Retrieval-Augmented Prompt Tuning with Dynamic Prompt Injection for Knowledge-Intensive NLP Tasks

Reaid Leawas
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
reid1980@oregonstate.edu

Albeairt Hierrera
Department of Computer Science, University of North Texas, Denton, TX, USA.
albert.herrera@unt.edu

Abstract

The rapid evolution of large language models has spurred interest in parameter-efficient adaptation methods, with prompt tuning standing out as a lightweight alternative to full fine-tuning. Concurrently, retrieval-augmented generation has demonstrated substantial improvements in knowledge-intensive natural language processing tasks by grounding outputs in external non-parametric memory. This paper presents a comprehensive systems-oriented analysis of retrieval-augmented prompt tuning augmented by dynamic prompt injection, a paradigm in which continuous soft prompts are continuously modulated by retrieved evidence at inference time. Drawing on recent advances in dense retrieval, soft prompt optimization, and selective prompt insertion, the work examines the architectural trade-offs, infrastructure requirements, deployment constraints, and socio-technical implications of integrating dynamic retrieval components into prompt-tuned language models. The discussion encompasses the design of retrieval indices, the gating policies that govern injection, the management of latency and resource budgets in production environments, and the challenges of maintaining robustness, fairness, and auditability. We further explore governance frameworks, sustainability considerations, and policy dimensions, including data provenance, copyright, energy consumption, and equitable access to the enabling infrastructure. By synthesizing perspectives from systems engineering, machine learning, and critical infrastructure studies, the paper contributes a holistic framework for evaluating retrieval-augmented prompt tuning as a cornerstone of next-generation knowledge-intensive applications. The analysis highlights that while dynamic prompt injection offers compelling accuracy and adaptability gains, its large-scale deployment demands careful coordination of model serving, retrieval backends, and regulatory compliance mechanisms, and calls for interdisciplinary collaboration to align technical design with societal values.

Keywords

retrieval-augmented generation, prompt tuning, dynamic prompt injection, knowledge-intensive NLP, large language models, parameter-efficient tuning, retrieval systems.

1. Introduction

The landscape of knowledge-intensive natural language processing (NLP) has been reshaped by two parallel developments: the emergence of massive pre-trained language models capable

of storing factual associations within their parameters, and the introduction of retrieval mechanisms that decouple factual grounding from parametric memory. Prompt tuning has gained traction as a parameter-efficient strategy that adapts frozen language models to downstream tasks by learning a small set of continuous prompt embeddings while keeping the core model weights fixed [1]. This approach dramatically reduces the storage and computational overhead associated with fine-tuning, enabling the sharing of a single model backbone across numerous tasks. Yet purely parametric knowledge remains limited by the temporal cutoff of pre-training corpora, the opacity of memorized facts, and the difficulty of updating knowledge without retraining. Retrieval-augmented generation [3] addresses these limitations by equipping language models with access to an external knowledge repository, often a large-scale document index searched via dense passage retrieval [15]. The resulting hybrid architectures exhibit enhanced factuality and the ability to incorporate fresh information, making them particularly suited for open-domain question answering, fact verification, and multi-hop reasoning.

Despite these advances, most retrieval-augmented language model designs either prepend retrieved text to the input sequence before encoding or use specialized cross-attention mechanisms that merge the retrieved content with the model's internal representations. Such integration strategies, while effective, often treat the retrieval step as a static pre-processing operation that does not interact with the prompt tuning process itself. A more flexible paradigm, which this paper terms retrieval-augmented prompt tuning with dynamic prompt injection, forges a tighter coupling: the continuous prompt embeddings that steer the model's behavior are themselves conditioned on the retrieved evidence, enabling the prompt to selectively incorporate, emphasize, or suppress retrieved information depending on the input context. This idea builds on recent work showing that prompt tokens can be inserted selectively rather than uniformly, leading to improved performance in prompt tuning [12]. When extended to retrieval-augmented settings, such selective injection mechanisms can determine not only whether retrieval is needed, but also how the retrieved content should modulate the prompt representation at different positions in the sequence.

This paper adopts a systems-level lens to investigate the architectural, infrastructural, and socio-technical dimensions of retrieval-augmented prompt tuning with dynamic injection. Rather than focusing narrowly on accuracy metrics, we probe the structural trade-offs that arise when retrieval backends are tightly integrated with continuously modulated prompts. These trade-offs span latency budgets, index freshness, resource elasticity, and the governance of external knowledge. As large language models are increasingly deployed in high-stakes domains such as healthcare, law, and public information services, it becomes imperative to assess how these architectures perform under adversarial retrieval settings, how they can be audited for provenance, and what infrastructure policies are required to ensure equitable and sustainable operation. The present analysis contributes a holistic perspective that situates dynamic prompt injection within the broader ecosystem of responsible AI deployment.

2. Background and Related Work

Parameter-efficient tuning methods have evolved rapidly since the observation that few-shot prompting with manually designed templates is effective in large language models [5]. Prompt tuning automates this by learning continuous embeddings that are prepended to the input, with performance scaling favorably as model size increases [1]. Prefix-tuning further demonstrated that inserting learned continuous vectors into all transformer layers can match

the quality of full fine-tuning under certain conditions [2]. Subsequent refinements, such as P-tuning v2, established that deep prompt tuning applied at multiple layers achieves competitive results across diverse scales and task categories [7]. These methods share the premise that a small set of tunable parameters can modulate a frozen model's behavior, thereby enabling efficient multi-task serving on shared infrastructure.

In parallel, retrieval-augmented generation systems have addressed the knowledge staleness and hallucination problems of parametric models by indexing large corpora and retrieving relevant passages on the fly [3]. Architectures such as RAG combine a dense passage retriever with a generative model via latent variable formulations, while kNN-LM augments next-token prediction by interpolating with distributions derived from nearest neighbor search over cached datastores [10]. RETRO scales retrieval to trillions of tokens by using a chunked cross-attention mechanism that attends to retrieved neighbors during generation [9]. Atlas further shows that joint pre-training of the retriever and generator from scratch yields strong few-shot capacities [8]. More recently, Self-RAG has proposed a self-reflective framework where the model learns to retrieve on demand, critique its own generations, and incorporate relevant passages selectively [11]. REPLUG treats the language model as a black box and uses a retriever to search a large corpus, then feeds the retrieved passages as context [13]. These works collectively illustrate the transition from static retrieval to dynamic, context-sensitive retrieval strategies.

The integration of retrieval with prompt tuning remains relatively underexplored. While it is straightforward to concatenate retrieved text with trainable soft prompts, the interaction is typically one-directional: the prompt affects how the model reads the concatenated evidence, but the evidence does not reshape the prompt itself. Dynamic prompt injection addresses this gap by allowing the prompt representation to be recomputed based on retrieved content. The idea resonates with recent investigations into learning to selectively insert prompts for better prompt tuning [12], which demonstrate that not all prompt tokens are equally useful and that insertion decisions can be optimized during training. When combined with retrieval, such selective injection can yield a system in which the prompt becomes an adaptive interface between the external knowledge base and the model's internal representations, a paradigm that we analyze in depth in the following sections.

3. Retrieval-Augmented Prompt Tuning Architecture

At the architectural level, a retrieval-augmented prompt tuning system with dynamic injection comprises four tightly interacting subsystems: a frozen foundational language model, a continuous prompt module, a retrieval backend, and an injection controller that governs how retrieved information modulates the prompt embeddings. The language model remains unchanged throughout serving, which simplifies model versioning and allows the same model checkpoint to serve hundreds of tasks via distinct prompt configurations. The prompt module consists of a small set of learned embedding vectors that are prepended or interleaved with the input sequence. In contrast to static prompt tuning, these embeddings are not constant for every inference request; instead, they are dynamically computed as a function of the retrieved context.

The retrieval backend typically relies on a dense vector index built from a corpus of documents, passages, or structured knowledge triples. During preprocessing, the corpus is encoded into fixed-dimensional vectors using a bi-encoder architecture, and the resulting vectors are stored in an approximate nearest neighbor index to enable sub-linear search [15]. The retriever can be a separate model trained jointly with the prompt parameters or kept

frozen to reduce training complexity; recent approaches such as unsupervised domain adaptation of retrievers and continual index updates have made it feasible to maintain high recall even as knowledge bases evolve. At inference time, the input query is encoded by the same retriever and the top-k most similar passages are fetched along with relevance scores and metadata.

The injection controller translates the retrieved passages into modifications of the continuous prompt. A common design involves a lightweight transformer or gating network that ingests the retrieved text embeddings and outputs a set of additive perturbation vectors for each soft prompt token, adjusting their values based on semantic alignment with the evidence. Alternatively, the controller may decide selectively which prompt positions should receive the retrieved signal, inserting additional prompt tokens derived from the retrieved content only where the model's existing prompt configuration is insufficient. This selective insertion mechanism [12] can be implemented through a learned policy that assigns a binary or scalar weight to each prompt slot, conditioned on the retrieved representations. Training the controller typically involves backpropagating through the frozen language model’s output distribution, using task-specific supervised signals or reinforcement learning objectives that reward faithful use of retrieval.

From a systems perspective, this architecture introduces a dependency chain that must be carefully managed: the retrieval latency adds to the end-to-end response time, the injection controller consumes additional compute, and the dynamic prompt must be recomputed for every query, negating some of the caching advantages of static prompt tuning. However, the architecture also opens opportunities for optimizing retrieval granularity, decoupling the retriever update cycle from the prompt update cycle, and distributing the retrieval index across multiple nodes to balance load.

4. Dynamic Prompt Injection Mechanism

Dynamic prompt injection is the mechanism by which retrieved evidence actively reshapes the representation of the continuous prompt, rather than simply being concatenated as additional context tokens. This mechanism can be decomposed into three conceptual stages: evidence encoding, injection planning, and prompt modulation. Evidence encoding transforms the retrieved passages into a dense representation that can interact with the soft prompt space. In many designs, the same dense encoder used for retrieval also produces the representations that will condition the injection, ensuring representational alignment. The injection planner decides when retrieval is necessary, which retrieved snippets are trustworthy, and how aggressively to modify the prompt. This planner can be a small neural network trained jointly with the prompt parameters, taking as input the encoded evidence and the current task embedding.

Prompt modulation applies the planned adjustments to the continuous prompt vectors. One strategy is additive conditioning, where each soft token vector receives a residual update computed as a learned linear or multi-layer transformation of the evidence encoding. This preserves the base prompt structure while allowing fine-grained adaptation. Another strategy is slot selection, inspired by findings that inserting prompts selectively at certain positions yields better performance than uniform insertion [12]. In the retrieval-augmented setting, the injection controller may predict a binary mask over the prompt sequence, keeping some prompt tokens at their default values while overwriting others with retrieved content-derived embeddings. This mixing of static and dynamic prompt components enables the model to retain its general task-tuning while still benefiting from instance-specific evidence. A further

refinement is hierarchical injection, where different layers of a deep prompt tuning architecture receive different conditioning signals, allowing early layers to access broader topical context and later layers to incorporate fine-grained factual details.

Training such a dynamic injection mechanism requires careful balancing of objectives. The loss function typically combines the standard language modeling or classification loss with auxiliary terms that encourage the injection planner to use retrieval parsimoniously, preventing the model from ignoring the retrieved content or over-relying on it in ways that introduce noise. Regularization techniques and scheduled dropout of retrieved evidence during training have proven effective in calibrating the dependence of the dynamic prompt on retrieval quality. At inference time, the injection planner can be simplified by thresholding or distillation to reduce latency, trading off some adaptivity for speed. The overall effect is a system that behaves like a continuously modulated soft interface between parametric memory and external knowledge, adjusting its inductive bias on a per-query basis while maintaining the parameter efficiency and multi-task serving advantages of prompt tuning.

5. Knowledge-Intensive Task Adaptation and System-Level Considerations

Knowledge-intensive tasks—including fact-checking, biomedical literature mining, legal document summarization, and open-domain question answering—impose stringent requirements on factual accuracy, timeliness, and source attribution. In these domains, retrieval-augmented prompt tuning with dynamic injection moves beyond static augmentation by enabling the model to reconfigure its operational prompt depending on the retrieved evidence's authority and relevance. For instance, when a user queries a rapidly evolving topic such as a recent public health guideline, the retrieval index can be updated daily without modifying the language model weights or the base prompt embeddings, and the injection controller learns to weight recent high-authority sources more heavily. This decoupling of knowledge updates from model retraining is a powerful property that aligns with continuous integration practices in deployed AI systems.

However, this adaptivity introduces system-level complications. The retrieval index must be maintained with attention to versioning and rollback capabilities, as erroneous or poisoned update batches can degrade downstream system behavior and erode user trust. The injection controller’s propensity to trust certain sources can encode biases present in the retrieval corpus, and its operation becomes entangled with the governance of the underlying knowledge base. Moreover, in multi-tenant environments where one model serves multiple organizations, the retrieval index and prompt configuration may need to be logically or physically partitioned to meet data residency requirements and prevent cross-tenant information leakage. The architectural decision to use a shared retrieval backend versus dedicated per-tenant indices has profound implications for cost, latency, and compliance with regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA).

From a deployment standpoint, dynamic prompt injection systems must coexist with model serving infrastructure that already handles batching, auto-scaling, and request routing. The injection controller adds a new critical path in the inference pipeline, requiring careful profiling to ensure that the added latency does not violate service level objectives. System designers may choose to precompute evidence embeddings for popular queries, cache frequently retrieved passages, or offload certain injection decisions to asynchronous processes. Such optimizations inevitably introduce trade-offs between freshness and speed, and must be documented in transparency reports if the system is deployed in regulated contexts.

6. Infrastructure, Deployment, and Scalability

Deploying retrieval-augmented prompt tuning with dynamic injection at scale demands a robust infrastructure that integrates a large-scale dense retrieval service with a language model serving platform. The retrieval service typically consists of an embedding model, a vector database or approximate nearest neighbor engine, and a document store that supplies raw text for evidence encoding. The prompts and injection controller add lightweight stateless components that can be collocated with the language model server or distributed as sidecar containers. In high-throughput settings, the retrieval index may be sharded across dozens of nodes, and query processing pipelines must balance the parallelism of retrieval with the sequential demands of transformer-based generation.

One of the central infrastructure challenges is managing the end-to-end latency budget. While retrieval latency can be brought below 50 milliseconds with optimized vector search engines, propagating retrieved content through the injection controller and recomputing the soft prompt adds overhead that scales with prompt length and model size. Techniques such as layer-wise injection caching, where only a subset of prompt tokens is updated, and early-exit gating help contain this overhead. Horizontal scaling of the retrieval layer is relatively straightforward; however, the dynamic prompt computation is tightly coupled to the per-request processing, making it less amenable to batching across diverse retrieval results. System architects must therefore decide between per-request dynamic prompt computation and approximate prompt clustering, where similar retrieval results are mapped to a small set of cached prompt configurations.

Storage and memory footprints constitute another scalability dimension. The language model and its prompt tables are measured in gigabytes, but the retrieval index for a corpus of several billion passages can span terabytes of memory. Hybrid memory-storage hierarchies, quantization of vector embeddings, and intelligent eviction policies for rarely accessed passages are essential to keep infrastructure costs within operational boundaries. Furthermore, the integration of dynamic injection with federated learning or geo-distributed deployments introduces questions about index consistency, model-prompt synchronization, and the feasibility of incremental index updates without service interruption. These engineering decisions are not purely technical; they influence who can deploy such systems and under what economic models, thereby shaping the landscape of AI access and competition.

7. Robustness, Fairness, and Governance

The inclusion of external retrieval into prompt-tuned language models enlarges the attack surface and amplifies concerns about robustness, fairness, and accountability. Retrieval indices can be manipulated via adversarial insertion of misleading documents, and dynamic injection controllers that lack robust calibration may amplify low-quality sources [21]. The selective prompt insertion mechanism, if trained on biased example corpora, can learn to preferentially activate prompts that reflect gender, racial, or cultural stereotypes present in the retrieval pool, thereby perpetuating systemic inequities. Fairness audits must extend beyond the language model itself to encompass the retrieval corpus, the retriever, and the injection policy, requiring a comprehensive socio-technical evaluation framework.

Governance frameworks for retrieval-augmented systems are still nascent. The ability to attribute generated content to retrieved evidence is a double-edged sword: on one hand, it enables source verification and facilitates external oversight; on the other hand, it exposes the system operator to liability if the retrieved documents contain defamatory or copyrighted

material. In the context of dynamic prompt injection, where the prompt itself is altered by retrieved content, the boundary between model output and retrieved evidence blurs, complicating auditing procedures. Policy interventions may require system developers to maintain verifiable logs of retrieval queries, injection decisions, and the resulting prompt states, as well as to implement kill switches that disable retrieval for certain sensitive query types. The deployment of such logging mechanisms must be balanced against privacy considerations, especially when user queries are themselves sensitive.

International regulatory regimes increasingly treat AI systems that combine generation with external data access as high-risk applications. The European Union AI Act, for instance, classifies certain AI practices as unacceptable or high-risk based on their potential to cause harm, and systems that generate factual claims without clear provenance can fall under strict transparency obligations. Retrieval-augmented prompt tuning with dynamic injection can support compliance by design if the injection controller not only modulates prompts but also emits attribution scores and retrieval confidence indicators. Such capabilities, however, require additional instrumentation and validation that must be integrated into the system's continuous monitoring pipeline.

8. Sustainability and Policy Implications

The environmental footprint of large language models has become a central concern in the AI ethics community [23]. While prompt tuning itself is energy-efficient compared to full fine-tuning, the addition of retrieval backends and dynamic injection controllers introduces supplementary computational costs that must be holistically assessed. The retrieval index's construction and maintenance involve embedding billions of documents and serving queries at scale, which can rival the energy consumption of model training over extended operational periods. At the same time, dynamic prompt injection can reduce the need for larger model sizes by compensating for parametric knowledge gaps with external evidence, potentially shifting the energy profile from training capital expenditure to inference operational expenditure. A comprehensive life-cycle assessment is needed to determine the net sustainability impact of such architectures.

Policy measures aimed at promoting sustainable AI may include energy efficiency standards for retrieval indices, incentives for sharing open-access retrieval corpora to avoid redundant indexing, and the development of carbon-aware scheduling mechanisms for retrieval-augmented serving. Public research infrastructures can play a pivotal role by maintaining verified, high-quality retrieval corpora that are continuously vetted for bias and misinformation, reducing the need for every organization to build its own energy-intensive indexing pipeline. Dynamic prompt injection further enables smaller, domain-adapted models to perform at parity with larger models on knowledge-intensive tasks, democratizing access to high-accuracy NLP and lowering the barrier for resource-constrained institutions. Policymakers should consider investment in shared utility-scale retrieval infrastructure as a public good, analogous to investments in broadband networks or digital libraries.

Intellectual property governance is another pressing concern. When retrieval-augmented systems inject copyrighted text fragments into prompts, the legal status of the resulting generated content and the transformations applied by the injection controller remain ambiguous. Open questions regarding fair use, derivative works, and database rights create uncertainty for commercial deployments and may delay adoption in sectors such as education and journalism. Collaborative frameworks involving rights holders, platform operators, and

regulatory bodies are needed to develop attribution protocols, opt-out mechanisms, and equitable remuneration models that do not stifle innovation while respecting creators' rights.

9. Discussion and Future Perspectives

Retrieval-augmented prompt tuning with dynamic prompt injection represents an emerging point of convergence between parameter-efficient adaptation and non-parametric memory. Looking forward, several research directions promise to deepen this integration. Multimodal retrieval, in which text prompts are modulated by images, tables, or structured database records, will extend the paradigm to richer forms of knowledge-intensive reasoning. Lifelong learning scenarios, where injection controllers adapt to distributional shifts over time without catastrophic forgetting, will require continual learning algorithms that jointly update prompt modules and retrieval policies. Federated retrieval, where evidence resides in siloed data stores across institutions, introduces security and privacy challenges that cryptographic techniques like private information retrieval may help address. The development of standardized benchmarks that measure not only task accuracy but also retrieval quality, prompt stability, latency, and fairness will be critical for guiding progress.

From a socio-technical standpoint, the trajectory of dynamic prompt injection systems will be shaped by regulatory evolution, public trust, and the economics of cloud infrastructure. Open-source implementations of retrieval-augmented prompt tuning toolkits could accelerate community-driven auditing and foster a diverse ecosystem of deployments, mitigating the risk of concentration in a few large technology firms. At the same time, the technical complexity of maintaining retrieval pipelines and ensuring responsible deployment may reinforce existing power asymmetries unless accompanied by capacity-building efforts and inclusive governance structures. The academic community, together with civil society and industry, must engage in ongoing dialogue to align the design of these systems with societal values and long-term human welfare.

10. Conclusion

This paper has presented a system-level analysis of retrieval-augmented prompt tuning augmented by dynamic prompt injection, a paradigm that fuses the efficiency of soft prompt adaptation with the factual grounding of external retrieval. By examining the architectural components, injection mechanisms, infrastructure scalability, robustness, fairness, governance, and sustainability dimensions, we have highlighted both the transformative potential and the multifaceted challenges of this approach. The tight coupling between retrieval and prompt modulation enables unprecedented adaptability for knowledge-intensive NLP tasks, but it also imposes demanding requirements on retrieval infrastructure, injection policy training, and accountable deployment practices. As the field moves toward deploying such hybrid systems in real-world settings, interdisciplinary collaboration among machine learning researchers, systems engineers, legal scholars, and policy experts will be essential to ensure that the benefits are broadly shared and that the risks are systematically managed. The framework developed here is intended to serve as a foundation for principled design and evaluation, guiding future efforts to build retrieval-augmented language systems that are not only accurate and efficient but also trustworthy, sustainable, and aligned with the public good.

References

1. Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3045–3059). Association for Computational Linguistics.

2. Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (pp. 4582–4597). Association for Computational Linguistics.

3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474). Curran Associates, Inc.

4. Gu, K., Dong, Y., Lee, K., & Liang, P. (2022). Training language models to retrieve and reason with world knowledge. In Advances in Neural Information Processing Systems (Vol. 35, pp. 27690–27706). Curran Associates, Inc.

5. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

6. 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.

7. Liu, X., Ji, K., Fu, Y., Tam, W. L., Du, Z., Yang, Z., & Tang, J. (2022). P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 61–79). Association for Computational Linguistics.

8. Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 24(251), 1–55.

9. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., … Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning (pp. 2206–2240). PMLR.

10. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2020). Generalization through memorization: Nearest neighbor language models. In International Conference on Learning Representations. OpenReview.net.

11. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations. OpenReview.net.

12. Zhu, W., & Tan, M. (2023, December). SPT: Learning to selectively insert prompts for better prompt tuning. In Proceedings of the 2023 conference on empirical methods in natural language processing (pp. 11862-11878).

13. Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., & Yih, W.-T. (2023). REPLUG: Retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652.

14. Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 1870–1879). Association for Computational Linguistics.

15. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769–6781). Association for Computational Linguistics.

16. Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P. N., Ahmed, J., & Overwijk, A. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In International Conference on Learning Representations. OpenReview.net.

17. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 2463–2473). Association for Computational Linguistics.

18. Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How can we know what language models know? Transactions of the Association for Computational Linguistics, 8, 423–438.

19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., 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. In Advances in Neural Information Processing Systems (Vol. 35, pp. 27730–27744). Curran Associates, Inc.

20. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2021). WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.

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

22. 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. S., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

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

24. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

25. 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., … Gabriel, I. (2022). Taxonomy of risks posed by language models. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229). Association for Computing Machinery.