Proceedings of the 3rd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2025.DL29003

From GPT to LLaMA: Tracing the Growth of Large Language Models

Jiarui Gu

Department of Mathmatical and Computer Science, University of Toronto Mississauga, Mississauga, Canada
niyou9ma@gmail.com

Abstract. Large Language Models (LLMs) have transformed natural language processing by scaling model parameters to unprecedented levels. This review traces the historical progression of LLM parameter sizes, from early pre-trained models with millions of parameters to today's multi-billion and even trillion-parameter systems. We examine key breakthroughs in scaling (e.g., the GPT series, PaLM, LLaMA), highlighting how increasing model size has led to emergent capabilities in language understanding and generation. We also discuss the engineering innovations, such as Transformer architectures and mixture-of-experts, that enabled these leaps in scale. A comparative analysis is provided, including a table and trend figure, to illustrate growth in parameter counts over time and across model families. We further explore the implications of model size on performance, emergent behaviors, and computational cost, noting scaling laws and diminishing returns. Finally, we discuss future directions, arguing that while scaling has driven progress, challenges in efficiency, alignment, and data quality will shape the next phase of LLM development.

Keywords: Large language models, Parameter scaling, Scaling laws, Emergent abilities, Model performance

1. Introduction

Large Language Models (LLMs) are a class of neural language models characterized by extremely large numbers of parameters (typically on the order of billions) trained on massive text corpora. The recent prominence of LLMs can be traced to the success of Transformer-based architectures and the demonstration that scaling model size leads to improved capabilities. Since the release of OpenAI's ChatGPT in late 2022, LLMs have gained widespread attention for their general-purpose language understanding and generation abilities. These abilities arise in part from the sheer scale of the models' parameters and training data, which allow LLMs to internalize vast amounts of linguistic and world knowledge. Indeed, theoretical and empirical scaling laws have predicted that larger models trained on more data achieve lower error rates, providing a recipe for progress in NLP through increasing model size.

However, the quest for ever-larger LLMs also introduces significant challenges. Training and deploying multi-billion-parameter models demand enormous computational resources and careful engineering. Furthermore, while LLMs exhibit impressive performance, their behavior (such as the

DOI: 10.54254/2753-8818/2025.DL29003

emergence of new capabilities) is not always a smooth or linear function of model scale. As a result, understanding the evolution of LLM parameter sizes – and the consequences of that evolution – is crucial for researchers and practitioners. This paper provides a technical review of how LLM parameter counts have grown historically, the breakthroughs and scaling techniques that enabled this growth, and the implications for model performance and cost. We focus on prominent LLM families (GPT, PaLM, LLaMA, etc.) and key milestones in scaling, anchoring our discussion in data drawn from recent survey literature. The remainder of this paper is organized as follows: Section 2 presents a historical overview of LLM developments with increasing model sizes. Section 3 discusses parameter scaling trends and architectural advancements that facilitated extreme-scale models. Section 4 provides a comparative analysis of parameter growth, including a summary table of major LLMs and a figure illustrating the trend of model sizes over time. Section 5 examines the impact of model size on performance and computational cost. Section 6 outlines discussion points and future directions, and Section 7 concludes the paper.

2. Historical overview of LLMs

Beyond the traditional paradigm of scaling up pre-trained models, researchers have begun to explore the use of LLMs as the core components of multi-agent systems. Multi-agent LLMs let multiple models work together. They divide tasks to solve complex problems, like software development, social simulations, and policy modeling. This shows that LLM history is not only about bigger parameters, but also about the new application paradigms [1].

In recent years, large language models have shown exponential growth in size [2]. Early neural language models in the 2010s, such as recurrent networks and the first Transformer-based models, had only tens or hundreds of millions of parameters. The introduction of the Transformer in 2017 marked a turning point, which enable greater scalability. In 2018, OpenAI release GPT-1 (Generative Pre-Training), a 110M–parameter Transformer decoder, demonstrating the power of unsupervised pre-training. At the same year, Google introduced BERT, which is an encoder-based Transformer with up to 340M parameters, set new benchmarks through bidirectional pre-training.These systems established the paradigm of large pre-trained language models (PLMs). However, their parameter counts remained below the billion scale.

A dramatic scale-up occurred with OpenAI's GPT-2 in 2019, which increased the parameter count to 1.5 billion. GPT-2 showed that a sufficiently large model trained on a large web corpus could generate fluent, coherent text and perform rudimentary zero-shot tasks. The following year saw the debut of GPT-3 (2020) with 175 billion parameters. GPT-3 was a watershed moment: it is widely considered the first true "large" LM, as its unprecedented size enabled qualitatively new capabilities such as robust few-shot in-context learning (the ability to perform tasks with only prompt examples instead of fine-tuning). This emergent ability was not observed in smaller predecessors, underscoring how scaling led to a breakthrough in model behavior. GPT-3's release demonstrated that very large models can achieve strong performance across a wide range of NLP tasks without task-specific training.

The race toward larger models continued after GPT-3. By late 2021, several organizations had trained models surpassing GPT-3 in scale. Microsoft/NVIDIA's Megatron-Turing NLG (MT-NLG) achieved 530 billion parameters in 2021, and Google's Switch Transformer/GShard initiative introduced a sparse mixture-of-experts model with effectively 1.2 trillion parameters (the GLaM model). Notably, GLaM's mixture-of-experts design meant only a subset of its 1.2T parameters are active for any given token, reducing computation compared to a dense model of equal size. By early 2022, Google's PaLM (Pathways Language Model) pushed the frontier for dense Transformers with

DOI: 10.54254/2753-8818/2025.DL29003

540 billion parameters, trained on 780 billion tokens of data. Around the same time, DeepMind unveiled Gopher (280B) and showed extensive evaluations, and later introduced Chinchilla – a 70B model that, despite having fewer parameters, was trained on 4× more data (1.4 trillion tokens) to outshine Gopher's performance by following an optimal scaling law. This highlighted that simply increasing parameters without increasing data was suboptimal, and that compute-optimal training might favor smaller models given a fixed compute budget.

By mid-2022, the open-source community also entered the arena: the BigScience project released BLOOM, a 176 billion–parameter multilingual model, as a fully open-access LLM. BLOOM's development underscored the collaborative effort to make large models more accessible. In 2023, we saw further refinement rather than huge increases in parameter count for dense LLMs. Meta AI's LLaMA (Feb 2023) provided models ranging from 7B to 65B parameters, trained on high-quality data, which achieved performance competitive with larger models. Notably, LLaMA-65B (65 billion) trained on 1.4T tokens matched or surpassed earlier 175B models in many benchmarks, showing the benefits of data quality and training efficiency. Google's PaLM 2 (May 2023) slightly reduced size to 340B parameters but trained on a massive 3.6 trillion token corpus, yielding strong multilingual and reasoning performance. Finally, OpenAI's GPT-4 (2023) arrived as the successor to GPT-3. While details are proprietary, GPT-4's model size has been reported to be on the order of trillions of parameters (approximately 1.7 trillion). This would make GPT-4 the largest publicly known dense LLM. In summary, from 2018 to 2023 the field moved from hundreds of millions of parameters to over a trillion, an increase by four orders of magnitude. Table 1 and Figure 1 illustrate this progression through select major models.

Table 1. The progression of major LLM models

Model	Release	Parameters	Training Data (tokens)	Notes
GPT-3	2020	175 billion	300 billion	First to demonstrate emergent few-shot learning.
Gopher (DeepMind)	2021	280 billion	300 billion	Extensive evaluation on 152 tasks.
Chinchilla (DeepMind)	2022	70 billion	1.4 trillion	Compute-optimal model; outperforms Gopher with fewer params.
PaLM (Google)	2022	540 billion	780 billion	Dense decoder-only Transformer (English and multilingual variants).
BLOOM (BigScience)	2022	176 billion	366 billion	Open-access multilingual model (collaborative project).
LLaMA 1 (Meta)	2023	65 billion	1.4 trillion	Released as foundation model for research (Apache 2.0 license).
PaLM 2 (Google)	2023	340 billion	3.6 trillion	Improved training efficiency and multilingual capability.
GPT-4 (OpenAI)	2023	1.76 trillion	13 trillion	Estimated; exact details closed. Multi-modal (image & text) capabilities.

DOI: 10.54254/2753-8818/2025.DL29003

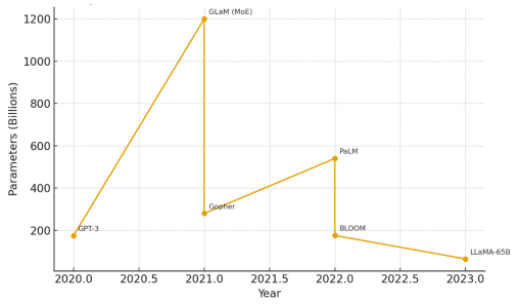

Figure 1. Growth of LLM parameter counts over time (highlighting record-high model sizes)

As shown in Figure 1, The curve illustrates the rapid increase in scale from ~0.1 billion (in 2018) to ~1700 billion (in 2023). Notably, the introduction of sparse mixture-of-experts (GLaM in 2021) enabled trillion-scale parameters (marked with), and the trend resumed with dense models by 2023. Data source: References [3,4].

In contrast to text-only scaling, multimodal large language models (MLLMs) have recently emerged, with examples such as GPT-4V capable of processing both text and images. These models exhibit new emergent abilities through multimodal instruction tuning and cross-modal representation learning, such as image description, OCR-free mathematical reasoning, and visual programming assistance. Multimodal extensions thus represent a parallel trajectory of progress, rather than a mere increase in parameter counts [5].

3. Parameter scaling and architectural advancements

The rapid escalation of model parameters has been driven both by empirical discoveries about scaling and by innovations in model architecture and training strategies. In this section, we examine how researchers have managed to scale LLMs to such extents, and what challenges and phenomena have emerged along the way.

3.1. Scaling laws and emergent behavior

A key driver of LLM growth is the principle of scaling laws. These laws describe how performance improves with larger models, more data, and higher compute. Kaplan et al. (2020) showed power-law error reductions as parameters and data increase. This suggested that bigger models perform better if data and compute grow together. Researchers adopted this view and scaled models by orders of magnitude. Many breakthroughs, as shown in Table 1, emerged from entering new scale regimes. For example, GPT-2 had 1.5B parameters, while GPT-3 expanded to 175B. This jump produced not only quantitative gains but also qualitative abilities. Such emergent behaviors were absent in smaller models and define the impact of scaling. Emergence is often unpredictable: certain capabilities seem to surface only after crossing a parameter threshold. As one survey notes, some performance gains are “predictable” with smooth scaling, while other gains appear as sharp discontinuities, signaling emergence of new skills like arithmetic or commonsense reasoning at a certain model size. GPT-3’s in-context learning is one example; more recently, models approaching a trillion parameters are reported to attain strong multi-step reasoning and knowledge integration that smaller models struggle with.

However, scaling laws also revealed diminishing returns in some regimes. Beyond a point, doubling parameters might yield less than a proportional gain in performance on certain benchmarks. An OpenAI study found that while larger models continue to improve representation

DOI: 10.54254/2753-8818/2025.DL29003

quality, the improvements can slow as size grows, especially if not matched with sufficient data. This revealed that, under a fixed compute budget, model size and training duration must be balanced. DeepMind's Chinchilla experiment explained this point. Chinchilla (70B) outperformed Gopher (280B) by training a smaller model on more data with equal compute. The result challenged the idea that increasing parameters alone ensures progress. It showed that scaling data and optimizing training are equally critical.

3.2. Architectural innovations enabling scale

Building LLMs with tens of billions of parameters required overcoming challenges in training efficiency and memory usage. Several key advancements have allowed researchers to push model sizes upward:

1. Transformer Efficiency: The Transformer architecture supports parallelization far better than recurrent networks. This property allowed models such as BERT and GPT to scale to hundreds of millions of parameters. By removing recurrent dependencies, Transformers enabled efficient training on distributed hardware. Subsequent optimizations further expanded feasible model size. Examples include efficient attention mechanisms and improved software frameworks such as DeepSpeed and Megatron-LM.

2. Model and Data Parallelism: Training extremely large models requires parallelism across many GPUs or TPUs. Model parallelism distributes different parts of the model across devices. Data parallelism divides training batches among processors. Pipeline parallelism and optimized communication strategies were developed to keep hundreds or thousands of accelerators effectively utilized. For instance, GPT-3's training reportedly used a cluster of 1024 GPUs. Google's PaLM was trained across 6144 TPU v4 chips, and other large models similarly required massive compute clusters.

3. Mixture-of-Experts (MoE): One of the most intriguing architectural advances for scaling is the Mixture-of-Experts approach. In an MoE Transformer, multiple parallel subnetworks ("experts") are trained, and a gating mechanism activates only a few experts per input. This sparsely activated model dramatically reduces the computation needed for a given model capacity. GLaM (Generalist Language Model) used an MoE architecture to reach 1.2 trillion parameters effectively, but with much lower training cost than an equally large dense model. As reported, GLaM achieved about 7× the size of GPT-3 while consuming only 1/3 of the energy for training and half the inference FLOPs. MoE models thus proved that ultra-high parameter counts were possible without a proportional increase in computational expense, by trading off utilization (not all parameters are used for each input). This approach has been adopted in various forms (Switch Transformers, BASE layers), although it introduces complexity in training dynamics.

4. Training Stabilization and Regularization: Training at extreme scale required solving optimization instabilities. Techniques such as precision reduction with FP16/BF16, gradient clipping, adjusted learning rate schedules, and improved initialization were crucial to stabilize learning. In addition, regularization strategies including dropout and weight decay, together with advanced optimizers such as AdamW and Adafactor, were carefully tuned for massive models. These incremental advancements collectively enabled the successful training of models with more than 100B parameters without divergence.

5. Infrastructure and Hardware: Lastly, the evolution of hardware, such as the transition from V100 GPUs to A100s and from TPU v3 to TPU v4, together with the development of specialized software frameworks for large-model training, proved instrumental. Distributed training libraries

DOI: 10.54254/2753-8818/2025.DL29003

and high-memory GPUs provided the foundation for scaling, making it possible to move from millions to trillions of parameters within only a few years.

In summary, the growth of LLMs was not simply adding more layers. Scaling required strict use of scaling laws to guide what to expand. It also needed innovations such as mixture-of-experts (MoE) for efficiency. Engineering solutions were essential to train on available hardware. These combined efforts shaped today's LLMs, which stand as achievements of both model design and systems engineering.

4. Comparative analysis of parameter growth

To illustrate LLM growth, we provide a comparative analysis of representative models. Table 1 in Section 2 lists major LLMs, their parameter counts, and training data scales. Several comparisons and trends stand out, which we discuss below:

1. Scaling within a Family: Within the GPT family, parameter counts rose from 1.5B in GPT-2 to 175B in GPT-3 and to an estimated 1.7T in GPT-4. Each leap added new capabilities. GPT-3 outperformed GPT-2 in zero-shot tasks, while GPT-4 is reported to show stronger reasoning and multi-modal abilities with image inputs. Similarly, Google's PaLM family progressed from 540B (PaLM) to a more refined 340B (PaLM 2) with better training efficiency and data, aiming for quality over sheer size.

2. Different Approaches to Scaling: Not all organizations prioritized the same strategy. OpenAI and Google initially focused on dense scaling (GPT-3, PaLM), whereas others like DeepMind and Google Brain explored sparse scaling via MoE (e.g., GLaM 1.2T, Switch Transformer) to push parameter counts higher. DeepMind's Chinchilla study took yet another approach: it scaled data rather than parameters to demonstrate an optimal trade-off. These comparisons indicate that "largest model" is not the only metric – how the model is scaled (and with how much data) is equally important. For example, Chinchilla (70B) outperformed Gopher (280B) by having 4× the training tokens despite being 1/4 the size.

3. Open-Source vs Proprietary Models: An interesting contrast is between open models like BLOOM (176B) or LLaMA (65B) and proprietary models like GPT-3/4 or PaLM. Open models often have slightly fewer parameters, partly due to resource limits, but they can be trained on custom data or fine-tuned by anyone. LLaMA's release (although under a research license) was significant because it showed that a carefully trained 65B model can reach the level of a much larger closed model. This suggests that pure size is not everything – optimizations in training and data quality allow smaller open models to compete.

4. Cross-Model Performance: It is informative to compare models of similar era but different origin. For instance, in 2022, OpenAI's 175B InstructGPT (fine-tuned GPT-3) and DeepMind's 70B Chinchilla and BigScience's 176B BLOOM all represented large models, but their performance on tasks like question answering or summarization varied due to differences in training strategy (instruction fine-tuning vs. pure pre-training, etc.). By 2023, LLaMA-65B (Meta) and PaLM-2 340B (Google) could be compared – PaLM-2 is larger and trained on more data, but LLaMA, when fine-tuned (e.g., LLaMA-65B tuned to ChatGPT-style), also demonstrated impressive performance. Generally, larger models tend to perform better, but a smaller well-tuned model can sometimes rival a bigger raw model.

5. Trends in Data Scaling: The table and figure also underscore the trend that training data has grown alongside parameters. Early models such as GPT-2 trained on about 40 GB of text, or roughly 10B tokens from WebText. GPT-3 expanded this to 300B tokens. GPT-4 reportedly used up to 13T token-equivalents, though many may not be unique. Larger datasets provide sufficient examples to

DOI: 10.54254/2753-8818/2025.DL29003

match growing capacity and help reduce overfitting. Recent practice has shifted from scraping massive corpora to curating higher-quality datasets. Newer models also integrate diverse sources such as code, dialogues, multilingual text, mathematics, and conversational data. PaLM-2 is one example of this trend.

In summary, the comparative landscape of LLMs reveals a rapid scaling race tempered by emerging best practices. Larger models generally unlock better performance and new abilities, but with diminishing returns and greater cost. Efficient strategies (like MoE or optimal data usage) have been developed to maximize performance for a given model size. Open models have trailed slightly in scale but shown that smart training can compensate for size to an extent. The next section delves into the practical implications of these massive models, in terms of what they can do and what it costs to use them.

However, as model capabilities improve, ensuring alignment with human values has become a critical challenge. Recent surveys highlight that LLMs risk generating biased, harmful, or adversarial outputs [6]. To mitigate these issues, methods such as reinforcement learning from human feedback (RLHF), constitutional AI, and task decomposition have been proposed. These works emphasize that scaling parameters alone cannot guarantee safety and reliability—alignment mechanisms must be integrated into both training and inference [7].

5. Implications of model size on performance and cost

Scaling up LLMs has clear benefits in terms of performance – but also comes with significant costs and some nuanced trade-offs. In this section, we discuss what extremely large model size means for an LLM's capabilities, as well as the computational, financial, and environmental costs associated with such models.

5.1. Performance and capabilities

Generally, increasing the parameter count of LLMs has enabled higher accuracy and new capabilities on a wide array of tasks. Larger models are better at capturing the complexities of language: they typically yield higher scores on benchmarks for language understanding (e.g., reading comprehension, commonsense reasoning) and generation (e.g., coherence, following instructions). For example, GPT-3's jump to 175B parameters allowed it to achieve strong results on tasks it was never explicitly trained on, simply via prompting. Many later works observed that certain complex skills (multi-step reasoning in math, understanding subtle humor, etc.) start to appear only in models beyond a certain size – an observation aligned with the concept of emergent abilities. In essence, big models seem to develop more robust internal representations and can interpolate or generalize in ways smaller models cannot.

However, bigger is not always straightforwardly better for every metric. One study on knowledge and facts in language models found that simply increasing parameters does not guarantee a proportional increase in "world knowledge" – at least not without the right data. For instance, a survey on LLM evaluation noted that a much smaller fine-tuned model (e.g., a 340M parameter BERT-based QA system) can outperform a zero-shot 175B model on certain knowledge-intensive questions [8]. In other words, specificity and training still matter; a focused smaller model can beat a generic large model if the task is narrow and data-rich. Another consideration is diminishing returns: beyond a point, each additional billion parameters might yield only minor gains on well-trodden benchmarks. This is why scaling laws emphasize that data and compute must grow together with parameters. Without this balance, models risk under-using their capacity.

DOI: 10.54254/2753-8818/2025.DL29003

Large models are generally more robust to distribution shifts and can process a wider range of inputs due to training on diverse data. They also perform better on tasks requiring abstraction or creativity, such as essay writing or coding, where smaller models often lose coherence. However, larger models can generate fluent but incorrect answers, creating an "illusion of competence," and they are harder to control. This challenge has driven research in alignment and fine-tuning with human feedback, extending beyond raw parameter growth. Overall, scaling has brought major gains in capability, enabling applications such as sophisticated dialogue in systems like ChatGPT, but it has also introduced new challenges for reliability and safety.

5.2. Computational and economic cost

The main drawback of very large LLMs is their extreme computational cost. Training GPT-3 (175B) was estimated to require thousands of petaflop/s-days, translating to millions of dollars in cloud GPU time. As model size has grown, training costs have risen superlinearly, making large-scale development increasingly expensive. For instance, while exact figures are not public, GPT-4's training likely cost tens of millions of USD in compute. The energy consumption is also non-trivial. It was reported that training GPT-3 consumed approximately 1.287 GWh of electricity, equating to about 502 metric tons of CO_2 emissions – roughly the carbon footprint of dozens of cars over their lifetimes. Although newer hardware is more efficient, GPT-4's greater size would have proportionally increased energy usage despite efficiency gains. These costs raise concerns about the environmental impact of blindly scaling models.

Beyond training, inference and deployment of large models are expensive as well. Running a 100+ billion parameter model demands significant GPU memory and runtime. For real-time applications (like serving a chatbot to millions of users), the computational cost can be prohibitive. For example, serving GPT-3-level models requires clusters of GPUs and results in high operational expenses. One direct implication is that only a handful of tech companies or well-funded organizations can afford to train the largest models from scratch, which has led to a concentration of power (and sparked efforts like BLOOM to democratize access). Another implication is the push towards model compression and efficient inference: techniques such as quantization, distillation, and low-rank adaptation are actively researched to make large models more tractable to deploy. There is also interest in smaller specialized models that can outperform a giant general model on a specific task with far fewer parameters (sometimes called the "small-model renaissance"), for use cases where deploying a gigantic model is not feasible.

5.3. Memory and latency

Large parameter counts mean large memory footprints. A 175B model in half-precision requires on the order of 350 GB of memory just to store the weights, not counting activations. GPT-3 was thus impossible to run on a single GPU and even inference required model-parallel techniques across multiple accelerators. The latest 1T+ models are beyond the RAM of even high-end GPU arrays, often relying on memory streaming or offloading techniques. This affects latency – it may take seconds to generate each response, which is borderline for interactive applications. Research into optimized model serving, including methods like model sharding and caching, has become critical for practical use of LLMs. Some applications have embraced client-server paradigms (e.g., via APIs) because running these models on consumer devices is currently infeasible.

Furthermore, a distinct line of research focuses on LLM-based autonomous agents. Unlike multi-agent systems, autonomous agents emphasize the capacity of a single LLM to perceive and act

DOI: 10.54254/2753-8818/2025.DL29003

autonomously, integrating memory, reasoning, and planning to operate effectively in open environments. This paradigm offers a new pathway toward artificial general intelligence (AGI) and is likely to become a central theme in future developments [9].

Another promising direction is the application of LLMs in recommendation systems. Unlike traditional approaches that rely on sparse interaction data, LLMs leverage strong text representations and external knowledge to provide zero-shot and few-shot recommendations. This not only enhances the interpretability of recommendations but also enables personalized interactions through natural language, broadening the application scope of LLMs [10].

In summary, the massive scale of LLMs has unlocked remarkable performance gains, but at the cost of very high computation and energy requirements. The industry now faces the trade-off between the benefits of scale and the rising costs. Proposed solutions include more efficient algorithms that reduce parameter needs, advances in hardware, and approaches like retrieval-augmented models such as RETRO, which pair smaller networks with large text databases. Looking ahead, efficiency and cost considerations are expected to shape the trajectory of LLM research as strongly as performance goals.

6. Discussion and future directions

The evolution of LLM parameter size prompts several questions about the future of large language models and whether the current trend of scaling can continue. In this section, we discuss some open challenges and potential directions for research in the era beyond trillion-parameter models.

6.1. Efficiency and optimality

One clear theme is that simply scaling parameters indefinitely is unsustainable. The compute-optimal training paradigm introduced by Hoffmann et al. (2022) suggests that for a given compute, there is an optimal model size and dataset size. Future research will likely focus on finding new optimal frontiers – for example, if we fix an enormous compute budget, what is the best trade-off between model size and data? Already, Chinchilla's result hints that GPT-3 was far larger than optimal for its training data amount, meaning future models could attain GPT-3 level performance (or better) with fewer parameters but more data. Smaller but better might become a mantra, as also evidenced by PaLM 2 focusing on data and quality improvements at 340B instead of going to a trillion. We anticipate algorithms for efficient training (e.g., better optimizers, parallelism, memory saving techniques) will be crucial for unlocking further gains without breaking the bank. In the near term, researchers call for more efficient training and inference algorithms to make LLMs affordable and widely accessible – echoing the sentiment that optimization, rather than brute-force scaling, is the way forward.

6.2. Specialized and multimodal LLMs

Another direction is the development of specialized LLMs rather than one-size-fits-all behemoths. For instance, domain-specific LLMs (in medicine, law, science) might not need to be trillion-parameter models if they leverage domain knowledge more efficiently. Fine-tuning or training medium-sized models (e.g. 10B–50B) on domain-specific data has shown impressive results that sometimes rival much larger general models. Additionally, the community is exploring modular approaches where multiple expert models (each smaller) can be combined or interact, instead of a single monolith. This is related to the idea of LLM-based multi-agent systems, where many smaller

DOI: 10.54254/2753-8818/2025.DL29003

LLM agents collaborate on tasks. Such approaches could bypass the need for a single giant model by distributing intelligence across a team of models with different specialties.

We are also witnessing a rise in multimodal LLMs – models like GPT-4 can accept images as input (GPT-4V) or others that integrate text with audio or video. These extensions often increase parameter count (for new modality-specific components), but future advancements might focus on parameter-efficient fusion of modalities. The success of models like CLIP and Flamingo suggests that language models can be extended to vision without simply scaling parameters drastically, by cleverly combining vision encoders with language decoders. Multimodal capabilities may be a more fruitful path than simply making text-only models larger. These models provide new functionality, such as describing images or controlling robots. Pure text models cannot achieve this. Such advances move LLMs closer to AI systems with a broader understanding of the world.

6.3. Addressing limitations: alignment and interpretability

As LLMs grow more powerful, issues of alignment with human values, controllability, and interpretability become critical. Many alignment issues are independent of parameter count. A 10B model can be as biased or toxic as a 100B model if trained on the same data. The stakes are higher for larger models because of their wide deployment and authoritative tone. As a result, research now emphasizes aligning model behavior with human intentions. Key methods include reinforcement learning from human feedback (RLHF), prompt conditioning, and rule-based "constitutional" AI. Surveys on LLM alignment stress that preventing imprecise, misleading, or harmful outputs is essential, even as capabilities grow. This may require reducing parameter focus in favor of safer behavior, or improving filtering and fine-tuning after training. Additionally, interpretability research aims to understand the internal workings of these gigantic models – an area that becomes harder as parameter count increases. New tools and methods to peer into the attention patterns or neuron activations of LLMs are needed so that researchers can diagnose why a model behaves a certain way or has certain failures. Progress in interpretability could inform better model architectures that achieve the same performance with fewer parameters by identifying redundancies or unused capacity in current models.

6.4. Democratization vs centralization

One concern for the future is the heavy resource requirement concentrating LLM development in a few companies. Projects like BLOOM, OpenLLaMA, and others in the open-source community are exploring ways to produce high-performing LLMs accessible to all, often by leveraging public research collaborations or by fine-tuning publicly released base models. There is a push to find techniques that allow replicating or approximating the big proprietary models at much lower cost – whether through distillation (compressing a large model into a smaller one) or through leveraging public data with slightly smaller models. The success of LLaMA, trained at 65B parameters with relatively modest cost, suggests that 50–100B may be a practical target for academic labs. Models of this scale are achievable and can be shared widely. A likely path forward is community-driven training, where groups pool resources, as in the BigScience project. Such efforts can produce open models that compete with private ones. They also promote inclusive AI development and enable broader research, as more people can study and extend these models.

DOI: 10.54254/2753-8818/2025.DL29003

6.5. New frontiers in modeling

Looking ahead, researchers are exploring architectural changes beyond brute-force parameter scaling. Retrieval-augmented models such as RETRO and Atlas integrate external databases, allowing the model to "look up" information instead of storing it all in parameters. This approach can cut parameter requirements significantly. Smaller models can match the performance of much larger ones by retrieving relevant knowledge from external corpora. Future LLMs might hybridize neural networks with symbolic or retrieval systems more deeply, achieving better performance without exorbitant parameter growth. Another frontier is continuous learning and adaptation: current LLMs are static after training and require full retraining for new data. If models can be made to update themselves in a more efficient online fashion, they might not need to grow larger to absorb new knowledge, instead periodically refreshing a portion of their weights.

In summary, while the past five years were largely about scaling up to unprecedented model sizes, the next phase of LLM research is likely to emphasize quality over quantity – making models smarter, safer, more efficient, and more versatile, rather than just bigger. There is consensus in the community that purely pursuing parameter count is yielding diminishing returns, and issues such as cost, environmental impact, and the need for human-aligned behavior call for a pivot in strategy. As one recent survey concluded, future progress will require focusing on "improving the accuracy and performance of these models, addressing their limitations, and exploring new ways to use them" rather than just scaling up. We expect to see innovative model designs, training methods, and interdisciplinary approaches (combining learning with knowledge bases, multimodal inputs, etc.) defining the state of the art in the coming years. The notion of what constitutes a "large" language model may shift from raw parameter count to a combination of factors that together produce powerful AI systems.

7. Conclusion

The development of large language models has been closely tied to the expansion of parameter sizes. Models grew from millions of parameters in the late 2010s to more than a trillion by 2023. This growth unlocked new capabilities and reshaped the field of natural language processing. Scaling models such as the GPT series and PaLM improved performance and produced emergent abilities absent in smaller models, confirming predictions from scaling laws. Technical advances—including the Transformer architecture, parallel training, and mixture-of-experts layers—made this expansion possible. Comparative studies show exponential growth in scale but also indicate that size alone is not sufficient. The quality of data, the design of training objectives, and efficiency considerations remain essential.

Large language models with hundreds of billions of parameters have become core tools with wide applications. Their growth, however, carries high computational cost and practical challenges. Larger models require more training data and greater compute resources, raising concerns about accessibility, economic expense, and environmental impact. The field now recognizes that progress must balance performance with sustainable and inclusive practices. Future work is expected to move from raw scaling toward smarter optimization. Likely directions include identifying optimal model sizes, improving data efficiency, and using retrieval augmentation to increase utility per parameter. Multimodal modeling, domain-specific fine-tuning, and stronger alignment methods are also set to shape the next generation more than sheer parameter count.

In conclusion, the growth of parameter size has been a central driver of advances in language models. Tracing this evolution helps explain current performance and indicates how future progress

DOI: 10.54254/2753-8818/2025.DL29003

may unfold. The era of rapid scaling may be reaching its peak, yet the lessons from pushing size to the limit remain valuable. These lessons will guide the design of models that are more intelligent, efficient, and safe. The history of large language models shows a clear theme: scaling matters, but it is not the only factor. Innovation in algorithms and responsible deployment will ultimately shape their impact on technology and society.

References

[1] Guo, T., Chen, X., Wang, Y., et al. (2023) Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. arXiv preprint arXiv: 2309.15025.
[2] Chang, Y., Wang, X., Wu, Y., et al. (2024) A Survey on Evaluation of Large Language Models. ACM Computing Surveys, 55(3), Article 39.
[3] Minaee, S., Mikolov, T., Nikzad, N., et al. (2023) Large Language Models: A Survey. arXiv preprint arXiv: 2402.06196.
[4] Zhou, K., Li, J., Tang, T., Hou, Y., et al. (2023) A Survey of Large Language Models. arXiv preprint arXiv: 2303.18223.
[5] Yin, S., Fu, C., Zhao, S., et al. (2024) A Survey on Multimodal Large Language Models. National Science Review, 11, nwae403.
[6] Hadi, M.U., Al Tashi, Q., Qureshi, R., et al. (2023) Large Language Models: A Comprehensive Survey of Applications, Challenges, Limitations, and Future Prospects. TechRxiv Preprint.
[7] Shen, T., Jin, R., Huang, Y., et al. (2023) Large Language Model Alignment: A Survey. arXiv preprint arXiv: 2309.15025.
[8] Wang, Y., Zhong, W., Li, L., et al. (2023) Aligning Large Language Models with Human: A Survey. arXiv preprint arXiv: 2307.12966.
[9] Wang, L., Ma, C., Feng, X., et al. (2024) A Survey on Large Language Model Based Autonomous Agents. Frontiers of Computer Science.
[10] Wu, L., Zheng, Z., Qiu, Z., et al. (2024) A Survey on Large Language Models for Recommendation. World Wide Web, 27(60).