, así como sus campos de aplicación, posteriormente se comenta sobre cada tecnología en lo particular.

Tabla 1

Distribución de estudios por tecnología LLM identificada

Tecnología	Estudios	Porcentaje	Detalles específicos
OpenAI GPT series	28	31.5%	GPT-4/GPT-4o: 15 estudios (aplicaciones médicas, educación, desarrollo software) GPT-3.5/ChatGPT: 13 estudios (principalmente educación y servicios cliente)
Anthropic Claude series	18	20.2%	Claude 3.5 Sonnet: 11 estudios (análisis legal, investigación científica) Claude Opus: 7 estudios (escritura académica, análisis complejo)
Google/DeepMind	16	18.0%	Gemini Pro/Ultra: 9 estudios (multimodalidad, análisis de datos) PaLM/Med-PaLM: 7 estudios (aplicaciones médicas especializadas)
Meta LLaMA series	15	16.9%	LLaMA-2/3: 12 estudios (código abierto, implementaciones locales) Code Llama: 3 estudios (desarrollo software especializado)
Modelos especializados	12	13.5%	DeepSeek series: 5 estudios (eficiencia computacional) Mistral AI: 4 estudios (modelos europeos) Otros modelos de código abierto: 3 estudios

Nota: Diseño propio a partir de la recopilación de datos conforme a la metodología.

OpenAI GPT

GPT-4.1 (enero 2025) representa el estado del arte actual en modelos comerciales, estableciendo nuevos estándares con 1M tokens de contexto expandible y 54.6% precisión en SWE-bench Verified destinado a resolución de problemas reales de software. GPT-4.1 incorpora mejoras fundamentales en razonamiento paso-a-paso y tasas reducidas de alucinación mediante técnicas de verificación internas.

GPT-4o (omni) introduce multimodalidad nativa con procesamiento integrado de texto, imagen, audio, y video desde capas base. Omni demuestra 88.7% en MMLU, mantiene paridad con GPT-4 en benchmarks textuales incorporando capacidades transmodales sin degradación de rendimiento.

Anthropic Claude

Claude 3.7 Sonnet establece nuevo paradigma con capacidades híbridas de razonamiento instantáneo y extendido, permitiendo conmutación dinámica entre modos según complejidad de tarea. Logra 70.3% en SWE-bench mediante metodología de razonamiento estructurado que supera aproximaciones de generación directa.

Claude 3.5 Sonnet mantiene equilibrio óptimo entre capacidad y eficiencia, estableciéndose como referencia destinada a aplicaciones comerciales que requieren confiabilidad consistente, exhibe tasas de alucinación reducidas (12% menor que competidores) y adherencia superior a instrucciones complejas.

Google Gemini

Gemini 2.5 Pro lidera clasificaciones actuales como primer "thinking model" comercial con “cadena de pensamiento”(“Chain-of-Thought “) integrada nativamente. Ocupa posición #1 en LMArena leaderboard, logrando 63.8% en SWE-bench y soporte nativo destinado a 100+ idiomas con rendimiento casi-nativo en 40+ idiomas prioritarios.

Meta LLaMA

LLaMA 4 (2025) marca avance decisivo en modelos de código abierto con capacidades multimodales nativas que compiten directamente con GPT-4o. Disponible en configuraciones 8B, 70B, y 405B parámetros, ofrece escalabilidad desde “dispositivos de borde” (“Edge Devices”) hasta centros de datos empresariales.

LLaMA 3.3 70B demuestra que modelos relativamente compactos pueden lograr rendimiento competitivo mediante optimizaciones de entrenamiento avanzadas y curación de datos superior.

DeepSeek series: eficiencia revolucionaria

DeepSeek-R1 (enero 2025) representa avance decisivo en eficiencia de entrenamiento con 671B parámetros totales (37B activos durante inferencia) logrando 97.3% en MATH-500 con costos de entrenamiento de $5.6M. R1 permite escalabilidad masiva manteniendo costos operacionales razonables.

Análisis comparativo de desempeño

Tabla 2

Comparación de Rendimiento en los Benchmarks Principales

Modelo	MMLU (%)	SWE-bench Verified (%)	MATH-500 (%)	HumanEval (%)	MedQA (%)	Capacidades Especiales
OpenAI o3	90.2	-	-	-	-	Razonamiento avanzado
GPT-4.1	88.7	54.6	52.9	85.0	-	Contexto 1M tokens
GPT-4	86.4	-	52.9	67.0	-	Multimodal integrado
Gemini 2.5 Pro	85.4	63.8	-	99.0	-	Thinking model nativo
Claude 3.7 Sonnet	84.2	70.3	71.1	92.0	-	Razonamiento híbrido
Claude 3.5 Sonnet	-	-	71.1	-	-	Alucinación reducida
DeepSeek-R1	-	-	97.3	-	-	Eficiencia extrema
LLaMA-3.3 70B	-	31.2	-	67.8	-	Código abierto
Med-PaLM 2	-	-	-	-	87.0	Especialización médica
GPT-3	43.9	-	14.5	29.0	-	Referencia histórica

Nota: Diseño propio a partir de la recopilación de datos conforme a la metodología.

MMLU (Massive Multitask Language Understanding) constituye la métrica fundamental orientada hacia evaluación de conocimiento académico, abarcando 57 materias desde matemáticas elementales hasta derecho profesional. Los líderes actuales incluyen OpenAI o3 (90.2%), Gemini 2.5 Pro (85.4%), Claude 3.7 Sonnet (84.2%), junto con GPT-4.1 (88.7%), representando convergencia en competencias de conocimiento factual entre modelos comerciales.

SWE-bench Verified establece estándar dorado orientado hacia competencias de ingeniería de software mediante resolución de problemas reales de GitHub. Claude 3.7 Sonnet domina con 70.3%, GPT-4.1 alcanza 54.6%, mientras que los de código abierto como LLaMA-3.3 logran 31.2%.

Métrica MATH evalúa competencias matemáticas desde nivel preparatoria hasta licenciatura. DeepSeek-R1 establece récord con 97.3% de precisión, superando significativamente a GPT-4 (52.9%) junto con Claude 3.5 Sonnet (71.1%).

Mapeo de aplicaciones actuales

Tabla 4

Síntesis de aplicaciones por sector

Sector	Estudios	Porcentaje	Tecnologías predominantes	Aplicaciones principales	Resultados promedio
Salud	26	29.2%	Med-PaLM 2, GPT-4, Claude 3.5	Diagnóstico asistido, documentación clínica, educación médica	83-87% precisión en benchmarks médicos
Educación	23	25.8%	GPT-4, ChatGPT, Claude	Tutoría personalizada, generación de contenido, evaluación automática	25-40% mejora en engagement estudiantil
Desarrollo de Software	21	23.6%	GitHub Copilot, Code Llama, GPT-4	Generación de código, debugging, documentación	35% incremento en productividad
Investigación Científica	12	13.5%	GPT-4, Claude Opus, modelos especializados	Revisiones sistemáticas, análisis de datos, hipótesis	58% reducción en tiempo de revisión
Otros sectores	7	7.9%	Diversos	Finanzas, legal, marketing, manufactura	Variable según sector

Nota: Diseño propio a partir de la recopilación de datos conforme a la metodología.

Atención médica

La atención sanitaria representa 30% de aplicaciones documentadas en literatura académica, acompañada de 94.4% de estudios enfocados en chatbots médicos orientados hacia educación del paciente junto con soporte de decisiones clínicas. Med-PaLM 2 establece precisión estado-del-arte en métricas médicas (MedQA 87.0%, MedMCQA 72.3%, PubMedQA 79.0%).

Métricas de éxito incluyen: 40% de reducción en tiempo de documentación clínica, 83% de precisión en extracción de información médica, junto con expansión a través de 29 especialidades médicas. Implementaciones notables incluyen modelos de soporte diagnóstico de Mayo Clinic junto con chatbots NHS orientados hacia tamizaje preliminar.

Investigación científica

Revisiones sistemáticas automatizadas muestran resultados transformadores: 58% reducción en carga laboral de tamizaje manual, 68% disminución en errores de extracción de información, junto con compresión temporal desde 67.3 semanas promedio hasta 2 semanas orientado hacia revisiones comprehensivas.

BrainGPT representa avance en predicción científica, superando especialistas humanos en predicción de resultados experimentales de neurociencia con 78% precisión versus 65% línea de referencia humana.

Desarrollo de software: transformación industrial

GitHub Copilot reporta 35% incremento en velocidad de programación entre usuarios. Aplicaciones comprehensivas incluyen generación de código desde lenguaje natural (soportando 200+ lenguajes de programación), refactorización automatizada (45% ahorro de tiempo reportado), junto con detección y corrección de errores (73% precisión identificando errores lógicos).

Métricas de adopción empresarial: 78% de desarrolladores encuestados emplean asistentes de codificación IA regularmente, acompañadas de ganancias de productividad oscilando 25-40% a través de diferentes tareas de programación.

Educación personalizada junto con democratización

Modelos de tutoría adaptativa demuestran 25-40% mejora en resultados de aprendizaje mediante algoritmos de personalización que ajustan dificultad junto con ritmo de contenido a necesidades estudiantiles individuales. Integración GPT-4 de Duolingo muestra 34% incremento en compromiso junto con 28% mejores porcentajes de retención comparado con métodos de instrucción tradicionales.

Acceso educativo global: soporte destinado a 100+ idiomas con rendimiento casi-nativo en 40+ idiomas prioritarios facilita educación en regiones desatendidas.

Tendencias identificadas

Convergencia hacia modelos de razonamiento avanzado

El año 2025 marca transición fundamental desde mejoras fundamentadas en escalamiento hacia competencias de razonamiento especializadas. OpenAI o-series, DeepSeek-R1, junto con Gemini 2.5 Pro representan nueva generación que emplea escalamiento de cómputo en tiempo de prueba orientado hacia "pensar" antes de responder, logrando mejoras considerables en matemáticas (97.3% MATH-500) junto con tareas de razonamiento científico.

Democratización mediante código abierto competitivo

Estrechamiento de brecha de desempeño entre modelos comerciales junto con código abierto demuestra democratización de competencias como se pudo notar en la investigación, DeepSeek-R1, QwQ-32B, junto con LLaMA-4 logran paridad competitiva en múltiples métricas.

Revolución de gastos de capacitación / adiestramiento ejemplificado por DeepSeek-V3 logrando desempeño competitivo con $5.6M versus estimados $50M+ orientado hacia competidores comerciales sugiere disrupción potencial de ventajas de recursos mantenidas por corporaciones principales.

Brechas de investigación identificadas

Evaluación de impacto longitudinal permanece insuficiente acompañada de estudios limitados examinando efectos del despliegue a largo plazo de LLM en educación, atención sanitaria, junto con productividad laboral, recordando que estas tecnologías se presentan en el año 2022, por lo cual aún es difícil dar una trazabilidad a largo plazo.

Marcos éticos estandarizados requieren desarrollo urgente, particularmente orientados hacia aplicaciones críticas que requieren garantías de seguridad.

Representación cultural junto con lingüística muestra sesgo geográfico significativo: perspectivas occidentales dominan información de adiestramiento junto con métricas de evaluación, mientras que 38.5% de estudios originan desde instituciones estadounidenses.

CONCLUSIONES

Este análisis sistemático fundamentado en metodología PRISMA muestra un ecosistema LLM en rápida maduración caracterizado por convergencia en competencias fundamentales entre modelos líderes comerciales, democratización mediante alternativas de código abierto competitivas, junto con transición paradigmática hacia competencias de razonamiento especializadas.

Los sectores donde las aplicaciones han alcanzado mayor madurez —atención sanitaria, educación e investigación científica— evidencian un valor transformador mensurable acompañado de mejoras de eficiencia del 40-68% junto con compresión temporal significativa en procesos complejos tradicionalmente intensivos en tiempo humano.

La multimodalidad nativa habilita una nueva clase de aplicaciones integradas que trascienden limitaciones de modelos tradicionales.

Desafíos persistentes incluyen alucinaciones (afectando 15-25% de resultados), sesgo cultural junto con demográfico documentado a través de múltiples estudios, además de problemas de reproducibilidad (38.2% de estudios reportan inconsistencias de salida). Marcos regulatorios junto con de seguridad retrasan competencias tecnológicas, requiriendo desarrollo urgente orientado hacia aplicaciones críticas.

Trabajos futuros deben priorizar el Desarrollo de métricas dinámicas resistentes a contaminación de información; (2) Marcos éticos estandarizados orientados hacia aplicaciones críticas; (3) Estudios de impacto longitudinal examinando efectos de despliegue del mundo real; (4) Democratización de competencias avanzadas mediante mejoras continuas de eficiencia; junto con (5) Diversidad cultural junto con lingüística en información de adiestramiento.

El sector evoluciona hacia IA práctica, implementable, junto con globalmente accesible que balancea avance de competencias con responsabilidad, eficiencia con desempeño, e innovación con consideraciones éticas. Esta revisión sistemática proporciona fundamentos orientados hacia toma de decisiones fundamentada en evidencia en prioridades de investigación,

desarrollo de políticas, junto con estrategias de despliegue responsable orientadas hacia la próxima década.

REFERENCIAS

Boiko, D. A., et al. (2023). "Autonomous chemical research with large language models." Nature, 619, 423-428. DOI: https://doi.org/10.1038/s41586-023-06792-0 .

Bommasani, R., et al. (2022). "Holistic Evaluation of Language Models (HELM)." Transactions on Machine Learning Research. DOI: https://doi.org/10.48550/arXiv.2211.09110 .

Brown, T., et al. (2020). "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems, 33, 1877-1901. DOI: https://doi.org/10.48550/arXiv.2005.14165 .

Chang, Y., et al. (2024). "A Survey on Evaluation of Large Language Models." ACM Transactions on Intelligent Systems and Technology, 15(3). DOI: https://doi.org/10.1145/3641289

García-Peñalvo, F.J. (2024). "La nueva realidad de la educación ante los avances de la inteligencia artificial generativa." RIED-Revista Iberoamericana de Educación a Distancia, 27(1), 15-31. DOI: https://doi.org/10.5944/ried.27.1.37716

Hendrycks, D., et al. (2020). "Measuring Massive Multitask Language Understanding." ICLR. DOI: https://doi.org/10.48550/arXiv.2009.03300 .

Hoffmann, J., et al. (2022). "Training Compute-Optimal Large Language Models." NeurIPS. DOI: https://doi.org/10.48550/arXiv.2203.15556 .

Luo, X., et al. (2024). "Large language models surpass human experts in predicting neuroscience results." Nature Human Behaviour. DOI: https://doi.org/10.1038/s41562-024-02046-9 .

McKinsey Global Institute. (2023, junio 14). The economic potential of generative AI: The next productivity frontier. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Mendoza-Castillo, L. et al. (2024). "Perspectiva de estudiantes de nivel medio superior respecto al uso de la inteligencia artificial generativa en su aprendizaje." Apertura, 16(1), 628-647. DOI: https://doi.org/10.32870/Ap.v16n1.2343

Ouyang, L., et al. (2022). "Training language models to follow instructions with human feedback." NeurIPS. DOI: https://doi.org/10.48550/arXiv.2203.02155

Singhal, K., et al. (2023). "Large language models encode clinical knowledge." Nature, 620, 172-180. DOI: https://doi.org/10.1038/s41586-023-06291-2

Vaswani, A., et al. (2017). "Attention Is All You Need." NIPS, 5998-6008. DOI: https://doi.org/10.5555/3295222.3295349

Wei, J., et al. (2022). "Emergent Abilities of Large Language Models." Transactions on Machine Learning Research. DOI: https://doi.org/10.48550/arXiv.2206.07682