

# Glossário



**AGI (Artificial General Intelligence):** futura inteligência artificial hipotética que igualaria ou superaria a inteligência humana em qualquer domínio intelectual, sendo capaz de realizar qualquer tarefa intelectual que um ser humano possa fazer.

**Alucinações:** geração de informações ou conteúdo por um LLM que parece plausível, mas não se baseia em fatos ou conhecimentos reais adquiridos durante o treinamento, levando a imprecisões ou invenções nas respostas do modelo.

**Avaliação humana:** processo de revisão e avaliação qualitativa do comportamento e dos resultados de um sistema de IA por especialistas e usuários, que complementa as métricas quantitativas e permite a detecção de erros, vieses ou comportamentos indesejados que podem passar despercebidos em uma avaliação puramente automática.

**CNN (Convolutional Neural Network):** tipo de rede neural especializada no processamento de dados com uma topologia de grade, como imagens ou séries temporais. As CNNs usam camadas convolucionais para extrair automaticamente recursos locais e abstratos dos dados e são amplamente usadas em tarefas de visão computacional e processamento de sinais.

**Dados de treinamento:** conjunto de exemplos usados para treinar um modelo de aprendizado de máquina, incluindo as entradas (features) e, no caso do aprendizado supervisionado, os rótulos ou as respostas esperadas. A qualidade e a diversidade desses dados são fundamentais para o desempenho e a generalização do modelo.

**Efeito Eliza:** fenômeno psicológico em que os usuários tendem a atribuir capacidades cognitivas e emocionais semelhantes às humanas aos sistemas de conversação baseados em IA, apesar de esses sistemas não terem nenhum entendimento real de linguagem ou inteligência geral.

**Embeddings:** representações densas e contínuas de elementos discretos (como palavras, frases ou documentos) em um espaço vetorial de alta dimensão, em que elementos semelhantes têm representações próximas. Eles são usados em LLMs para capturar relações semânticas e sintáticas entre elementos de linguagem.

**Ética em IA:** disciplina que estuda os princípios morais, os valores e as diretrizes que devem orientar o desenvolvimento, a implantação e o uso de sistemas de inteligência artificial, com o objetivo de garantir que eles sejam benéficos, justos, transparentes e alinhados aos valores humanos.

**Explicabilidade (XAI, eXplainable AI):** propriedade de um modelo de IA que se refere à sua capacidade de fornecer explicações compreensíveis para humanos sobre seu funcionamento interno, o raciocínio por trás de suas previsões e os fatores que influenciam suas decisões.

**Few-shot learning:** capacidade de um modelo de aprendizado de máquina, especialmente LLMs, de aprender a executar uma nova tarefa com poucos exemplos (de um a algumas dezenas), aproveitando o conhecimento prévio adquirido durante o pré-treinamento em grandes quantidades de dados.

**Fine-tuning:** técnica para adaptar um modelo de linguagem pré-treinado a uma tarefa específica por meio de treinamento adicional com um conjunto de dados menor e mais especializado para essa tarefa. Ela permite aproveitar o conhecimento geral do modelo e ajustá-lo para obter alto desempenho em aplicativos específicos.

**Hacking ético:** prática de testar e desafiar um sistema de IA de forma controlada e autorizada, com o objetivo de identificar vulnerabilidades, falhas, vieses ou comportamentos indesejados e, em seguida, corrigi-los para melhorar a segurança e a robustez do sistema.



**Instruction tuning:** técnica de ajuste fino para LLM que consiste em fornecer ao modelo instruções, perguntas e exemplos de respostas esperadas, com o objetivo de alinhar seu comportamento com as expectativas e preferências dos usuários em um domínio específico.

**Inteligência artificial (IA):** campo da ciência e engenharia da computação dedicado ao desenvolvimento de sistemas capazes de realizar tarefas que normalmente exigem inteligência humana, como aprendizado, raciocínio, percepção, interação com linguagem natural e solução de problemas.

**Inteligência Artificial Generativa (GenAI):** subcampo da IA que se concentra na criação de modelos e algoritmos capazes de gerar conteúdo novo e original, como texto, imagens, vídeo, áudio, código-fonte ou projetos em 3D, aprendendo padrões e recursos de um conjunto de dados de treinamento.

**Large Language Models (LLM):** modelos de aprendizagem profunda especializados em processamento e geração de linguagem natural, treinados em grandes quantidades de texto e com um grande número de parâmetros (de milhões a bilhões), capazes de executar várias tarefas linguísticas com um alto nível de compreensão e coerência.

**LLMOps (Large Language Model Operations):** conjunto de práticas, ferramentas e processos para gerenciar de forma eficiente e dimensionável todo o ciclo de vida dos LLMs em ambientes de produção, abrangendo treinamento, implantação, monitoramento, atualização e governança desses modelos.

**Machine learning:** ramo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem que os sistemas aprendam e melhorem automaticamente por meio da experiência, sem serem explicitamente programados para isso.

**Machine unlearning:** conjunto de técnicas para remover ou "desaprender" seletivamente determinadas informações ou vieses indesejados de um modelo de aprendizado de máquina já treinado, sem a necessidade de treiná-lo novamente do zero, o que permite cumprir os requisitos de privacidade ou corrigir comportamentos indesejados.

**Métricas quantitativas:** medidas numéricas padronizadas usadas para avaliar de forma objetiva e consistente o desempenho de um modelo de IA em tarefas específicas, como exatidão, integridade, precisão ou eficiência.

**Modelo generativo:** tipo de modelo de aprendizado de máquina projetado para aprender a distribuição de probabilidade subjacente de um conjunto de dados e gerar novas amostras que sejam semelhantes aos dados de treinamento, podendo criar conteúdo novo e realista.

**Pré-treinamento:** estágio inicial do treinamento do LLM no qual um grande corpus de texto não estruturado e não rotulado é usado para que o modelo aprenda representações e padrões gerais do idioma, adquirindo um conhecimento amplo e robusto que pode ser adaptado a tarefas específicas por meio de ajuste fino.

**Privacidade diferencial:** técnica criptográfica usada para compartilhar informações agregadas sobre um conjunto de dados e, ao mesmo tempo, proteger a privacidade dos indivíduos presentes nesses dados, introduzindo ruído aleatório que dificulta a identificação de entradas individuais a partir dos resultados da análise.

**Prompt engineering:** disciplina que se concentra em projetar, otimizar e adaptar prompts (entradas de texto) para obter os melhores resultados possíveis dos LLMs em tarefas específicas, aproveitando técnicas como a inclusão de exemplos, especificação de formatos ou orientação passo a passo.

**Quantização:** técnica usada para reduzir o tamanho e acelerar a inferência de LLMs, reduzindo a precisão numérica dos pesos do modelo, passando de números de ponto flutuante para representações de menor precisão, como números inteiros ou de ponto fixo.

**Regulação da IA:** conjunto de leis, regulamentações, normas e diretrizes estabelecidas por governos e organizações para garantir que o desenvolvimento, a implantação e o uso de sistemas de inteligência artificial sejam conduzidos de forma responsável, segura, ética e alinhada com os valores e direitos fundamentais da sociedade.

**Retrieval-Augmented Generation (RAG):** técnica usada em LLMs que consiste em recuperar informações relevantes de uma base de conhecimento externa antes de gerar uma resposta, combinando assim a capacidade de acessar informações estruturadas com a geração de linguagem natural coerente e fluente.

**RNN (Recurrent Neural Network):** tipo de rede neural projetada para processar sequências de dados, como texto ou séries temporais. Diferentemente das redes neurais feedforward, as RNNs têm conexões recorrentes que lhes permitem manter o estado interno e capturar dependências temporais. Variantes como LSTM e GRU têm sido amplamente usadas em tarefas de processamento de linguagem natural antes do surgimento dos transformers.

**Segurança (AI safety):** disciplina que se concentra na identificação, prevenção e mitigação dos possíveis riscos associados ao desenvolvimento e ao uso de sistemas avançados de IA, tanto a curto quanto a longo prazo, incluindo riscos de segurança, vieses, erros, uso indevido ou consequências não intencionais.

**Testes A/B:** método experimental usado para comparar o desempenho de duas versões diferentes de um sistema de IA (A e B) ou entre um sistema de IA e uma abordagem alternativa (como um humano ou um modelo de base), a fim de determinar qual apresenta melhor desempenho de acordo com métricas predefinidas.

**Token:** unidade discreta na qual um texto é dividido para processamento por um modelo de linguagem. Os tokens podem ser palavras, subpalavras ou caracteres, e são a entrada básica para o treinamento e a inferência do LLM.

**Tokenização:** processo de conversão de texto em uma sequência de tokens. A escolha da estratégia de tokenização tem um impacto significativo sobre o desempenho e a eficiência do modelo.

**Transformers:** arquitetura de rede neural profunda que usa mecanismos de atenção para processar e gerar sequências em paralelo, em vez de sequencialmente como as RNNs. Ela permite capturar dependências contextuais e de longo prazo, sendo a arquitetura dominante para LLMs e definindo o estado da arte em várias tarefas de processamento de linguagem natural.

**Validação:** processo abrangente e multidisciplinar para avaliar um sistema de IA, especialmente o LLM, em termos de desempenho, robustez, segurança, imparcialidade, explicabilidade e alinhamento com requisitos e valores éticos e sociais, combinando métricas quantitativas e avaliação qualitativa por especialistas e usuários.

**Viés:** tendência sistemática de um modelo de aprendizado de máquina de produzir resultados que favorecem ou desfavorecem injustamente determinados grupos ou indivíduos, devido a características sensíveis, como gênero, etnia, idade ou orientação sexual, e geralmente é o resultado de vieses nos dados de treinamento ou de decisões abaixo do ideal durante o desenvolvimento do modelo.

## Referências



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 Misusage 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 - Desafios na interpretabilidade de modelos.. <https://www.managementsolutions.com/pt-br/microsites/whitepapers/explainable-artificial-intelligence>

Management Solutions (2022). Auto Machine Learning, rumo à automação dos modelos <https://www.managementsolutions.com/pt-br/publicacoes-e-eventos/informes-sectoriais/white-papers/auto-machine-learning-rumo-a-automacao-dos-modelos>

Management Solutions (2014). Model Risk Management: Quantitative and Qualitative Aspects. <https://www.managementsolutions.com/pt-br/publicacoes-e-eventos/informes-sectoriais/white-papers/model-risk-management>

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](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](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