A Survey of Large Language Models: Architectures, Applications, Challenges, and Future Directions

Ganjeme Kane
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
benjaminlane82@colostate.edu

Mork Neran
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
markwork@ku.edu

Abstract

Large language models have rapidly shifted from specialized natural language processing components to general-purpose computational infrastructures that support a broad range of institutional, commercial, and public-sector applications. This survey examines the field from a system-level perspective, emphasizing the structural trade-offs that arise across model architectures, training infrastructures, deployment environments, and governance regimes. It begins by tracing the conceptual foundations of large language models, including the rise of self-supervised pretraining, unified text-to-text interfaces, and emergent capabilities. The discussion then moves to architectural choices such as dense and sparse parameterization, context length, retrieval augmentation, and tool integration, showing how these decisions shape inference cost, robustness, and accountability. Training infrastructures are analyzed as distributed systems in which data selection, parallelization, post-training, and continuous adaptation introduce complex dependencies. The survey further addresses alignment, robustness, safety, fairness, privacy, and truthfulness as emergent properties of entire sociotechnical pipelines rather than isolated model characteristics. Governance and policy implications are considered alongside sustainability and operational deployment, with attention to energy consumption, carbon reporting, auditability, and regulatory compliance. Future directions are explored through the lens of multimodal extension, agentic workflows, long-horizon reasoning, open science, and participatory evaluation. Throughout, the survey stresses that large language models cannot be understood purely as machine learning artifacts; they must be analyzed as infrastructures embedded within data supply chains, organizational workflows, legal systems, and public values. The conclusion identifies systemic research priorities for building more reliable, equitable, and sustainable language model ecosystems.

Keywords

large language models; transformer architectures; scaling; alignment; governance; sustainability; fairness; socio-technical systems.

1. Introduction

Large language models have become central to contemporary artificial intelligence research and practice. They are increasingly treated not as task-specific natural language processing components but as general-purpose substrates that can be adapted to many downstream uses. The foundation model perspective [1] captures this shift by emphasizing that a single model trained on broad data can support diverse applications through prompting, fine-tuning, or

external tool use. This development has transformed how organizations approach machine learning system design. Instead of training isolated models for each task, system architects now reason about shared pretrained representations, adaptation interfaces, operational costs, data dependencies, and downstream accountability. The significance of large language models therefore extends beyond accuracy benchmarks. Their deployment raises structural questions about infrastructure, robustness, fairness, energy consumption, privacy, and regulatory compliance. This survey adopts a systems orientation, treating large language models as sociotechnical infrastructures rather than isolated statistical predictors.

Early evidence for this transition appeared with few-shot learning in very large autoregressive models [2]. Practitioners could provide a small number of demonstrations in the prompt and obtain useful behavior across translation, question answering, and classification without updating parameters. In parallel, bidirectional pretraining [3] demonstrated that masked language modeling could produce high-quality contextual representations for language understanding tasks, leading to widespread industrial adoption. The underlying sequence modeling paradigm was enabled by self-attention [4], which replaced recurrence and convolution with a mechanism that computes dependencies across all positions in a sequence. Unified text-to-text formulations [5] further consolidated diverse tasks into a single interface, reducing the amount of task-specific engineering required for transfer learning. As models grew, researchers observed emergent abilities [6] that were not explicitly programmed and that became visible only after particular scale thresholds. These observations have motivated several surveys [7] that document technical progress, evaluation practices, and unresolved risks. This paper builds on that literature by emphasizing the system-level trade-offs that shape whether large language models can be deployed responsibly and sustainably.

2. Architectural Foundations and System-Level Design Trade-offs

Current large language models are predominantly based on transformer architectures, but the design space has expanded considerably. One central trade-off concerns dense versus sparse parameterization. Compute-optimal scaling analyses [8] indicate that model size and training data quantity should grow in a balanced manner, challenging earlier assumptions that emphasized parameter count alone. Dense models such as PaLM [9] demonstrate that very large decoder-only architectures can perform well across multilingual and reasoning tasks when trained on a heterogeneous mixture of text and code. At the same time, the engineering costs of dense scaling motivate more efficient approaches. Open models such as LLaMA [10] show that carefully curated data and longer training can yield competitive performance with reduced inference cost, illustrating that parameter count alone does not determine capability. Sparse mixture-of-expert designs [11] go further by activating only a subset of parameters per token, reducing computation while preserving model capacity. These architectural alternatives create different deployment profiles: dense models may be simpler to parallelize but more expensive at inference; sparse models may offer lower latency per token but introduce routing complexity, load balancing concerns, and communication overhead.

System-level trade-offs also involve context length, attention complexity, and memory layout. Autoregressive decoding is sequential in nature, which makes inference latency and memory bandwidth critical constraints for production systems. Encoder-decoder models [5] provide a natural interface for conditional text generation and can be efficient for tasks with fixed input-output mappings, but they are less common in recent general-purpose systems that favor decoder-only designs. The choice of positional encoding, normalization scheme, and activation function has direct implications for numerical stability during distributed training

and for quantization after deployment. Quantization, pruning, and distillation are not merely post-hoc optimizations; they influence the failure modes of the system, the interpretability of internal representations, and the ease with which a model can be updated or corrected. Consequently, architecture selection cannot be separated from the operational environment in which the model will serve.

Another architectural consideration is the interface to external tools and memory. Systems increasingly augment language models with retrieval modules, calculators, code interpreters, or database connectors. These additions change the trust boundary of the system: a language model that calls external APIs can propagate errors or introduce security vulnerabilities if the outputs are not validated. Architectural extensions such as retrieval-augmented generation modify the capacity-accuracy trade-off by allowing models to consult non-parametric knowledge sources, but they require careful orchestration of chunking, indexing, authorization, and versioning. System architects must therefore evaluate whether a monolingual, multilingual, code-oriented, or tool-augmented model is appropriate for a given deployment context. The decision is not purely technical; it also shapes accountability, because a system that autonomously queries external resources may produce decisions whose provenance is harder to trace.

3. Training Infrastructures, Data, and Scaling Dynamics

Training a large language model is a distributed systems problem as much as a machine learning problem. Models are partitioned across accelerators using data, tensor, and pipeline parallelism, and each parallelization strategy introduces distinct communication bottlenecks and failure recovery requirements. The data mix also strongly shapes model behavior. Training corpora now include web pages, books, code, scientific articles, and dialogue, but these sources differ in quality, legality, and representational coverage. Deduplication and filtering can improve data efficiency, but they can also remove dialectal or minority perspectivesDeduplication and filtering can improve data efficiency, but they can also remove dialectal or minority perspectives if curation is driven only by fluency or popularity metrics. The construction of training corpora therefore raises deep questions about representation, consent, and historical bias. Sensitive information, licensed content, and user-generated text may appear in the data mix, often without clear provenance. The absence of systematic documentation for web-scale corpora has been identified as a limitation for responsible use [15]. When models memorize and later reproduce such content, they can create legal and ethical liabilities for downstream operators. Rigorous data governance must include provenance tracking, license auditing, and ongoing contamination detection, because the model is only as trustworthy as the data pipeline that produced it.

Training infrastructure also shapes the ability to audit and reproduce model behavior. Distributed training runs involve many nondeterministic events, including hardware failures, gradient noise, and scheduling decisions, which can introduce subtle variation between runs. This variation complicates the comparison of systems and the attribution of failures. Furthermore, post-training procedures such as instruction tuning and reinforcement learning from human feedback depend on human annotations whose quality and diversity are difficult to verify at scale [12]. These procedures can shift model behavior in desirable directions, but they also introduce new sources of bias because the annotator pool may not represent all user communities. Consequently, training is not a one-time event but an ongoing socio-technical process that includes data selection, pretraining, alignment, evaluation, and continuous adaptation.

4. Alignment, Robustness, and Safety

Alignment is the process of shaping model behavior to better reflect human intentions, but the concept is contested and context dependent. Instruction tuning has become a standard method for making models more usable across tasks [12]. By training on diverse prompts and human demonstrations, models learn to follow instructions more reliably, yet this does not guarantee that their responses are truthful or harmless. Reinforcement learning with human preference data can further encourage helpfulness and reduce obvious toxicity, but it is expensive and subject to annotator disagreement. Constitutional approaches attempt to scale oversight by having the model critique its own outputs against a set of principles [13]. While this reduces direct human labeling cost, it raises concerns about whether the principles are sufficiently explicit, contestable, and inclusive. Alignment is therefore not a single optimization target; it is a continuous negotiation among competing values such as accuracy, politeness, safety, privacy, and user autonomy.

Robustness failures in large language models are often discovered after deployment rather than anticipated before release. Models can be manipulated through adversarial prompts, produce confident but false statements, or leak sensitive training data. Extraction attacks have shown that it is sometimes possible to recover memorized training examples, including names, addresses, and other personal information [19]. This challenges the assumption that fine-tuning or instruction tuning fully removes sensitive content. Calibration and uncertainty estimation remain difficult for autoregressive models because token-level probabilities do not directly capture the reliability of a factual claim. Truthfulness benchmarks have revealed that larger models do not automatically become more truthful and can sometimes produce more plausible falsehoods [25]. These observations suggest that robustness and safety cannot be achieved solely by scale; they require explicit evaluation and mitigation at the system level.

Fairness in language models is complicated by the fact that harms may arise from associations, stereotypes, or the erasure of minority language varieties. Audits have demonstrated persistent negative associations in model outputs even when the model is not explicitly prompted with demographic attributes [20]. Because language models are used in hiring assistance, content moderation, customer support, and educational tools, biased outputs can reproduce structural inequality. System designers must decide whether to address harms through data curation, representation constraints, post-training debiasing, or controlled generation interfaces. Each intervention has limitations. Removing certain terms from training data can reduce one form of bias while hiding others, and debiasing methods may distort legitimate linguistic distinctions. A comprehensive approach therefore requires disaggregated evaluation, participatory input from affected communities, and transparent reporting of known limitations.

5. Deployment, Sustainability, and Operational Challenges

Deploying large language models in production environments involves a complex stack of inference optimization, hardware management, and continuous monitoring. The autoregressive generation process is memory bandwidth bound, which means that throughput and latency depend heavily on the parallelization of attention and the efficiency of key-value cache management. Quantization can reduce memory requirements and energy per token, but it may introduce new error modes that are not captured by standard perplexity benchmarks. Serving systems must also support batching, caching, and routing across multiple model replicas, which creates trade-offs between cost and response time. Organizations that deploy models as internal tools or public APIs must decide whether to host open-source models or rely on external providers. This decision affects data governance, vendor lock-in, auditability,

and the ability to customize behavior for specific domains. Operational robustness requires fallback strategies when the model refuses, loops, or produces harmful content, and these strategies are often more important than the raw capability of the underlying model.

Sustainability has emerged as a key operational constraint. The energy consumed during training and inference can be substantial, and regional electricity mixes determine the associated carbon emissions. Studies of large multilingual models have estimated significant carbon footprints for a single training run, with additional emissions from ongoing inference and experimentation [16]. Efficiency improvements do not automatically reduce overall energy use if demand increases in response, a phenomenon known as the rebound effect. System designers should therefore measure not only peak compute but also cumulative energy across the model lifecycle, including data storage, networking, cooling, and hardware manufacturing. Reporting practices vary widely, making it difficult to compare the sustainability of different deployment choices. Standardized energy and carbon accounting would support more responsible decision making and prevent sustainability claims from becoming greenwashing.

The maintenance phase of a language model system is often undervalued in research but dominant in practice. Models drift as user behavior changes, new terminology emerges, and the external world shifts. Continuous adaptation through periodic fine-tuning can improve relevance but may also destabilize previously aligned behavior. Monitoring must extend beyond aggregate performance to include failure cases, demographic patterns, and user satisfaction. Logging user interactions raises privacy concerns, and data minimization principles may conflict with the desire to collect feedback for improvement. System designers need explicit policies for retention, deletion, and anonymization. Deployment therefore requires collaboration across machine learning engineering, security, legal, and product teams, and the success of a language model system depends on this broader organizational infrastructure rather than on the model checkpoint alone.

6. Governance, Policy, and Accountability

Large language models are increasingly subject to regulatory attention, but existing legal frameworks are fragmented. Data protection laws may restrict the collection and use of personal information in training corpora, while intellectual property regimes are still unsettled with respect to model outputs and memorized content. Sector-specific rules in health, finance, and education add further constraints, and cross-border deployment introduces conflicts among jurisdictions. The governance challenge is not simply to add new rules but to create mechanisms for accountability that are technically meaningful and institutionally enforceable. Transparency reporting can help external stakeholders understand model capabilities, limitations, and known risks. However, transparency without auditability can become a box-ticking exercise. Independent evaluation and red teaming have been proposed as ways to surface failures before they cause harm, but they require access to models, data, and evaluation infrastructure that many organizations are unwilling to share.

Liability is a central governance question. When a language model produces incorrect medical advice, discriminatory hiring recommendations, or defamatory statements, it is unclear who bears responsibility: the model developer, the data provider, the deployment operator, or the user who acted on the output. Contracts often allocate risk to the deployer, but deployers may lack the technical means to verify model behavior. This mismatch between responsibility and capability creates a governance gap. Audit requirements could help by forcing developers to disclose training data provenance, evaluation results, and failure modes.

At the same time, excessive disclosure may create security risks or enable malicious use. Governance thus requires balancing transparency, security, and accountability in a way that is sensitive to the specific context of deployment.

Global governance is further complicated by the concentration of advanced model development in a small number of institutions. This concentration raises concerns about economic power, technological dependency, and the exclusion of underrepresented linguistic communities. Open model releases can broaden access and enable independent research, but they also remove the control points that centralized systems can use to enforce safety policies. Policy responses range from voluntary codes of conduct to mandatory impact assessments and licensing regimes. The effectiveness of these approaches depends on the availability of standardized evaluation methods, the capacity of regulators to interpret technical evidence, and the willingness of industry actors to engage in good faith. Multistakeholder processes that include civil society, academic researchers, and affected communities are essential for building governance arrangements that are legitimate and durable.

7. Future Directions

Future language model systems are likely to become more multimodal, agentic, and deeply integrated into organizational workflows. Multimodal training can align text with images, audio, and video, enabling richer interaction and new forms of accessibility. However, multimodal systems inherit the risks of text-only models while adding new failure modes related to visual perception, sensory grounding, and cross-modal bias. Agentic systems that plan, call external tools, and modify their environment introduce further challenges in reliability and oversight. The ability to use tools can improve factual accuracy and task completion, but it also expands the attack surface and complicates accountability. Research on reasoning has shown that prompting models to produce intermediate steps can improve performance on arithmetic and symbolic tasks [17]. Yet the intermediate steps are not guaranteed to be faithful to the actual computation, and they can create a false sense of transparency. Future work should therefore develop methods for verifying reasoning trails and for distinguishing between plausible post hoc explanations and causally meaningful computations.

Scaling will continue, but the research frontier is shifting from raw parameter count to systemic efficiency, controllability, and evaluative depth. Inverse scaling results have documented cases where larger models perform worse on certain tasks, challenging the assumption that scale uniformly improves capability [18]. This observation supports the view that evaluation must be designed to probe not only aggregate benchmark scores but also the specific conditions under which models fail. Benchmark saturation is a persistent problem, as models quickly achieve high performance on static test sets while remaining brittle on adversarial or distribution-shifted inputs. Dynamic evaluation, human-in-the-loop auditing, and domain-specific task suites are needed to keep pace with model development. Open science and reproducible research will be essential for independent verification, but open access to model weights must be balanced against safety concerns related to dual use.

The most important future direction may be the development of evaluation frameworks that treat language model systems as embedded in social and institutional contexts. Current evaluation often reduces a system to a single model checkpoint tested on decontextualized prompts. Real-world impact depends on the interface, the user population, the data pipeline, the moderation layer, and the broader incentive structures of the deploying organization. Participatory methods can bring affected communities into the evaluation process, surfacing

harms that are invisible to technical benchmarks. The emergence of regulatory requirements for algorithmic impact assessments may accelerate this shift, but technical researchers must develop methods that are rigorous enough to support policy decisions. Sustainability goals should also be integrated into the evaluation of new systems, not treated as afterthoughts. The trajectory of the field will be shaped as much by these governance and evaluation choices as by improvements in pretraining objectives or model architectures.

8. Conclusion

This survey has examined large language models from a system-level perspective, emphasizing that their success depends on the interaction of architectures, data infrastructures, training processes, deployment environments, and governance regimes. The transformer has enabled impressive capabilities, but its limitations are inseparable from the way models are trained, adapted, and used. Architectural choices such as dense or sparse parameterization shape cost, robustness, and auditability. Training data selection determines whose language and knowledge are represented, and post-training alignment introduces value judgments that require explicit institutional oversight. Deployment and sustainability challenges are not peripheral but central to the viability of these systems. Governance must reconcile the concentration of power with the need for public accountability, transparency, and fairness. Future progress will require moving beyond narrow performance benchmarks toward evaluation that captures real-world impact, long-term reliability, environmental cost, and social legitimacy. Researchers, engineers, and policymakers share a responsibility to build language model infrastructures that are not only capable but also trustworthy, inclusive, and sustainable.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171–4186.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

5. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

6. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., &

Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.

7. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

8. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2022). Training compute-optimal large language models. Advances in Neural Information Processing Systems, 35, 30016–30030.

9. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240), 1–113.

10. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

11. Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120), 1–39.

12. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

13. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

14. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

15. Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 1286–1305.

16. Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. Journal of Machine Learning Research, 24(253), 1–15.

17. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

18. Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., Elhage, N., Showk, S. E., Fort, S., Hatfield-Dodds, Z., Henighan, T., Johnston, S., Jones, A., Joseph, N., Kernan, J., Kravec, S., ... Clark, J. (2022). Predictability and surprise in large generative models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1747–1764.

19. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. Proceedings of the 30th USENIX Security Symposium, 2633–2650.

20. Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 298–306.

21. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Isaac, W. (2022). Taxonomy of risks posed by language models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 214–229.

22. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

23. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations.

24. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761.

25. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 3214–3252.