## Language Models as Generative Search Engines

In today's digital landscape, the sheer volume of information dispersed across the Internet makes finding useful answers increasingly tedious. Traditional search engines often return hundreds of long and noisy web pages, requiring users to sift through them manually. In contrast, large language models (LMs) can interpret and generate free-form text, demonstrating promise as personalized interfaces for information seeking. This emerging paradigm of *generative search engines*[1] holds immense potential to revolutionize how we access information; however, deploying them at scale requires addressing significant shortcomings in their utility, reliability, and efficiency. **I develop principled techniques and evaluations that both transform language models into next-generation information-seeking tools and strengthen their general-purpose utility.**

- **Enabling language models for effective information seeking.** Compared to traditional search engines, the next-generation information-seeking paradigm will be equipped with enhanced semantic search for accurately locating relevant sources, which will then be synthesized into succinct, informative answers by powerful language models. I have developed key components for this new paradigm, starting with SimCSE [1], which introduced a simple objective to turn any pre-trained language model into powerful text embeddings. It has since served as the cornerstone for state-of-the-art embedding models, enabling crucial applications of semantic search and beyond. I then proposed ALCE [2], the first framework to automatically evaluate generative search engines, especially on the ability of language models to accurately synthesize and cite sources retrieved from the web. A significant challenge in this paradigm is reliably reasoning over long and diverse contexts of thousands of sources. To address this, I designed thoughtful evaluations [3] and training methods [4] that culminated in the development of ProLong [5]—a long-context model that surpasses the industry state of the art while using much less computational resource.
- **Innovating training and adaptation methods for language models.** Improving the efficiency and quality of language models will be key to realizing the above vision, and I have worked on simple yet enduring techniques to achieve this goal. Back in 2020, my work in LM-BFF [6] demonstrated the importance of using a carefully tuned prompt to drastically reducing the amount of task-specific data needed. The practice of prompt engineering is still widely useful today. Moreover, I developed MeZO [7], a zeroth-order optimizer that substantially reduces the memory overhead for fine-tuning with little performance degradation, enabling applications like on-device customization. I also co-developed a useful pipeline to produce small yet capable models from existing large ones, yielding the popular Sheared-Llama [8] models and inspiring similar efforts across industry. Additionally, I have introduced novel pre-training strategies, such as accelerated masked language model training [9] and conditional pre-training [10], the former having influenced several industry-standard practices.

## Enabling Language Models for Effective Information Seeking

Advanced AI models have introduced an exciting shift in how humans gather, sift through, and aggregate information. This new information-seeking paradigm, termed "generative search engines", requires designing and improving several key components: enhanced semantic search, reliable long-context processing, and the synthesis and attribution of relevant sources. I have made significant contributions to the foundational techniques and evaluations that support each of these components.

**Adapting language models to semantic text embeddings.** Text embedding models, which encode text sequences into continuous vectors, can overcome the limitations of traditional keyword-based representations of text and unlock many applications, such as semantic search, clustering, and recommendation systems. In **SimCSE** [1], I designed a simple yet powerful contrastive learning algorithm that transforms any pre-trained language model into a high-performance text embedding function by pushing similar sentences together while pulling apart different ones. This approach only required unsupervised data but matched the performance of leading supervised methods which used intensive human annotation. Adding such supervision boosts the efficacy of SimCSE even further, causing it to outperform OpenAI's embedding API despite using a 500× smaller model [14].

[1] Commercial products like Perplexity and Google Search's AI overview demonstrate this concept.

The impact of SimCSE has been substantial, with over 3,200 citations and 22 million model downloads; its methodology now serves as the foundation for many commercial embedding products from companies like OpenAI, Microsoft, Google, and Nvidia [15, 16, 17]. I am particularly excited about leveraging these embedding models to aid scientific research. Recently, we introduced **LitSearch** [11], the first literature retrieval benchmark built on manually verified paper-seeking questions. Our findings reveal that embedding models significantly outperform traditional keyword-based retrieval, showing their remarkable potential to accelerate scientific discovery.

**Benchmarking verifiability as source attribution.** Modern language models excel at providing fluent natural language answers; yet to transform them into trustworthy tools for information seeking, *verifiability* is the essential trait—every part of the model's response should be backed by a cited source that users can examine. The commercial success of generative search engines like Perplexity reflects growing demand for AI-powered information-seeking tools, but a lack of evaluations besides labor-intensive human verification makes it hard to confidently scale these systems.

To bridge this gap, I built **ALCE** [2] in the spring of 2023, the first automatic benchmark for generations with citations. We developed holistic model-based evaluation to enable an accurate characterization of model performance, covering fluency, factual correctness, and citation quality. We benchmarked a series of frontier LMs, and found that even GPT-4, the strongest language model at the time, lacked complete citation support 50% of the time. ALCE has set a solid foundation for attribution evaluation in information seeking, inspiring numerous subsequent studies [18].

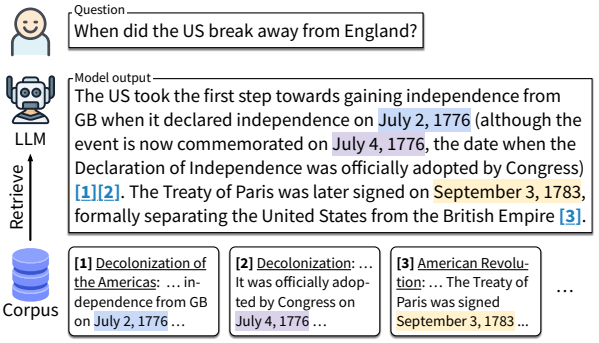


Figure 1: ALCE requires systems to search for relevant sources and synthesize them into coherent answers with citations to references.

ALCE provided useful insights for model development: we observed that most models degrade quickly with more retrieved documents, and those with more sophisticated instruction tuning perform better. This underscores the need for reliable long-context processing and instruction following of language models to build effective generative search engines.

**Enabling language models to process long and diverse contexts.** Building on these insights, I shifted my focus to the broader challenge of making LMs better at accessing information buried in long and diverse contexts. Long-context LMs have the potential to unlock a myriad of new applications besides generative search engines, including long document summarization and software engineering agents. But many challenges lie ahead, including the difficulty of long-context evaluation, the high costs of processing long contexts, and the lack of high-quality training data.

We began by addressing the absence of a reliable long-context evaluation and developed **HELMET** [3], a comprehensive benchmark that covers diverse long-context applications, including ALCE, with robust model-based metrics. We benchmarked 51 long-context LMs and found that HELMET provided more consistent and meaningful signals compared to previous benchmarks.

To address the efficiency challenge, I co-developed **CEPE** [4], a long-context adaptation method by employing a small encoder to process long inputs chunk by chunk. While highly efficient, CEPE and many other open-source models experience significant performance drops on challenging long-context tasks from HELMET. The insights led to the development of **ProLong** [5], a comprehensive recipe for continually training and instruction-tuning short-context LMs to effectively handle long inputs. We conducted thoughtful experiments to find the best practice in each component, and some of our results challenged widespread beliefs in the open-source community: (1) we found that combining code repositories and long books with high-quality shorter data is essential to improving both long and short-context performance; (2) training on documents longer than the intended deployment length yields better performance compared to training only at the deployment length; (3) contrary to previous findings, simply using diverse short instruction data yields better long-context performance than using synthetic long instruction data. Our final model achieves the state-

of-the-art performance and context lengths among 10B-scale open-source models. Notably, ProLong surpasses even Meta's long-context model, Llama-3.1-8B, with only 5% of its long-context data budget, thanks to our comprehensive recipe study. Since their release two months ago, ProLong models and data have gained remarkable popularity, with over 130,000 downloads.

**Evaluating instruction following.** The free-form nature of LM-powered information seeking demands that models strictly obey human instructions. Yet instruction following was largely overlooked in model development, which more often focused on "vibe checking" [19]. In the fall of 2023, we built **LLMBar** [12] to benchmark instruction following by using an adversarially constructed and manually verified pairwise preference dataset. Since then, instruction following has become a crucial focus in language model development and is centered in public benchmarks in 2024. Due to its high quality and challenging nature, LLMBar is adopted in RewardBench [20], the canonical evaluation for reward models in reinforcement learning from human feedback (RLHF) pipelines [21].

## Innovating Training and Adaptation Methods for Language Models

Language models have enabled numerous new applications such as generative search engines. Their wide deployment calls for both efficient training methods and compact models that maintain versatility and capability. I have developed principled techniques for this purpose that are proven to be useful in real-world applications and have been widely adopted by both academia and industry.

**Fine-tuning language models with minimal data and compute.** In 2020, the predominant paradigm for using language models is to fine-tune them on massive amounts of task-specific data. It is costly to collect such data but the models struggle if there are not at least thousands of training examples. In **LM-BFF** [6], I designed a novel prompting paradigm—reformulating any task as a next word prediction problem—and pioneered the idea of automatically searching prompts. This enabled us to fine-tune small language models with as few as 32 examples, achieving accuracies that would require at least thousands of examples by the previous approach to match. Our work is one of the earliest in *prompt engineering* and has served as a catalyst for numerous subsequent research in this area, including influential work such as FLAN [22].

Another challenge that fine-tuning faces is the ever-growing model sizes and the associated high memory cost. For example, using a $30K H100 GPU only allows one to fine-tune a 3B parameter model with a standard Adam optimizer. We developed **MeZO** [7], a zeroth-order optimizer with clever in-place operations that only requires two forward passes and same memory footprint as inference. We demonstrated that MeZO can achieve performance competitive with standard backpropagation (up to 12× memory reduction) and argued for the importance of prompt-based fine-tuning (previous section) with theoretical justifications. As companies increasingly pursue on-device LM customization, MeZO offers a practical solution for memory efficient fine-tuning.

**Producing capable small models efficiently via structured pruning.** There is a significant demand for capable smaller models, but training them from scratch is still resource-intensive. To address this, we developed **Sheared-Llama** [8], a method to derive strong small models from large pre-trained ones by targeted structured pruning and continued training. We created 1B and 3B models that outperform (then) state-of-the-art models of similar sizes, with only 3% of their training budget. Sheared-Llama has gained tremendous popularity in the open-source community with over 700,000 downloads, and offered an efficient solution for producing capable small models—adopted by both Nvidia's Minitron [23] and Meta's Llama-3.2 [24].

**Accelerating language model pre-training via objective innovations.** Before 2023, the prevailing paradigm for pre-training was BERT-like [25] masked language modeling (MLM). MLM masks out and predicts a portion of text; a fixed masking ratio of 15% is always applied. However, our research [9] demonstrated that it is important to tailor masking rates to specific settings: we revealed that masking 40% could achieve comparable performance on certain tasks with only half the compute compared to using 15%; even an 80% masking rate yielded non-trivial results. As the first paper studying masking rates, our work has since influenced industry practices, with Google's UL2 adopting 50% masking [26] and MosaicML's BERT training recipe using 30% masking [27].

In [10], we introduced a novel pre-training paradigm called "**conditional pre-training**" to address the challenge of diverse domains and varying qualities in data. It simply prepends a "condition" to the document, such as URLs (e.g., `wikipedia.org`); learning the relationship between the condition and the document will help models better interpret noisy pre-training distributions and improve data efficiency. Our experiments showed that combining conditional pre-training with a simple data annealing strategy can achieve comparable performance to standard language modeling with 33% less data, demonstrating tremendous promise in accelerating pre-training.

## Future Work

Building on my existing research in advancing how LMs source information, my future work will focus on enabling them to execute tasks and act on information—facilitating new exciting applications from personal assistants to autonomous research agents. This evolution will be driven by (1) continuous advancements in fundamental capabilities of LMs, (2) expanded means to interpret rich contexts, such as computer UIs, and (3) improved reliabilities in generating extended reasoning traces and action sequences. Here is how I plan to address these challenges.

**Better scaling via architecture innovations.** The ever-scaling data and compute used in training language models have led to emerging capabilities and applications; yet the industry is witnessing a deceleration of AI development due to exhausted resources and depleted data [28]. I believe that the next wave of AI advancement lies in more efficient scaling via architecture innovations beyond the predominant Transformer paradigm [29]. In an ongoing work, I am exploring a simple method to transform a full-attention long-context LM into a hybrid of full and local-attention, leading to 70%-90% memory saving while maintaining the performance. I will also argue that the focus of architecture research should not only be on efficient replacement of Transformers, but also on combining them with more expressive components: for example, research shows that non-linear recurrent architectures are more proficient at reasoning tasks [30] while full attention is better at in-context recall [31]. A hybrid architecture that combines all will strike a balance between efficiency and expressivity and hold the promise to a better scaling law. While most industry efforts remain focused on scaling Transformers, academia and open science can significantly impact future language models through architecture innovations.

**Screenshots as the unified input interface.** A vast amount of human knowledge is encoded in modalities beyond text, such as images and videos. However, even the state-of-the-art multimodal models often cannot use the visual information effectively [32, 33]. I believe the fundamental limitation of existing models lies in the separation of language and vision. In [13], we explored a new paradigm termed *screenshot language models*, casting both images and text in a single "screenshot" input. We show that by using a carefully designed objective, screenshot LMs can reach a similar language understanding ability as a text-only LM while only reading text from images. The appeal of this paradigm lies in its ability to natively and seamlessly integrate text and image understanding, promising exciting new applications like multimodal document understanding and computer use [34]. The distinct nature of screenshot LMs compared to traditional multimodal LMs presents unique challenges in scaling, and I plan to continue exploring better objectives, data, and architectures to improve screenshot LMs.

**Training language models that can reliably plan, reason, and act.** Language models that can not only understand the digital world but also plan, reason, and act within it represent a crucial step towards general-purpose intelligence. These capabilities will manifest through *long-form generations*, which present a significant challenge as current models are not yet fully equipped to handle them reliably. I believe that the key to overcoming this challenge lies in *weak-to-strong synthetic data*: new training data that are easy to generate following a procedure but are hard to learn by models—for example, we can construct complex reasoning traces following a search algorithm, and then use them to train the language models. I plan to explore advanced methods for constructing such data, combined with innovative objectives, to boost the capabilities of language models to reliably and faithfully execute long-horizon tasks.

## Publications/Preprints by the Applicant

[1] **Tianyu Gao***, Xingcheng Yao*, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910, 2021.

[2] **Tianyu Gao**, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *EMNLP*, pages 6465–6488, 2023.

[3] Howard Yen, **Tianyu Gao**, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izasak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

[4] Howard Yen, **Tianyu Gao**, and Danqi Chen. Long-context language modeling with parallel context encoding. In *ACL*, pages 2588–2610, 2024.

[5] **Tianyu Gao***, Alexander Wettig*, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*, 2024.

[6] **Tianyu Gao***, Adam Fisch*, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL-IJCNLP*, pages 3816–3830, 2021.

[7] Sadhika Malladi*, **Tianyu Gao***, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *NeurIPS*, 2023.

[8] Mengzhou Xia, **Tianyu Gao**, Zhiyuan Zeng, and Danqi Chen. Sheared LLaMA: Accelerating language model pre-training via structured pruning. In *ICLR*, 2024.

[9] Alexander Wettig*, **Tianyu Gao***, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? In *EACL*, pages 2985–3000, 2023.

[10] **Tianyu Gao**, Alexander Wettig, Luxi He, Yihe Dong, and Danqi Chen. Accelerating language model scaling via simple conditional pre-training. 2024.

[11] Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and **Tianyu Gao**. Litsearch: A retrieval benchmark for scientific literature search. In *EMNLP*, 2024.

[12] Zhiyuan Zeng, Jiatong Yu, **Tianyu Gao**, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *ICLR*, 2024.

[13] **Tianyu Gao**, Zirui Wang, Adithya Bhaskar, and Danqi Chen. Improving language understanding from screenshots. *arXiv preprint arXiv:2402.14073*, 2024.

## Publications/Preprints by Others

[14] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

[15] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[16] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

[17] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024.

[18] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024.

[19] Arnav Gudibande, Eric Wallace, Charlie Victor Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In *ICLR*, 2024.

[20] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

[21] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[22] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *ICLR*, 2021.

[23] Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *arXiv preprint arXiv:2407.14679*, 2024.

[24] Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, September 2024.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[26] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. Ul2: Unifying language learning paradigms. In *ICLR*, 2023.

[27] Jacob Portes, Alex Trott, Daniel King, Sam Havens, and Erica Ji Yuen. Mosaicbert: Pretraining bert from scratch for $20, March 2023.

[28] Stephanie Palazzolo, Erin Woo, and Amir Efrati. Openai shifts strategy as rate of 'gpt' ai improvements slows, 2024.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017.

[30] William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *ICML*, 2024.

[31] Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and eran malach. Repeat after me: Transformers are better than state space models at copying. In *ICML*, 2024.

[32] Zhangyang Qi, Ye Fang, Mengchen Zhang, Zeyi Sun, Tong Wu, Ziwei Liu, Dahua Lin, Jiaqi Wang, and Hengshuang Zhao. Gemini vs gpt-4v: A preliminary comparison and combination of vision-language models through qualitative cases. *arXiv preprint arXiv:2312.15011*, 2023.

[33] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024.

[34] Anthropic. Developing a computer use model, 2024.