Glosario

AGI *(Artificial General Intelligence):* inteligencia artificial hipotética futura que igualaría o superaría la inteligencia humana en cualquier dominio intelectual, siendo capaz de realizar cualquier tarea intelectual que un ser humano puede hacer.

Alucinaciones: generación de información o contenido por parte de un LLM que parece plausible pero que no se basa en hechos reales o en el conocimiento adquirido durante el entrenamiento, llevando a inexactitudes o invenciones en las respuestas del modelo.

CNN *(Convolutional Neural Network)*: tipo de red neuronal especializada en procesar datos con una topología de cuadrícula, como imágenes o series temporales. Las CNN utilizan capas de convolución para extraer automáticamente características locales y abstractas de los datos, y son ampliamente utilizadas en tareas de visión por computador y procesamiento de señales.

***Cuantización*:** técnica utilizada para reducir el tamaño y acelerar la inferencia de los LLM, que consiste en reducir la precisión numérica de los pesos del modelo, pasando de números en coma flotante a representaciones de menor precisión, como enteros o números en coma fija.

Datos de entrenamiento: conjunto de ejemplos utilizados para entrenar un modelo de aprendizaje automático, que incluyen las entradas (*features*) y, en el caso del aprendizaje supervisado, las etiquetas o respuestas esperadas. La calidad y diversidad de estos datos es crucial para el rendimiento y la generalización del modelo.

Efecto Eliza: fenómeno psicológico por el cual los usuarios tienden a atribuir capacidades cognitivas y emocionales similares a las humanas a los sistemas de conversación basados en IA, a pesar de que estos sistemas no poseen una comprensión real del lenguaje ni inteligencia general.

Embeddings: representaciones densas y continuas de elementos discretos (como palabras, frases o documentos) en un espacio vectorial de alta dimensión, donde elementos similares tienen representaciones cercanas. Se utilizan en los LLM para capturar relaciones semánticas y sintácticas entre los elementos del lenguaje.

Ética de la IA: disciplina que estudia los principios morales, valores y directrices que deben guiar el desarrollo, despliegue y uso de los sistemas de inteligencia artificial, con el objetivo de garantizar que sean beneficiosos, justos, transparentes y alineados con los valores humanos.

Evaluación humana: proceso de revisión y valoración cualitativa del comportamiento y resultados de un sistema de IA por parte de expertos y usuarios, que complementa las métricas cuantitativas y permite detectar errores, sesgos o comportamientos indeseados que podrían pasar desapercibidos en una evaluación puramente automática.

Explicabilidad (XAI, *eXplainable AI*): propiedad de un modelo de IA que se refiere a su capacidad para proporcionar explicaciones comprensibles para los humanos sobre su funcionamiento interno, el razonamiento detrás de sus predicciones y los factores que influyen en sus decisiones.

Few-shot learning: capacidad de un modelo de aprendizaje automático, especialmente los LLM, para aprender a realizar una nueva tarea a partir de pocos ejemplos (desde uno hasta unas decenas), aprovechando el conocimiento previo adquirido durante el preentrenamiento en grandes cantidades de datos.

Fine-tuning: técnica para adaptar un modelo de lenguaje preentrenado a una tarea específica, mediante el entrenamiento adicional con un conjunto de datos más pequeño y especializado en esa tarea. Permite aprovechar el conocimiento general del modelo y ajustarlo para obtener un alto rendimiento en aplicaciones concretas.

***Hacking* ético:** práctica de probar y desafiar un sistema de IA de manera controlada y con permiso, con el objetivo de identificar vulnerabilidades, fallos, sesgos o comportamientos no deseados, para posteriormente corregirlos y mejorar la seguridad y robustez del sistema.

Instruction tuning: técnica de ajuste fino para LLM que consiste en proporcionar al modelo instrucciones, preguntas y ejemplos de respuestas esperadas, con el objetivo de alinear su comportamiento con las expectativas y preferencias de los usuarios en un dominio específico.

Inteligencia Artificial (IA): campo de la informática y la ingeniería que se dedica al desarrollo de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento, la percepción, la interacción en lenguaje natural y la resolución de problemas.

Inteligencia artificial generativa (GenAI): subcampo de la IA que se enfoca en la creación de modelos y algoritmos capaces de generar contenido nuevo y original, como texto, imágenes, vídeo, audio, código fuente o diseños 3D, aprendiendo patrones y características a partir de un conjunto de datos de entrenamiento.

***Large Language Models* (LLM):** modelos de aprendizaje profundo especializados en el procesamiento y generación de lenguaje natural, entrenados en enormes cantidades de texto y con un gran número de parámetros (desde millones hasta billones), capaces de realizar diversas tareas lingüísticas con un alto nivel de comprensión y coherencia.

LLMOps *(Large Language Model Operations):* conjunto de prácticas, herramientas y procesos para gestionar de manera eficiente y escalable el ciclo de vida completo de los LLM en entornos de producción, abarcando el entrenamiento, despliegue, monitorización, actualización y gobierno de estos modelos.

Machine learning: rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a los sistemas aprender y mejorar automáticamente a través de la experiencia, sin ser programados explícitamente para ello.

Machine unlearning: conjunto de técnicas para eliminar o "desaprender" de manera selectiva cierta información o sesgos indeseados de un modelo de aprendizaje automático ya entrenado, sin necesidad de reentrenarlo desde cero, permitiendo cumplir con requisitos de privacidad o corregir comportamientos no deseados.

Métricas cuantitativas: medidas numéricas estandarizadas utilizadas para evaluar de manera objetiva y consistente el rendimiento de un modelo de IA en tareas específicas, como la precisión, la exhaustividad, la exactitud o la eficiencia.

Modelo generativo: tipo de modelo de aprendizaje automático diseñado para aprender la distribución de probabilidad subyacente a un conjunto de datos y generar nuevas muestras que sean similares a los datos de entrenamiento, pudiendo crear contenido nuevo y realista.

Preentrenamiento: etapa inicial del entrenamiento de un LLM en la que se utiliza un gran corpus de texto no estructurado y sin etiquetar para que el modelo aprenda representaciones generales y patrones del lenguaje, adquiriendo un conocimiento amplio y robusto que luego puede ser adaptado a tareas específicas mediante *fine-tuning*.

Privacidad diferencial: técnica criptográfica utilizada para compartir información agregada sobre un conjunto de datos, mientras se protege la privacidad de los individuos presentes en esos datos, introduciendo un ruido aleatorio que dificulta la identificación de entradas individuales a partir de los resultados del análisis.

Prompt engineering: disciplina que se enfoca en diseñar, optimizar y adaptar los *prompts* (entradas de texto) para obtener los mejores resultados posibles de los LLM en tareas específicas, aprovechando técnicas como la inclusión de ejemplos, la especificación de formatos o la orientación paso a paso.

Pruebas A/B: método experimental utilizado para comparar el rendimiento de dos versiones diferentes de un sistema de IA (A y B) o entre un sistema de IA y un enfoque alternativo (como un humano o un modelo base), con el objetivo de determinar cuál funciona mejor según métricas predefinidas.

Regulación de la IA: conjunto de leyes, normativas, estándares y directrices establecidos por gobiernos y organizaciones para garantizar que el desarrollo, despliegue y uso de los sistemas de inteligencia artificial se realice de manera responsable, segura, ética y alineada con los valores y derechos fundamentales de la sociedad.

Retrieval-Augmented Generation (RAG): técnica utilizada en los LLM que consiste en recuperar información relevante de una base de conocimientos externa antes de generar una respuesta, combinando así la capacidad de acceso a información estructurada con la generación de lenguaje natural coherente y fluido.

RNN (Recurrent Neural Network): tipo de red neuronal diseñada para procesar secuencias de datos, como texto o series temporales. A diferencia de las redes neuronales *feedforward*, las RNN tienen conexiones recurrentes que les permiten mantener un estado interno y capturar dependencias temporales. Variantes como LSTM y GRU han sido ampliamente utilizadas en tareas de procesamiento del lenguaje natural antes del auge de los *transformers*.

Seguridad *(AI safety):* disciplina que se enfoca en identificar, prevenir y mitigar los riesgos potenciales asociados con el desarrollo y uso de sistemas de IA avanzados, tanto a corto como a largo plazo, incluyendo riesgos de seguridad, sesgos, errores, mal uso o consecuencias no deseadas.

Sesgo: tendencia sistemática de un modelo de aprendizaje automático a producir resultados que favorecen o perjudican injustamente a ciertos grupos o individuos, debido a características sensibles como el género, la etnia, la edad o la orientación sexual, y que suele ser resultado de sesgos presentes en los datos de entrenamiento o de decisiones subóptimas durante el desarrollo del modelo.

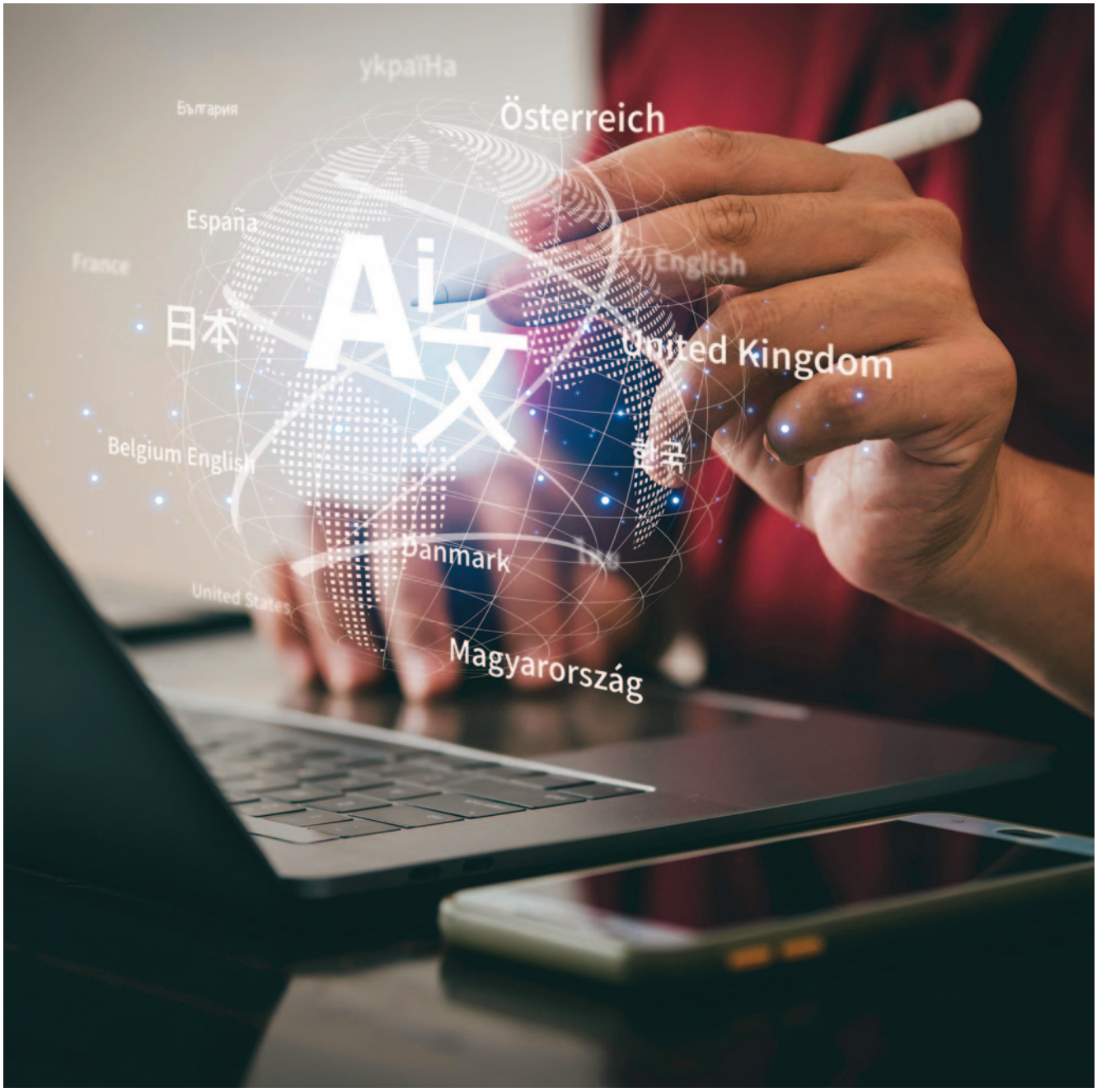
Token: unidad discreta en la que se divide un texto para su procesamiento por parte de un modelo de lenguaje. Los tokens pueden ser palabras, subpalabras o caracteres, y constituyen la entrada básica para el entrenamiento y la inferencia de los LLM.

***Tokenización*:** proceso de convertir un texto en una secuencia de tokens. La elección de la estrategia de *tokenización* tiene un impacto significativo en el rendimiento y la eficiencia del modelo.

Transformers: arquitectura de red neuronal profunda que utiliza mecanismos de atención para procesar y generar secuencias de forma paralela, en lugar de secuencialmente como las RNNs. Permite capturar dependencias a largo plazo y contextuales, siendo la arquitectura dominante para los LLM y estableciendo el estado del arte en diversas tareas de procesamiento del lenguaje natural.

Validación: proceso integral y multidisciplinar para evaluar un sistema de IA, especialmente LLM, en términos de rendimiento, robustez, seguridad, equidad, explicabilidad y alineación con los requisitos y valores éticos y sociales, combinando métricas cuantitativas y evaluación cualitativa por parte de expertos y usuarios.

Bibliografía



Abhyankar, R. et al. (2024). APIServe: Efficient API Support for Large-Language Model Inferencing. https://arxiv.org/abs/2402.01869. arXiv:2402.01869v1

Alabdulmohsin, I. et al. (2024). CLIP the Bias: How Useful is Balancing Data in Multimodal Learning? https://arxiv.org/html/2403.04547v1. arXiv:2403.04547v1

Banerjee, I., et al. (2023). MLOps with enhanced performance control and observability. https://arxiv.org/abs/2302.01061. arXiv:2302.01061v1

Bengio, Y. et al. (2003). A Neural Probabilistic Language Model. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Bréal, M. (1883). Les lois intellectuelles du langage fragment de sémantique. Annuaire de l'Association pour l'encouragement des études grecques en France. Vol. 17 (1883), pp. 132-142. https://www.jstor.org/stable/44253893

Cambon, A. et al. (2023). Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity. A first update from Microsoft's research initiative on AI and Productivity.

Chen, D. et al. (2023). Data-Juicer: A One-Stop Data Processing System for Large Language Models. https://arxiv.org/abs/2309.02033. arXiv:2309.02033v3

Chen, Y. et al. (2023). LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. https://arxiv.org/abs/2309.12307. arXiv:2309.12307v3

Chiang, C. et al. (2023). Can Large Language Models Be an Alternative to Human Evaluations? https://arxiv.org/abs/2305.01937. arXiv:2305.01937v1

Chu, T., Song, Z., Yang, C. (2023). How to Protect Copyright Data in Optimization of Large Language Models? https://arxiv.org/abs/2308.12247. arXiv:2308.12247v1

CIO (2023). Chief AI Officer: What it takes to land the C-suite's hottest new job. https://www.cio.com/article/657977/chief-ai-officer-what-it-takes-to-land-the-c-suites-hottest-new-job.html

Cui, Q. et al. (2022). Contrastive Vision-Language Pre-training with Limited Resources. https://arxiv.org/abs/2112.09331. arXiv:2112.09331v3

CommetML. https://www.comet.com/site/.

Datta, T. et al. (2023). Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook. https://arxiv.org/abs/2303.06223. arXiv:2303.06223v1

Dettmers, T. et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs https://arxiv.org/abs/2305.14314. arXiv:2305.14314v1

Devlin, J. et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805. arXiv:1810.04805v2

Duan, J. et al. (2023). Shifting attention to relevance: towards the uncertainty estimation of large language models. https://arxiv.org/abs/2307.01379. arXiv:2307.01379v2

Dun, C. et al. (2024). Sweeping Heterogeneity with Smart MoPs: Mixture of Prompts for LLM Task Adaptation. https://arxiv.org/abs/2310.02842. arXiv:2310.02842v2

Elazar, Y. et al. (2021). Measuring and Improving Consistency in Pretrained Language Models. https://aclanthology.org/2021.tacl-1.60/.

Euronews (2023). 2023 was the year AI went mainstream. It was also the year we started to panic about it. https://www.euronews.com/next/2023/12/27/2023-was-the-year-ai-went-mainstream-it-was-also-the-year-we-started-to-panic-about-it

European Parliament (2024). Artificial Intelligence Act / European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). https://artificialintelligenceact.eu/; https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

European Commission (2024). Knowledge Center on Interpretation. https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/what-large-language-model

Fisher, M., Campagna, G., Choi, E., Lam, M. S., Freund, S. N., Yahav, E.,. (2021). DIY Assistant: A Multi-modal End-User Programmable Virtual Assistant. https://dl.acm.org/doi/10.1145/3453483.3454046.

Gartner (2023). What is generative AI? https://www.gartner.com/en/topics/generative-ai

Google DeepMind (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. Meredith Ringel Morris; Jascha Sohl-Dickstein; Noah Fiedel; Tris Warkentin; Allan Dafoe; Aleksandra Faust; Clement Farabet; and Shane Legg. arXiv:2311.02462v1

Google + Implement (2023). The economic opportunity of generative AI in D9+. An Implement Consulting Group study commissioned by Google.

Gozalo-Brizuela, R., y Garrido-Merchán, E.C. (2023). A survey of Generative AI Applications. https://ar5iv.labs.arxiv.org/html/2306.02781

Guo, Z. et al. (2023). Evaluating Large Language Models: A Comprehensive Survey. https://arxiv.org/pdf/2310.19736.pdf. arXiv:2310.19736v3

Guzman, F. et al. (2015). How do Humans Evaluate Machine Translation. https://aclanthology.org/W15-3059.pdf.

Fu, HY. et al. (2023). Estimating Large Language Model Capabilities without Labeled Test Data. https://arxiv.org/abs/2305.14802. arXiv:2305.14802v2

Fu, X. et al (2024). Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization? https://arxiv.org/abs/2402.00841. arXiv:2402.00841

Goyal, S. et al (2024). LLMGuard: Guarding Against Unsafe LLM Behavior. https://arxiv.org/abs/2403.00826. arXiv:2403.00826v1

Hendrycks, D. et al (2021). Measuring Massive Multitask Language Understanding. https://arxiv.org/abs/2009.03300. arXiv:2009.03300v3

Huang, L. et al. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. https://arxiv.org/abs/2311.05232. arXiv:2311.05232v1

Hugging Face Datasets (2024). CodeParrot. https://huggingface.co/codeparrot.

IAPP (2024). Global AI Law and Policy Tracker. https://iapp.org/resources/article/global-ai-legislation-tracker/

iDanae 2T23 (2023): Large Language Models: una nueva era en la inteligencia artificial. Cátedra iDanae. Newsletter trimestral 2T23. http://www.idanae-stem.com/

iDanae 1T24 (2024): Hacia una inteligencia artificial sostenible. Cátedra iDanae. Newsletter trimestral 1T24. http://www.idanae-stem.com/

Imperial, JM., et al. (2023). Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. https://arxiv.org/abs/2309.05454. arXiv:2309.05454v2

IndesIA (2024). Barómetro de adopción de la inteligencia artificial en las pymes españolas. https://www.indesia.org/wp-content/uploads/2024/04/IndesIA.-Barometro-de-adopcion-de-la-inteligencia-artificial-en-las-pymes-espanolas-Edicion-2024.pdf

Jang et al. (2022). Knowledge unlearning for mitigating privacy risks in language models. https://arxiv.org/abs/2210.01504. arXiv:2210.01504.

Jia, C. et al (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. https://arxiv.org/abs/2102.05918. arXiv:2102.05918v2

Kahng, M. et al. (2024). LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. https://arxiv.org/abs/2402.10524. arXiv:2402.10524v1

Kuchnik, M. et al. (2023). Validating Large Language Models with Realm. https://arxiv.org/abs/2211.15458. arXiv:2211.15458v2

Kudo, T., Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. https://arxiv.org/abs/1808.06226. arXiv:1808.06226v1

Lam, M. (2018). https://profiles.stanford.edu/monica-lam?tab=publications. Keeping the Internet Open with an Open-Source Virtual Assistant.

Lee, C. et al (2024). OrchestraLLM: Efficient Orchestration of Language Models for Dialogue State Tracking. https://arxiv.org/html/2311.09758v2. arXiv:2311.09758v2

Lee, J. et al. (2022). Seq2Seq-SC: End-to-End Semantic Communication Systems with Pre-trained Language Model. https://arxiv.org/abs/2210.15237. arXiv:2210.15237v2

Lester, B. et al. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. https://arxiv.org/abs/2104.08691. arXiv:2104.08691v2

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. https://arxiv.org/abs/2005.11401

Li, H. et al. (2024). Digger: Detecting Copyright Content Mis-usage in Large Language Model Training. https://arxiv.org/abs/2401.00676. arXiv:2401.00676v1

Li, S. et al (2024). Evaluating Quantized Large Language Models. https://arxiv.org/abs/2402.18158. arXiv:2402.18158v1

Li, Y. et al (2023). A Survey on Fairness in Large Language Models. https://arxiv.org/abs/2308.10149. arXiv:2308.10149.

Liang, P. et al. (2023). Holistic Evaluation of Language Models. https://arxiv.org/abs/2211.09110. arXiv:2211.09110v2

Liu, T. et al (2022). Autoregressive Structured Prediction with Language Models. https://arxiv.org/abs/2210.14698. arXiv:2210.14698v2

Liu, Y. et al (2024). Datasets for Large Language Models: A Comprehensive Survey. https://arxiv.org/abs/2402.18041. arXiv:2402.18041v1

Liu, Y. et al (2023). Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models. https://arxiv.org/pdf/2308.07847.pdf. arXiv:2308.07847v1

Luo, Y. et al. (2023). An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. https://arxiv.org/pdf/2308.08747.pdf. arXiv:2308.08747v3

Management Solutions (2023). Explainable Artificial Intelligence (XAI): desafíos en la interpretabilidad de los modelos. https://www.managementsolutions.com/en/microsites/whitepapers/explainable-artificial-intelligence

Management Solutions (2022). AutoML, hacia la automatización de los modelos. https://www.managementsolutions.com/es/publicaciones-y-eventos/informes-sectoriales/white-papers/auto-machine-learning-hacia-la-automatizacion-de-los-modelos

Management Solutions (2014). Model Risk Management: Quantitative and Qualitative Aspects. Model Risk Management: Quantitative and qualitative aspects | Management Solutions

Meeus, M. et al. (2024). Copyright Traps for Large Language Models. https://arxiv.org/abs/2402.09363. arXiv:2402.09363v1

Mehta, S.V. et al. (2023). An Empirical Investigation of the Role of Pre-training in Lifelong Learning. https://arxiv.org/abs/2112.09153. arXiv:2112.09153v2

Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781. arXiv:1301.3781v3.

Minaee, S. et al. (2024). Large Language Models: A Survey. https://arxiv.org/abs/2402.06196. arXiv:2402.06196v2

MindsDB (2024). A Comparative Analysis of Leading Large Language Models. https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models

Mökander, J. et al. (2023). Auditing large language models: a three-layered approach. arXiv:2302.08500v2

Nasr, M., et al. (2023). https://arxiv.org/pdf/2311.17035.pdf. arXiv:2311.17035v1

Neelakantan, A. et al. (2022). Text and Code Embeddings by Contrastive Pre-Training. https://arxiv.org/abs/2201.10005. arXiv:2201.10005v1

NIST (2023). AI Risk Management Framework | NIST. https://www.nist.gov/itl/ai-risk-management-framework

Oneto, L., Chiappa, S. (2020). Fairness in Machine Learning. 2012.15816.pdf (arxiv.org) arXiv:2012.15816v1

OpenAI (2024). Prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering

Ovadia, O. et al (2024). Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. https://arxiv.org/pdf/2312.05934.pdf. arXiv:2312.05934v3

Pankajakshan, R. et al (2024). Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal. https://arxiv.org/html/2403.13309v1. arXiv:2403.13309v1.

Parikh, A. P., et al. (2016). A Decomposable Attention Model for Natural Language Inference. https://arxiv.org/abs/1606.01933. arXiv:1606.01933v2

Penedo, G. et al (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. https://arxiv.org/abs/2306.01116. arXiv:2306.01116v1

Pew Research Center (2023). Experts Predict the Best and Worst Changes in Digital Life by 2035.

Project Gutenberg (2024). https://www.gutenberg.org/.

Rae, JW, et al (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. https://arxiv.org/abs/2112.11446. arXiv:2112.11446

Rafailov, R. et al (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. https://arxiv.org/abs/2305.18290. arXiv:2305.18290v2

Rejeleene, R.; Xu, X.; Talburt, J.; (2024). Towards Trustable Language Models: Investigating Information Quality of Large Language Models. https://arxiv.org/abs/2401.13086. arXiv:2401.13086v1

Risk.net. (2024). The bank quant who wants to stop gen AI hallucinating. https://www.risk.net/risk-management/7959062/the-bank-quant-who-wants-to-stop-gen-ai-hallucinating.

Sachdeva, N., et al (2024). How to Train Data-Efficient LLMs. https://arxiv.org/html/2402.09668v1. arXiv:2402.09668v1

Samsi, S., et al (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. https://arxiv.org/pdf/2310.03003.pdf. arXiv:2310.03003v1

Sarti, G. et al (2023). Inseq: An Interpretability Toolkit for Sequence Generation Models. [2302.13942] Inseq: An Interpretability Toolkit for Sequence Generation Models (arxiv.org). arXiv:2302.13942v3

Searle, J. (1980). Minds, Brains, and Programs. The Behavioral and Brain Sciences, vol. 3. Cambridge University Press. https://web.archive.org/web/20010221025515/http://www.bbsonline.org/Preprints/OldArchive/bbs.searle2.html

Shaikh, O. et al. (2022). On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. https://arxiv.org/abs/2212.08061. arXiv:2212.08061v2

SHAP documentation. https://shap.readthedocs.io/

Shaw, P. et al (2018). Self-Attention with Relative Position Representations. https://arxiv.org/abs/1803.02155. arXiv:1803.02155v2

Sherstinsky, A. (2018). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. https://arxiv.org/abs/1808.03314. arXiv:1808.03314v10

Shi, W. et al (2024). Detecting pretraining data from large language models. https://arxiv.org/abs/2310.16789. arXiv:2310.16789v3

Singh, C. et al (2024). Rethinking Interpretability in the Era of Large Language Models. https://arxiv.org/abs/2402.01761. arXiv:2402.01761v1

Sinha, K. et al (2021). Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. https://arxiv.org/abs/2104.06644. arXiv:2104.06644v2

Soskek (2019). BookCorpus. https://github.com/soskek/bookcorpus.

Su, J., et al (2021). Roformer: Enhanced transformer with rotary position embedding. https://arxiv.org/abs/2104.09864. arXiv:2104.09864.

Sutskever, I. et al (2014). Sequence to Sequence Learning with Neural Networks. https://arxiv.org/abs/1409.3215. arXiv:1409.3215v3

The Next Web (2023). When will AGI arrive? Here's what our tech lords predict. https://thenextweb.com/news/when-will-agi-arrive-tech-experts-predict-artificial-general-intelligence

Tian, Y. et al (2024). TinyLLM: Learning a Small Student from Multiple Large Language Models. https://arxiv.org/abs/2402.04616. arXiv:2402.04616

Tirumala, K. et al. (2023). D4: Improving LLM Pretraining via Document De-Duplication and Diversification. https://arxiv.org/abs/2308.12284. arXiv:2308.12284v1

UK Government (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Vartziotis, T. et al (2024). Learn to Code Sustainably: An Empirical Study on LLM-based Green Code Generation. https://arxiv.org/html/2403.03344v1. arXiv:2403.03344v1.

Vaswani, A. et al. (2017). Attention Is All You Need. https://arxiv.org/abs/1706.03762

Wan, Z. et al (2024). Efficient Large Language Models: A Survey. https://arxiv.org/pdf/2312.03863.pdf. arXiv:2312.03863v3

Wang, Q. et al (2024). LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools. [2401.12576] LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools (arxiv.org). arXiv:2401.12576v1

Wang, Y. et al (2024). Two-stage LLM Fine-tuning with Less Specialization and More Generalization. https://arxiv.org/html/2211.00635v3. arXiv:2211.00635v3

Wei, J. et al (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://arxiv.org/abs/2201.11903. arXiv:2201.11903v6

Wenzek, G., et al (2019). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. https://arxiv.org/abs/1911.00359. arXiv:1911.00359v2

Wettig, A. et al. (2024). QuRating: Selecting High-Quality Data for Training Language Models. https://arxiv.org/abs/2402.09739. arXiv:2402.09739v1

Weights & Biases: The AI Developer Platform (wandb.ai). https://wandb.ai/site

Wikipedia (2024). Dumps. https://dumps.wikimedia.org/zhwiki/latest/.

Wired (2023). OpenAI's CEO Says the Age of Giant AI Models Is Already Over. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. https://dl.acm.org/doi/10.1145/365153.365168

White House (2022). Blueprint for an AI Bill Of Rights. Making Automated Systems Work for the American People. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Wu, X. et al. (2023). Depn: Detecting and editing privacy neurons in pretrained language models. https://arxiv.org/abs/2310.20138. arXiv:2310.20138.

Xin Zhao, W., et al. (2023). A Survey of Large Language Models. https://arxiv.org/abs/2303.18223. arXiv:2303.18223v13

Xu, L. et al. (2023). Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. https://arxiv.org/pdf/2312.12148.pdf. arXiv:2312.12148v1

Xu, Y. et al. (2021). Non-Autoregressive Text Generation with Pre-trained Language Models. https://aclanthology.org/2021.eacl-main.18/

Xu, Z. et al. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. https://arxiv.org/abs/2401.11817. arXiv:2401.11817v1

Yang, J. et al. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. https://arxiv.org/abs/2304.13712. arXiv:2304.13712v2

Yidiz, C. et al (2024). Investigating Continual Pretraining in Large Language Models: Insights and Implications. https://arxiv.org/html/2402.17400v1. arXiv:2402.17400v1

Yu, C. et al. (2023). Unlearning bias in language models by partitioning gradients. https://aclanthology.org/2023.findings-acl.375.pdf.

Yogarajan, V., et al (2023). Tackling Bias in Pre-trained Language Models: Current Trends and Under-represented Societies. https://arxiv.org/pdf/2312.01509.pdf. arXiv:2312.01509v1

Zaharia, M. et al (2018). Accelerating the Machine Learning Lifecycle with MLflow. https://people.eecs.berkeley.edu/~matei/papers/2018/ieee_mlflow.pdf.

Zeng, Y., et al (2023). CLIP2: Contrastive Language-Image-Point Pretraining from Real-World Point Cloud Data. https://arxiv.org/abs/2303.12417. arXiv:2303.12417v2

Zhang, B. et al (2024). When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. https://arxiv.org/abs/2402.17193. arXiv:2402.17193v1

Zhang, L. et al (2024). Enhancing Large Language Model Performance To Answer Questions and Extract Information More Accurately. https://arxiv.org/html/2402.01722v1. arXiv:2402.01722v1.

Zhang, S. et al (2023). Instruction Tuning for Large Language Models: A Survey. https://www.researchgate.net/publication/373263398_Instruction_Tuning_for_Large_Language_Models_A_Survey.

Zhang, Y. et al (2024). Bias Mitigation in Fine-tuning Pre-trained Models for Enhanced Fairness and Efficiency. https://arxiv.org/html/2403.00625v1. arXiv:2403.00625v1

Zhao, B., et al (2023). Tuning LayerNorm in Attention: Towards Efficient Multi-Modal LLM Finetuning. https://arxiv.org/abs/2312.11420. arXiv:2312.11420v1

Zhou, C. et al (2023). LIMA: Less Is More for Alignment. https://arxiv.org/abs/2305.11206. arXiv:2305.11206v1

Zhou, N., et al (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. https://arxiv.org/abs/2105.06558. arXiv:2105.06558v1