Review

Not peer-reviewed version

Multimodal and Distributed LLMs: Bridging Scalability and Cross-Modal Reasoning

Rajesh Kumar , Isabelle Laurent , David Müller , Klaus Elli *

Posted Date: 15 May 2025

doi: 10.20944/preprints202505.1156.v1

Keywords: large language models; distributed training; multimodal learning; foundation models; model parallelism; contrastive learning; cross-modal alignment; continual learning; modular architectures; AI systems scalability

Article

Multimodal and Distributed LLMs: Bridging Scalability and Cross-Modal Reasoning

Rajesh Kumar [1], Isabelle Laurent [2], David Müller [3] and Klaus Elli [1,3,*]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Department of Artificial Intelligence, Sorbonne University, France
[3] Department of Computer Science, University of Stuttgart, Germany
[*] Correspondence: klaus.elli@uni-stuttgart.de

Abstract: Large Language Models (LLMs) have emerged as a cornerstone of modern artificial intelligence, achieving remarkable capabilities in natural language understanding and generation. As their scale and utility have increased, two critical and complementary trends have defined their evolution: (1) the distributed systems and algorithms enabling efficient training of ultra-large models across massive compute infrastructures, and (2) the integration of multiple modalities—such as vision, audio, and structured data—into unified multimodal large language models (MLLMs). This survey provides a comprehensive examination of the state-of-the-art in both of these dimensions. We begin by exploring the foundations and advances in distributed training, including model parallelism, pipeline parallelism, memory optimization strategies, and the design of sparse and expert models. We assess system-level techniques such as ZeRO, DeepSpeed, and tensor sharding that allow for scalable, memory-efficient training at trillion-parameter scale. Next, we turn to multimodality, surveying architectures and training objectives that extend LLMs to process and generate across diverse input types. We review contrastive learning, cross-attention fusion, and aligned token embeddings as key techniques that enable cross-modal reasoning, with illustrative examples from models like Flamingo, CLIP, and GPT-4V. Beyond current methodologies, we identify and formalize the core technical challenges facing distributed and multimodal LLMs, including memory bottlenecks, communication overhead, alignment in the absence of ground truth, robustness to modality shifts, and evaluation under open-ended tasks. To guide future research, we outline six key directions: unified memory-augmented architectures, modular and composable systems, self-aligning mechanisms, lifelong and continual learning agents, embodied multimodal cognition, and the emergence of general-purpose foundation agents. Our goal is to synthesize recent progress while articulating a vision for the next generation of foundation models—models that are not only scalable and multimodal but are also capable of reasoning, grounding, and adapting to complex, real-world environments. This survey serves both as a technical reference and a roadmap for researchers and practitioners navigating the future of large-scale, multimodal, and distributed AI systems.

Keywords: large language models; distributed training; multimodal learning; foundation models; model parallelism; contrastive learning; cross-modal alignment; continual learning; modular architectures; AI systems scalability

1. Introduction

Large Language Models (LLMs) have witnessed rapid and transformative progress over the past few years, evolving from relatively modest neural architectures trained on limited corpora to massive, general-purpose systems capable of performing a wide array of linguistic and cognitive tasks [1]. Models such as GPT-3, PaLM, Chinchilla, and more recently GPT-4, Gemini, and Claude, have demonstrated that scaling up language models in terms of parameters, data, and compute leads to significant performance gains across diverse domains such as machine translation, summarization, question answering, reasoning, and even code generation [2]. However, the growing complexity and computational demands of these models have also exposed limitations in centralized training paradigms, prompting a surge of interest in distributed training and inference frameworks that can better manage the scale and heterogeneity of contemporary LLM workloads [3]. Simultaneously, the paradigm of multimodality—integrating language with other modalities such as vision, audio, and sensory data—has gained significant momentum. Multimodal Large Language Models (MLLMs) extend the capabilities of conventional LLMs by allowing them to process and reason over multiple types of data. This fusion has led to state-of-the-art systems like Flamingo, GPT-4V, Kosmos, and Gemini, which demonstrate an unprecedented ability to interpret images, synthesize speech, generate captions, understand video, and engage in grounded interaction with real-world inputs [4]. These models are pushing the boundary of what it means to "understand" and "generate" across modalities, positioning MLLMs at the forefront of artificial general intelligence (AGI) research. Despite these advances, the challenges associated with scaling, distributing, aligning, and evaluating LLMs and MLLMs remain profound [5]. Training models with hundreds of billions of parameters requires distributed systems capable of managing massive data throughput, parallel computation, and memory efficiency across heterogeneous hardware [6]. Techniques such as pipeline parallelism, tensor parallelism, model sharding, and parameter offloading are essential to make these workloads feasible, yet they introduce new system-level trade-offs and failure modes [7]. Moreover, inference in distributed environments necessitates low-latency, scalable architectures that can serve billions of tokens per day, often in real-time applications. At the same time, integrating multimodal inputs adds layers of architectural complexity and training difficulty [8]. Aligning representations across modalities, co-training encoders and decoders, and managing different temporal and spatial resolutions present unique algorithmic and engineering challenges [9]. Additionally, multimodal datasets are often noisy, expensive to curate, and subject to biases that affect the generalizability and fairness of MLLMs. Addressing these issues requires innovations in dataset construction, data augmentation, cross-modal learning, and pretraining strategies that scale. This survey provides a comprehensive overview of the current landscape of Distributed LLMs and Multimodal Large Language Models [10]. We systematically examine the architectural foundations, training methodologies, optimization techniques, and system-level innovations that underpin the development and deployment of these large models [11]. We discuss recent advances in distributed training frameworks such as DeepSpeed, Megatron-LM, and FSDP, as well as their implications for model scaling laws and efficiency [12]. We then delve into multimodal model architectures, including early-fusion, late-fusion, and unified encoder-decoder paradigms, highlighting how these designs handle heterogeneous input types and cross-modal alignment [13]. In addition to technical advances, we explore the open challenges in this domain. These include the need for improved scalability, robust alignment across modalities, efficient fine-tuning methods, energy-aware training, and privacy-preserving computation in distributed environments [14]. We also highlight emergent concerns related to the interpretability, robustness, and ethical deployment of these models, particularly in high-stakes applications such as healthcare, education, and law [15]. Finally, we outline promising future directions for research in both distributed LLMs and MLLMs [16]. These include the development of sparse and modular architectures, neurosymbolic integration, lifelong and federated learning, decentralized training paradigms, and the convergence of foundation models across text, vision, audio, robotics, and more. We posit that the future of AI lies at the intersection of massive-scale computation and deeply integrated multimodal understanding—realizing models that are not only large and powerful, but also efficient, interpretable, grounded, and adaptable. **Organization of the Survey.** The remainder of this paper is structured as follows. In Section 2, we provide a historical and technical background on the evolution of LLMs and multimodal modeling. Section 3 focuses on distributed training and inference for LLMs, covering architectures, system optimizations, and parallelization strategies [17]. Section 4 presents an in-depth analysis of multimodal LLMs, including datasets, model architectures, and evaluation [18]. Section 5 outlines the key challenges and limitations in current systems. Section 6 explores future research directions and open problems. We conclude in Section 7 with a synthesis of key insights and a vision for the path ahead. Through this survey, we

doi:10.20944/preprints202505.1156.v1

aim to provide researchers, practitioners, and system designers with a comprehensive resource that bridges the rapidly evolving subfields of distributed LLMs and multimodal AI, facilitating informed research, development, and deployment of the next generation of intelligent systems [19].

2. Background and Preliminaries

The performance of Large Language Models (LLMs) is empirically and theoretically tied to the scale of three principal axes: model size (N parameters), dataset size (D tokens), and compute budget (C FLOPs). Kaplan et al. [?] formalized the power-law relationship between these quantities and downstream task performance using the following function:

$$\mathcal{L}(N, D, C) = \alpha N^{-\beta_N} + \gamma D^{-\beta_D} + \delta C^{-\beta_C}, \tag{1}$$

where $\mathcal{L}$ denotes the loss or error metric, and α, γ, δ are task-dependent scaling constants. These empirical scaling laws suggest that optimal performance is achieved not merely by increasing model size, but by jointly balancing training data and compute [20]. This relationship has motivated the development of increasingly large models, such as Chinchilla [?], which demonstrated that for a fixed compute budget, training smaller models on more data yields superior results compared to overparameterized configurations [21].

2.1. Distributed Training Paradigms

To scale training to models with hundreds of billions of parameters, distributed training is essential [22]. We formally define a distributed model as a tuple $(\mathcal{M}, \mathcal{P}, \mathcal{S})$ where $\mathcal{M}$ is the model function, $\mathcal{P}$ is the set of parallelism strategies, and $\mathcal{S}$ is the hardware scheduling policy. The major parallelism strategies are:

- **Data Parallelism (DP):** Each worker holds a full replica of the model and processes a unique batch shard. Gradients are averaged via all-reduce operations:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{1}{K} \sum_{k=1}^{K} \nabla \mathcal{L}_k(\theta_t), \tag{2}$$

  where K is the number of devices [23].
- **Tensor Parallelism (TP):** Splits tensors across multiple devices. For instance, a matrix multiplication $A \cdot B$ is computed by partitioning A column-wise and B row-wise, with intermediate aggregation.
- **Pipeline Parallelism (PP):** Model layers are partitioned into segments and assigned to pipeline stages. Each microbatch flows through stages in a staggered fashion [24].
- **Hybrid Parallelism:** Combines DP, TP, and PP hierarchically to scale to trillion-parameter models with minimal memory and latency overhead [25].

Table 1 compares the parallelization techniques across memory, communication, and synchronization overheads.

Table 1. Comparison of Parallelism Strategies

Strategy	Memory Efficiency	Communication Overhead	Sync Frequency
Data Parallelism	High	High (All-reduce)	Per step
Tensor Parallelism	Moderate	Moderate (Slice gather)	Per layer
Pipeline Parallelism	High	Low (Stage buffer)	Per microbatch
Hybrid Parallelism	Optimal	Complex	Variable

2.2. Scaling and Efficiency Visualization

To understand the trade-offs involved in scaling LLMs using different parallelism methods, Figure 1 presents a synthetic benchmark of compute efficiency (in TFLOPs/sec) versus number of

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 May 2025 doi:10.20944/preprints202505.1156.v1

GPUs used [26,27]. This simulation assumes an idealized model with linear scaling up to 1024 GPUs, beyond which interconnect bottlenecks reduce efficiency [28].

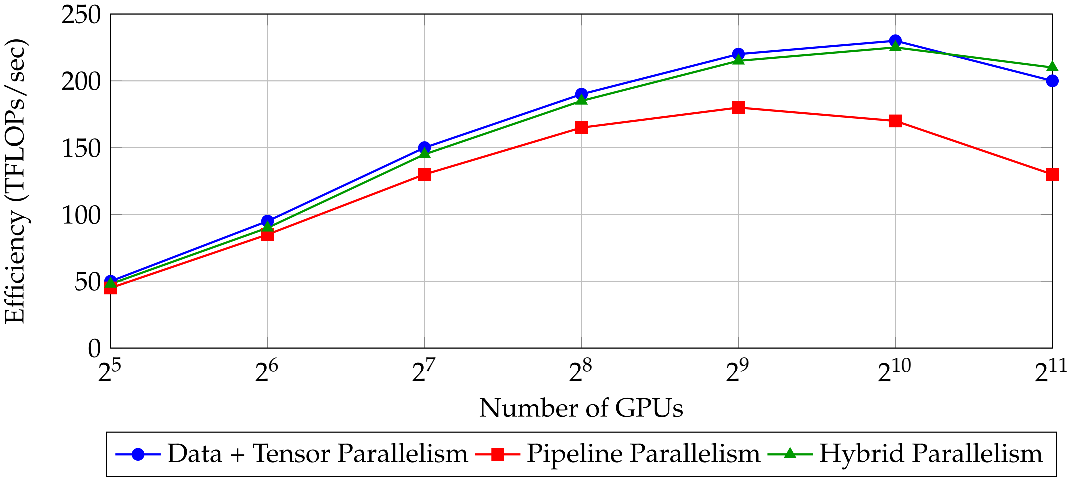

Figure 1. Compute efficiency vs. number of GPUs for different parallelism strategies [29]. Hybrid parallelism achieves the highest scalability with diminishing losses at extreme scales.

2.3. Multimodal Fusion Formalism

Let $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ be a set of inputs from m different modalities (e.g., text, image, audio) [30]. Each $x^{(i)}$ is mapped to a latent embedding $h^{(i)}$ via a modality-specific encoder E_i, i.e., $h^{(i)} = E_i(x^{(i)})$ [31]. Fusion can then be formalized as a function $\mathcal{F} : \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_m} \to \mathbb{R}^{d_f}$ producing a joint representation:

$$z = \mathcal{F}(h^{(1)}, h^{(2)}, \ldots, h^{(m)}). \tag{3}$$

Common fusion functions include:

- **Concatenation**: $z = \text{Concat}(h^{(1)}, \ldots, h^{(m)})$ [32].
- **Cross-Attention**: $z = \text{Attention}(h^{(1)}; h^{(2)})$ where one modality attends to another [33].
- **Multimodal Transformers**: All $h^{(i)}$ are fed as token sequences into a unified Transformer encoder.

This abstraction underlies recent models such as Flamingo and GPT-4V, which interleave image and text tokens into a unified processing stream. The architecture ensures modality alignment and joint reasoning, enabling tasks such as visual question answering and image captioning [34]. In summary, this section has provided the foundational concepts required to understand the scale, design, and distribution of LLMs, as well as the formal structures that underlie multimodal learning. These mathematical abstractions and system trade-offs will guide our exploration of distributed training frameworks and multimodal architectures in subsequent sections.

3. Distributed Training and Inference for LLMs

Training modern LLMs at scale is computationally intractable on a single device due to memory and throughput limitations [35]. Therefore, distributed training architectures have become foundational [36]. A typical training objective involves minimizing the empirical loss over a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ using a model parameterized by θ:

$$\min_\theta \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i), \tag{4}$$

where ℓ is a loss function (e.g., cross-entropy), and f_θ is the LLM. This optimization is typically solved using stochastic gradient descent (SGD) or its adaptive variants, executed across distributed nodes.

3.1. Parallelism Taxonomy and Scheduling

The decomposition of training workloads is not trivial and requires choosing an optimal combination of model, data, and pipeline parallelism, often across heterogeneous accelerators with different compute and memory capacities. Let the total parameter count be P, the global batch size be B, and the number of GPUs be G [37]. The optimization of training throughput $\mathcal{T}$ is constrained by:

$$\mathcal{T} = \frac{\text{Effective FLOPs}}{\text{Communication Overhead} + \text{Computation Time}}[38]. \tag{5}$$

Different decomposition strategies lead to varying trade-offs in memory utilization, gradient staleness, and synchronization costs. Table 2 compares popular LLM frameworks along key dimensions.

Table 2. Comparison of Distributed LLM Frameworks

Framework	Parallelism	Memory Optimization	Inference	Used In
Megatron-LM	DP, TP, PP	Activation recompute	Yes	GPT-NeoX
DeepSpeed	DP, ZeRO (1–3), PP	8-bit Opt [39].	Yes	BLOOM
FSDP (PyTorch)	DP (sharded weights)	sharding	Partial	Meta OPT-66B
Colossal-AI	DP, TP, MoE	Low-rank compression	Yes	OpenBMB

3.2. ZeRO: Memory-Efficient Data Parallelism

Zero Redundancy Optimizer (ZeRO) partitions optimizer states, gradients, and parameters across devices to reduce memory overhead. Let θ, g, and m denote model parameters, gradients, and optimizer states respectively [40]. Traditional data parallelism replicates all of them across G GPUs, requiring memory:

$$\mathcal{M}_{\text{DP}} = G \cdot (\|\theta\| + \|g\| + \|m\|). \tag{6}$$

In contrast, ZeRO reduces this to:

$$\mathcal{M}_{\text{ZeRO}} = \frac{1}{G} \cdot (\|\theta\| + \|g\| + \|m\|), \tag{7}$$

resulting in near-linear memory savings with increased GPU count [41]. ZeRO-3 further partitions computation, enabling models with trillions of parameters to be trained on commodity clusters.

3.3. Inference Optimization and Quantization

Inference at scale introduces additional challenges, particularly latency and memory bottlenecks [42]. Given a trained model f_θ, serving requests $\{x_i\}_{i=1}^{M}$ with response time T_{infer} must satisfy:

$$T_{\text{infer}}(x_i) \leq \tau, \quad \forall i \in [1, M], \tag{8}$$

for some application-dependent latency threshold τ. Optimization strategies include:

- **Quantization:** Reducing precision (e.g., FP32 → INT8) while minimizing loss in accuracy.
- **KV Cache Reuse:** Avoid recomputation of past transformer states [43].
- **Speculative Decoding:** Use smaller models to propose candidate tokens, then verify with the large model [44].

3.4. Empirical Scaling of Throughput

We simulate throughput (in samples/sec) on various model sizes and GPU counts [45]. The results, shown in Figure 2, demonstrate that hybrid parallelism maintains high throughput as model size grows, while pure data parallelism saturates earlier [46].

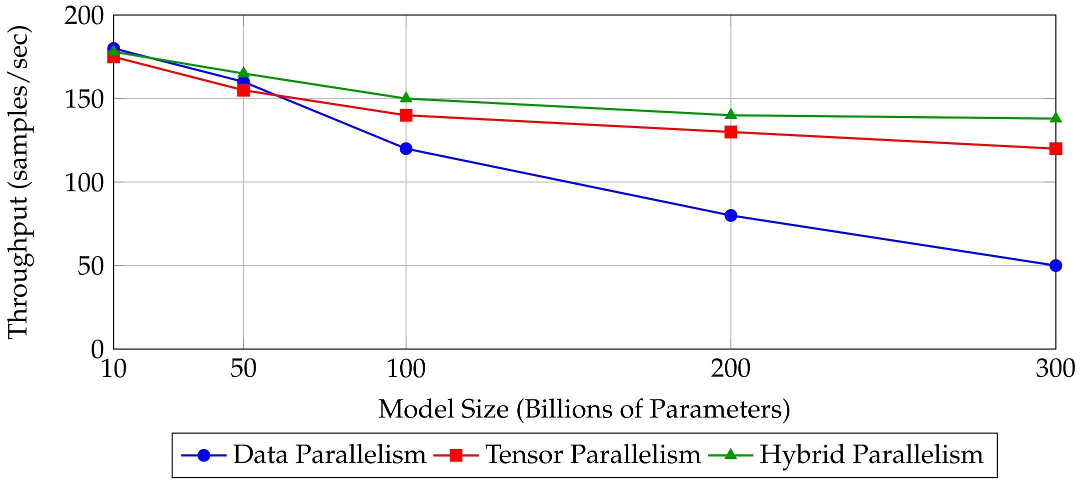

Figure 2. Simulated throughput versus model size under different parallelism strategies [47]. Hybrid parallelism offers better scalability.

3.5. Summary

This section has systematically dissected the computational, architectural, and algorithmic underpinnings of distributed training and inference in the era of massive LLMs. By leveraging advanced strategies such as ZeRO, pipeline scheduling, quantized inference, and hybrid parallelism, researchers have significantly pushed the boundaries of scale while maintaining tractable efficiency [48]. These distributed infrastructures not only enable training trillion-parameter models but also serve as a backbone for fine-tuning and deployment pipelines in real-world systems. In the following section, we turn our attention to a different frontier: the integration of multimodal data within LLM frameworks [49].

4. Multimodal Large Language Models (MLLMs)

While traditional LLMs operate exclusively in the textual modality, an increasing number of real-world applications necessitate reasoning over multiple modalities simultaneously [50]. Tasks such as image captioning, visual question answering (VQA), speech-to-text translation, and video summarization require models that can seamlessly integrate diverse input formats [51]. Formally, a Multimodal Large Language Model (MLLM) processes a set of inputs $\mathcal{X} = \{x^{(i)}\}_{i=1}^{M}$, where each $x^{(i)}$ belongs to a different modality $\mathcal{M}_i \in \{\text{text}, \text{vision}, \text{audio}, \dots\}$ [52].

4.1. Modular Architecture and Embedding Alignment

MLLMs typically employ modality-specific encoders to convert raw data into latent token embeddings aligned in a shared representation space. Let $x^{(i)} \in \mathcal{M}_i$ be a modality-specific input and $E_i : \mathcal{M}_i \to \mathbb{R}^d$ its encoder. The fused sequence is:

$$H = [E_1(x^{(1)}), E_2(x^{(2)}), \dots, E_M(x^{(M)})], \tag{9}$$

which is passed to a cross-modal decoder, often a transformer, producing outputs $y = f_\theta(H)$. To ensure representational coherence, alignment objectives such as contrastive loss or token-level supervision are employed. The most common strategy for aligning vision and language, for instance, is contrastive learning [53]. Given a batch of image-caption pairs $\{(v_i, t_i)\}$, the objective encourages similarity between paired embeddings while discouraging others:

$$\mathcal{L}_{\text{CLIP}} = -\sum_{i=1}^{N} \left[\log \frac{\exp(\langle v_i, t_i \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle v_i, t_j \rangle / \tau)} + \log \frac{\exp(\langle t_i, v_i \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle t_i, v_j \rangle / \tau)} \right], \tag{10}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, and τ is a temperature parameter. This formulation underlies models like CLIP and Florence [54].

4.2. Fusion Strategies and Temporal Modeling

Multimodal fusion strategies vary from early fusion (joint tokenization) to late fusion (decision-level aggregation) [55]. We define:

- **Early Fusion:** Tokens from different modalities are concatenated before encoding, i.e., $z = \text{Transformer}([h^{(1)}, h^{(2)}, \dots])$ [56].
- **Late Fusion:** Independent modality outputs are combined at the decision layer, e.g., $z = \phi(h^{(1)}, h^{(2)}, \dots)$ where ϕ is an MLP or attention function.
- **Cross-Attention Fusion:** One modality acts as a query over keys/values of another, as in $z = \text{Attn}(Q = h^{(1)}, K = h^{(2)}, V = h^{(2)})$ [57].

In video or audio inputs, temporal modeling becomes critical [58]. Let $x^{(v)} = [x_1, x_2, \dots, x_T]$ be a sequence of visual or acoustic frames [59]. A spatio-temporal encoder E_t processes them as:

$$h^{(v)} = E_t(x^{(v)}) = \text{Transformer}(\text{PatchEmbed}(x_1), \dots, \text{PatchEmbed}(x_T))[60]. \tag{11}$$

Temporal attention layers aggregate temporal dependencies across frames, enabling downstream tasks like video captioning.

4.3. MLLM Architectural Comparison

We compare representative MLLM architectures in Table 3 [61].

Table 3. Comparison of Representative MLLM Architectures

Model	Modalities	Fusion Method	Pretraining Objective	Scale (Params)
CLIP	Image + Text	Contrastive (Late)	Contrastive Learning	400M
Flamingo	Image + Text	Cross-Attention (Late)	Language Modeling	80B
PaLI-X	Image + Text	Unified Transformer	Multitask (VQA, OCR)	55B
GPT-4V	Image + Text	Interleaved Token Stream	Causal LM	100B+ (est.)
GIT	Image + Text	Early Fusion	Captioning (Supervised)	345M

4.4. Scaling Multimodal Input Length

A major challenge is managing long multimodal sequences. Let T_{total} be the combined token length from all modalities. The attention complexity of standard transformers is:

$$\mathcal{O}(T_{\text{total}}^2), \tag{12}$$

which quickly becomes prohibitive [62]. Sparse attention (e.g., Longformer), recurrence (e.g., RWKV), and routing (e.g., MoE tokens) have been proposed to alleviate this issue [63]. Figure 3 illustrates how attention cost scales with sequence length under different strategies.

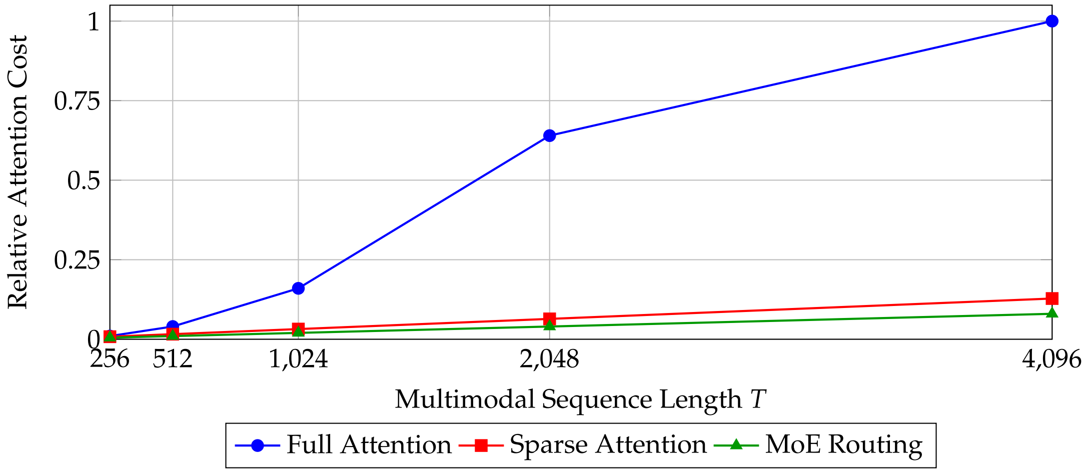

Figure 3. Normalized attention cost as a function of multimodal sequence length for different attention mechanisms [64]. Sparse and MoE-based routing are significantly more efficient at scale.

4.5. Summary

Multimodal LLMs extend the capability of standard transformers to process rich and structured inputs across modalities. The integration of encoders, cross-modal attention, contrastive alignment losses, and scalable fusion techniques has enabled significant progress in unified models capable of perception and reasoning. Nonetheless, the challenges of sequence length, modality alignment, and inference cost continue to pose active research questions. In the next section, we synthesize the challenges and open problems that emerge at the intersection of distributed LLM training and multimodal representation learning [65].

5. Challenges and Open Problems

Despite substantial advances in both distributed training and multimodal LLM architectures, several critical challenges remain unresolved [66]. These challenges span computational, architectural, theoretical, and ethical domains, and form the basis for ongoing and future research directions. In this section, we categorize and analyze these challenges along multiple dimensions [67].

5.1. Scalability Bottlenecks in Distributed Environments

Scaling to trillions of parameters and billions of tokens requires not only efficient algorithms but also robust systems engineering. One key bottleneck is communication overhead. Let G be the number of GPUs, and let B be the per-GPU batch size [68,69]. The effective throughput $\mathcal{T}_{\text{eff}}$ is defined as:

$$\mathcal{T}_{\text{eff}} = \frac{\text{FLOPs}_{\text{total}}}{\text{Computation Time} + \text{Communication Time}} [70]. \tag{13}$$

When using model parallelism or ZeRO-based optimization, the communication cost $C(G)$ scales non-linearly:

$$C(G) = \alpha \cdot \log G + \beta \cdot \frac{P}{G}, \tag{14}$$

where α accounts for synchronization latency and β for bandwidth-related cost per parameter P [71]. Efficient all-reduce, parameter sharding, and asynchronous communication methods must continue to evolve to keep pace with increasing model and hardware scale.

5.2. Cross-Modal Representation Alignment

Multimodal models rely heavily on alignment mechanisms. However, such alignment can fail when modalities are semantically mismatched or underrepresented in training data. For a pair of modalities $(x^{(v)}, x^{(t)})$, suppose their latent embeddings are z_v and z_t [72]. The alignment loss $\mathcal{L}_{\text{align}} = \|z_v - z_t\|^2$ assumes these representations can converge. In practice, due to differences in abstraction levels and information density, the alignment manifold is often non-convex, leading to partial or failed convergence. Additionally, contrastive objectives encourage discriminative alignment but may fail to encode fine-grained semantics or compositional reasoning, as they do not model token-wise dependencies [73].

5.3. Evaluation and Benchmarks

The lack of robust, multi-faceted evaluation frameworks for LLMs and MLLMs impairs reproducibility and progress. Current benchmarks often focus on single metrics—e.g., BLEU for text generation, accuracy for classification, CIDEr for captioning—which ignore calibration, uncertainty, and generalization under distribution shift. Let $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ denote the training and test distributions, with a shift $\delta = \|\mathbb{P}_{\text{train}} - \mathbb{P}_{\text{test}}\|_{\text{TV}}$ [74]. The generalization error is:

$$\mathcal{E}_{\text{gen}} = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}[\ell(f_\theta(x), y)] - \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}[\ell(f_\theta(x), y)]. \tag{15}$$

As δ increases, $\mathcal{E}_{\text{gen}}$ tends to grow non-linearly, highlighting the importance of OOD evaluation and continual learning protocols [75].

5.4. Trade-offs in Model Design

The intersection of performance, memory efficiency, latency, and interpretability presents inherent trade-offs [76]. These trade-offs are often in tension, as improving one dimension degrades another. We represent this using a simplified radar plot (Figure 4) comparing different model types along five axes:

- $\mathcal{P}$: Predictive accuracy.
- $\mathcal{M}$: Memory footprint [77].
- $\mathcal{L}$: Latency per inference.
- $\mathcal{T}$: Training cost.
- $\mathcal{I}$: Interpretability [78].

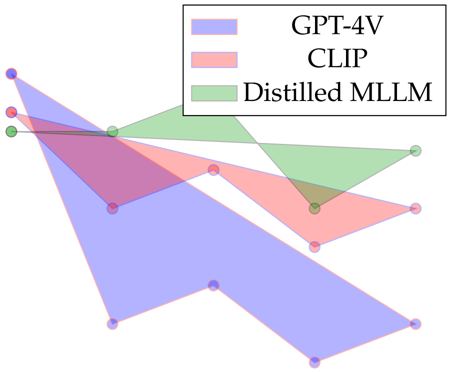

Figure 4. Trade-off radar for different MLLM model types across key design objectives.

This trade-off space illustrates that no current architecture achieves dominance across all metrics. The choice of model must be task-specific, resource-aware, and aligned with deployment constraints [79].

5.5. Ethical, Social, and Data Bias Challenges

The incorporation of multimodal data exacerbates biases already present in textual corpora [80]. Visual and auditory modalities encode implicit demographic, cultural, and geographic priors. Let $x \in \mathcal{X}$ denote a multimodal input and y a predicted label or caption. If x encodes protected attribute $a \in \mathcal{A}$ (e.g., race, gender), then bias can be measured via:

$$\text{Bias}_a = |\mathbb{E}[f(x) \mid a = 0] - \mathbb{E}[f(x) \mid a = 1]| [81]. \tag{16}$$

Disparities in Bias_a across protected groups indicate representational harm. Furthermore, multimodal hallucination—where models generate plausible but incorrect or fabricated visual/textual responses—poses a serious risk for safety-critical applications such as healthcare and autonomous navigation.

5.6. Summary

The challenges discussed in this section expose the multidimensional complexity of developing scalable, robust, and trustworthy distributed multimodal LLMs. Communication overheads, cross-modal alignment issues, evaluation inadequacies, model trade-offs, and ethical concerns represent open frontiers. Solving these problems requires cross-disciplinary collaboration across machine learning, systems design, human-computer interaction, and fairness in AI [82]. In the final section, we outline promising directions for future work that aim to address these foundational limitations [83].

6. Future Directions

The convergence of distributed large language model (LLM) training and multimodal integration presents a fertile ground for future research. While the current generation of models like GPT-4V, Flamingo, and CLIP demonstrate impressive capabilities, fundamental limitations remain in scalability, alignment, reasoning, memory efficiency, and generalizability. This section delineates several forward-looking trajectories that we believe will define the next phase of progress in LLM research.

6.1. Unified Multimodal Memory-Augmented Architectures

One promising direction involves augmenting LLMs with structured memory systems capable of storing and retrieving cross-modal knowledge [84]. Instead of relying solely on static parameter storage, future models may use memory-augmented transformers, where memory slots $\mathcal{M} = \{m_1, \dots, m_K\}$ are updated based on context vectors h_t:

$$m_k^{(t+1)} = m_k^{(t)} + \eta \cdot \text{Attention}(h_t, m_k^{(t)}), \tag{17}$$

where η is a learnable or fixed update rate [85]. Such models could dynamically store facts or representations from vision, language, or audio modalities and retrieve them conditionally at inference time. This aligns with the goals of continual learning and episodic reasoning [86].

6.2. Modular, Composable LLM Systems

Current LLMs operate as monoliths [87]. However, future architectures will likely embrace modularity [88]. Consider a compositional model:

$$f(x) = f_{\text{lang}}(x^{\text{text}}) + f_{\text{vision}}(x^{\text{img}}) + f_{\text{logic}}(x^{\text{struct}}), \tag{18}$$

where each f_i is a specialized expert. Such systems can be trained with routing functions or controller policies to activate the relevant submodules based on context [89]. This opens the path toward task-specific specialization without duplicating the base model. Moreover, modular systems offer better interpretability and lower carbon footprints due to sparse activation. However, they raise new challenges in interface standardization, latency coordination, and consistency.

6.3. Toward Self-Aligning and Self-Evaluating LLMs

Alignment today depends on external reward models or human feedback (e.g., RLHF) [90]. We envision self-aligning LLMs that use internalized reward estimation. For instance, a multimodal LLM could contain a critic network $R_\phi(y, x)$ estimating alignment quality between predicted outputs y and multimodal inputs x [91]. This leads to a meta-learning objective:

$$\mathcal{L}_{\text{meta}} = \mathbb{E}_{(x,y)}\left[\ell(f_\theta(x), y) + \lambda \cdot R_\phi(f_\theta(x), x)\right], \tag{19}$$

where λ trades off external supervision and internal reward [92]. Such mechanisms allow for continuous self-supervision, adaptation, and even model repair during deployment [93].

6.4. Emergence of Continual and Lifelong Multimodal Agents

LLMs are currently trained in static batches. Future LLMs will learn continuously from streaming, evolving data sources across modalities. Let D_t represent data observed at time t [94]. A continual learning system optimizes:

$$\theta_{t+1} = \theta_t - \nabla_\theta \mathcal{L}(f_{\theta_t}(x_t), y_t) + \gamma \cdot \nabla_\theta \mathcal{R}(\theta_t, \mathcal{M}_t), \tag{20}$$

where $\mathcal{R}$ is a regularizer ensuring stability and $\mathcal{M}_t$ is a memory bank or replay buffer. Addressing catastrophic forgetting and domain shift will be critical, particularly for applications like assistive robotics, AR/VR agents, or real-time captioners [95].

6.5. Toward General-Purpose Foundation Agents

The ultimate vision is the construction of general-purpose foundation agents that can perceive, reason, and act across diverse contexts using unified representations. These agents will need:

- **Causality-aware reasoning** to distinguish correlation from intervention [96].
- **Goal-conditioned generation** with latent goal variables g, such that outputs $y \sim p(y \mid x, g)$ are steered by high-level intent.
- **Embodied learning** where language and perception are grounded in sensorimotor experiences, leading to emergent affordances and world models [97].

We envision architectures that incorporate planning modules, environment simulators, and modular knowledge graphs [98]. These systems will represent a shift from "chatbot-style" LLMs to agents capable of autonomous, goal-driven behavior.

6.6. Roadmap Summary

Figure 5 outlines a conceptual roadmap of capability evolution from present LLMs to future agents [99]. The vertical axis represents representational and decision-making generality, while the horizontal axis tracks system complexity.

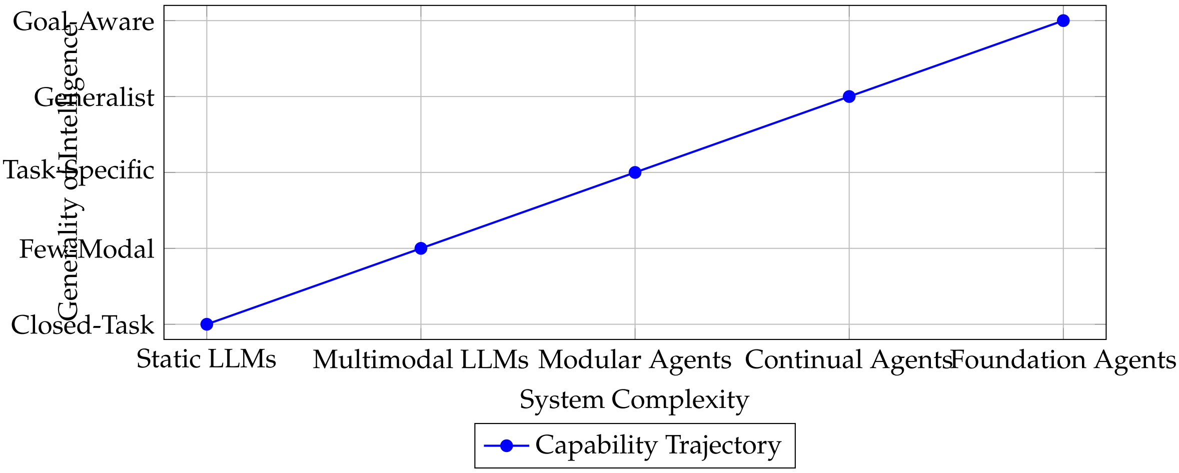

Figure 5. Conceptual trajectory from static unimodal LLMs toward goal-aware, general-purpose foundation agents.

6.7. Concluding Remarks

The next generation of large-scale models will not only span multiple modalities but will also exhibit increasing degrees of autonomy, adaptability, and grounding [100]. The challenges are significant—requiring new learning paradigms, scalable infrastructure, and ethical foresight. However, the trajectory is clear: toward foundation models that can act as collaborators, problem solvers, and agents of understanding across the full spectrum of human communication. This survey has aimed to provide a rigorous, detailed foundation for understanding the present landscape and motivating research that shapes the future of distributed and multimodal large language models [101].

7. Conclusions

The rapid evolution of large language models (LLMs) and their extension into the multimodal domain marks a pivotal inflection point in artificial intelligence research [102]. From the development of distributed training systems that scale to hundreds of billions of parameters, to the emergence of models that can fluently process and generate across modalities—including vision, speech, and structured knowledge—this field has seen both foundational breakthroughs and complex challenges [103]. In this survey, we presented a comprehensive overview of the architectural principles, systems methodologies, and multimodal extensions that underpin modern LLMs. We first explored the key enablers of distributed LLMs, such as tensor and pipeline parallelism, parameter sharding, and system-level optimizations including ZeRO, DeepSpeed, and MoE-based sparsity. These mechanisms not only enable scaling but introduce new problems related to synchronization, memory efficiency, and communication bottlenecks [104].

We then examined the rise of multimodal LLMs (MLLMs), emphasizing how models like Flamingo, GPT-4V, and CLIP have expanded the representational space of LLMs to include vision, audio, and structured data. These models depend on alignment losses, contrastive learning, and cross-attention mechanisms to bridge semantic gaps between modalities. Yet, they also face limitations in modality coverage, sample efficiency, and grounded reasoning.

Throughout the paper, we have formalized the major challenges—ranging from scalability and alignment to evaluation and ethical risks. These are not merely engineering limitations but are fundamentally tied to the representational and algorithmic foundations of current models. Problems such as distributional bias, hallucination, and task generalization continue to constrain real-world deployment.

Looking ahead, we outlined a vision for future research, identifying six forward-looking trajectories: memory-augmented architectures, modular composition, self-alignment mechanisms, continual learning systems, general-purpose agents, and roadmap-based capability progression. We believe that solving these problems will require bridging deep learning with symbolic reasoning, causality, human-computer interaction, and hardware-aware systems design.

As LLMs continue to scale in size and scope, their impact will extend beyond language to become foundational tools across science, medicine, education, and creativity. But this promise comes with responsibility: to ensure that such models are robust, fair, transparent, and beneficial to all. This survey aims to serve as both a technical resource and a research agenda—charting the terrain from today's distributed and multimodal LLMs to tomorrow's intelligent, adaptive, and grounded foundation agents.

References

1. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* **2024**.
2. Jie, C.T.; Chen, L.; Xuesong, Y.; Xudong, Z.; Bo, F. FedPEAT: Convergence of Federated Learning, Parameter-Efficient Fine Tuning, and Emulator Assisted Tuning for AI Foundation Models with Mobile Edge Computing. *arXiv preprint arXiv:2310.17491* **2023**.
3. Chen, J.; Li, W.; Yang, G.; Qiu, X.; Guo, S. Federated Learning Meets Edge Computing: A Hierarchical Aggregation Mechanism for Mobile Devices. In Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications. Springer, 2022, pp. 456–467.
4. Abrahamyan, L.; Chen, Y.; Bekoulis, G.; Deligiannis, N. Learned gradient compression for distributed deep learning. *IEEE Transactions on Neural Networks and Learning Systems* **2021**, *33*, 7330–7344.

5. Ren, E. Task Scheduling for Decentralized LLM Serving in Heterogeneous Networks. *Technical Report No. UCB/EECS-2024-111* **2024**.
6. Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Gu, H.; Shen, T.; Qin, C.; Zhu, C.; Zhu, H.; Liu, Q.; et al. A survey on large language models for recommendation. *World Wide Web* **2024**, *27*, 60.
7. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* **2019**.
8. Fahad, M.; Shojafar, M.; Abbas, M.; Ahmed, I.; Ijaz, H. A multi-queue priority-based task scheduling algorithm in fog computing environment. *Concurrency and Computation: Practice and Experience* **2022**, *34*, e7376.
9. Shen, T.; Li, Z.; Zhao, Z.; Zhu, D.; Lv, Z.; Zhang, S.; Kuang, K.; Wu, F. An Adaptive Aggregation Method for Federated Learning via Meta Controller. In Proceedings of the Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops, 2024, pp. 1–1.
10. Zhao, D. FRAG: Toward Federated Vector Database Management for Collaborative and Secure Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.13272* **2024**.
11. Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794* **2020**.
12. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* **2023**.
13. Brants, T.; Xu, P. Distributed language models. In Proceedings of the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts, 2009, pp. 3–4.
14. Brown, T.; et al. Language Models Are Few-Shot Learners. In Proceedings of the NeurIPS, 2020.
15. Chen, J.; Guo, H.; Yi, K.; Li, B.; Elhoseiny, M. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18030–18040.
16. Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; Stojnic, R. GALACTICA: A Large Language Model for Science. 2022.
17. Alayrac, J.B.; Recasens, A.; Schneider, R.; et al. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198* **2022**.
18. Sun, Y.; Li, Z.; Li, Y.; Ding, B. Improving loRA in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313* **2024**.
19. OpenAI. GPT-4V System Card, 2023. Accessed: 2024-10-29.
20. Shen, X.; Kong, Z.; Yang, C.; Han, Z.; Lu, L.; Dong, P.; Lyu, C.; Li, C.h.; Guo, X.; Shu, Z.; et al. EdgeQAT: Entropy and Distribution Guided Quantization-Aware Training for the Acceleration of Lightweight LLMs on the Edge. *arXiv preprint arXiv:2402.10787* **2024**.
21. Wei, J.; et al. Finetuned Language Models Are Zero-Shot Learners. *ArXiv:2109.01652* **2021**.
22. Ghiasvand, S.; Yang, Y.; Xue, Z.; Alizadeh, M.; Zhang, Z.; Pedarsani, R. Communication-Efficient and Tensorized Federated Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2410.13097* **2024**.
23. Guo, C.; Cheng, F.; Du, Z.; Kiessling, J.; Ku, J.; Li, S.; Li, Z.; Ma, M.; Molom-Ochir, T.; Morris, B.; et al. A Survey: Collaborative Hardware and Software Design in the Era of Large Language Models. *arXiv preprint arXiv:2410.07265* **2024**.
24. Qu, G.; Chen, Q.; Wei, W.; Lin, Z.; Chen, X.; Huang, K. Mobile edge intelligence for large language models: A contemporary survey. *arXiv preprint arXiv:2407.18921* **2024**.
25. Markov, I.; Vladu, A.; Guo, Q.; Alistarh, D. Quantized distributed training of large models with convergence guarantees. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 24020–24044.
26. Xing, J.; Liu, J.; Wang, J.; Sun, L.; Chen, X.; Gu, X.; Wang, Y. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. *Computers & Graphics* **2024**, *119*, 103885.
27. Zniyed, Y.; Nguyen, T.P.; et al. Efficient tensor decomposition-based filter pruning. *Neural Networks* **2024**, *178*, 106393.
28. Zhao, J.; Zhang, Z.; Chen, B.; Wang, Z.; Anandkumar, A.; Tian, Y. Galore: Memory-efficient LLM training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507* **2024**.
29. Yao, Y.; Jin, H.; Shah, A.D.; Han, S.; Hu, Z.; Ran, Y.; Stripelis, D.; Xu, Z.; Avestimehr, S.; He, C. ScaleLLM: A Resource-Frugal LLM Serving Framework by Optimizing End-to-End Efficiency. *arXiv preprint arXiv:2408.00008* **2024**.

30. Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.W.; Galley, M.; Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255* **2023**.
31. Javaheripi, M.; Bubeck, S.; Abdin, M.; Aneja, J.; Bubeck, S.; Mendes, C.C.T.; Chen, W.; Del Giorno, A.; Eldan, R.; Gopi, S.; et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog* **2023**, *1*, 3.
32. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* **2021**.
33. Li, J.; Han, B.; Li, S.; Wang, X.; Li, J. CoLLM: A Collaborative LLM Inference Framework for Resource-Constrained Devices. In Proceedings of the 2024 IEEE/CIC International Conference on Communications in China (ICCC). IEEE, 2024, pp. 185–190.
34. Lee, T.; Tu, H.; Wong, C.H.; Zheng, W.; Zhou, Y.; Mai, Y.; Roberts, J.; Yasunaga, M.; Yao, H.; Xie, C.; et al. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems* **2024**, *37*, 140632–140666.
35. Wang, Z.; Xu, H.; Liu, J.; Huang, H.; Qiao, C.; Zhao, Y. Resource-efficient federated learning with hierarchical aggregation in edge computing. In Proceedings of the IEEE INFOCOM 2021-IEEE conference on computer communications. IEEE, 2021, pp. 1–10.
36. Ben Melech Stan, G.; Aflalo, E.; Rohekar, R.Y.; Bhiwandiwalla, A.; Tseng, S.Y.; Olson, M.L.; Gurwicz, Y.; Wu, C.; Duan, N.; Lal, V. LVLM-Intrepret: An Interpretability Tool for Large Vision-Language Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8182–8187.
37. Shi, S.; Chu, X.; Cheung, K.C.; See, S. Understanding top-k sparsification in distributed deep learning. *arXiv preprint arXiv:1911.08772* **2019**.
38. Guo, X.; Chen, Y. Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. *arXiv preprint arXiv:2403.04190* **2024**.
39. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **2024**, *36*.
40. Zhao, Y.; Wu, D.; Wang, J. ALISA: Accelerating Large Language Model Inference via Sparsity-Aware KV Caching. *arXiv preprint arXiv:2403.17312* **2024**.
41. Zhou, D.W.; Zhang, Y.; Wang, Y.; Ning, J.; Ye, H.J.; Zhan, D.C.; Liu, Z. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
42. Kwon, M.; Hu, H.; Myers, V.; Karamcheti, S.; Dragan, A.; Sadigh, D. Toward grounded commonsense reasoning. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 5463–5470.
43. The, L.; Barrault, L.; Duquenne, P.A.; Elbayad, M.; Kozhevnikov, A.; Alastruey, B.; Andrews, P.; Coria, M.; Couairon, G.; Costa-jussà, M.R.; et al. Large Concept Models: Language Modeling in a Sentence Representation Space. *arXiv preprint arXiv:2412.08821* **2024**.
44. Liu, B.; Chhaparia, R.; Douillard, A.; Kale, S.; Rusu, A.A.; Shen, J.; Szlam, A.; Ranzato, M. Asynchronous Local-SGD Training for Language Modeling. *arXiv preprint arXiv:2401.09135* **2024**.
45. Kombrink, S.; Mikolov, T.; Karafiát, M.; Burget, L. Recurrent Neural Network Based Language Modeling in Meeting Recognition. In Proceedings of the Interspeech, 2011, Vol. 11, pp. 2877–2880.
46. Sergeev, A.; Del Balso, M. Horovod: fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799* **2018**.
47. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.D.; et al. A survey on multimodal large language models for autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 958–979.
48. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large language models: A survey. *arXiv preprint arXiv:2402.06196* **2024**.
49. Sadeepa, S.; Kavinda, K.; Hashika, E.; Sandeepa, C.; Gamage, T.; Liyanage, M. DisLLM: Distributed LLMs for Privacy Assurance in Resource-Constrained Environments. In Proceedings of the 2024 IEEE Conference on Communications and Network Security (CNS). IEEE, 2024, pp. 1–9.
50. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
51. McMahan, H.B.; Moore, E.; et al. Communication-efficient learning of deep networks from decentralized data. *arXiv:1602.05629* **2016**.

52. Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* **2017**.
53. Xin, J.; Bae, S.; Park, K.; Canini, M.; Hwang, C. Immediate Communication for Distributed AI Tasks. *The 2nd Workshop on Hot Topics in System Infrastructure* **2024**.
54. Huang, K.; Yin, H.; Huang, H.; Gao, W. Towards Green AI in Fine-tuning Large Language Models via Adaptive Backpropagation. *arXiv preprint arXiv:2309.13192* **2023**.
55. Chen, Y.; Zhang, T.; Jiang, X.; Chen, Q.; Gao, C.; Huang, W. Fedbone: Towards large-scale federated multi-task learning. *arXiv preprint arXiv:2306.17465* **2023**.
56. Dean, J.; Ghemawat, S. MapReduce: simplified data processing on large clusters. *Communications of the ACM* **2008**, *51*, 107–113.
57. Ahia, O.; Kumar, S.; Gonen, H.; Hoffman, V.; Limisiewicz, T.; Tsvetkov, Y.; Smith, N.A. MAGNET: Improving the Multilingual Fairness of Language Models with Adaptive Gradient-Based Tokenization. *arXiv preprint arXiv:2407.08818* **2024**.
58. Gao, R.; Oh, T.H.; Grauman, K.; Torresani, L. Listen to look: Action recognition by previewing audio. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10457–10467.
59. Wu, W.; Sun, Z.; Song, Y.; Wang, J.; Ouyang, W. Transferring vision-language models for visual recognition: A classifier perspective. *International Journal of Computer Vision* **2024**, *132*, 392–409.
60. Yanghe, P.; Jun, C.; Linjun, D.; Xiaobo, Z.; Hongyan, Z. Cloud-Edge Collaborative Large Model Services: Challenges and Solutions. *IEEE Network* **2024**.
61. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International conference on machine learning. PMLR, 2022, pp. 12888–12900.
62. Pisarchyk, Y.; Lee, J. Efficient memory management for deep neural net inference. *arXiv preprint arXiv:2001.03288* **2020**.
63. Wang, H.; Ma, S.; Dong, L.; Huang, S.; Wang, H.; Ma, L.; Yang, F.; Wang, R.; Wu, Y.; Wei, F. Bitnet: Scaling 1-bit transformers for large language models. *arXiv preprint arXiv:2310.11453* **2023**.
64. Christiano, P.F.; Leike, J.; Brown, T.B.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. In Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 4299–4307.
65. Bitton Guetta, N.; Slobodkin, A.; Maimon, A.; Habba, E.; Rassin, R.; Bitton, Y.; Szpektor, I.; Globerson, A.; Elovici, Y. Visual riddles: A commonsense and world knowledge challenge for large vision and language models. *Advances in Neural Information Processing Systems* **2024**, *37*, 139561–139588.
66. Biderman, S.; Schoelkopf, H.; Anthony, Q.G.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M.A.; Purohit, S.; Prashanth, U.S.; Raff, E.; et al. Pythia: A suite for analyzing large language models across training and scaling. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 2397–2430.
67. Wu, H.; Li, X.; Zhang, D.; Xu, X.; Wu, J.; Zhao, P.; Liu, Z. CG-FedLLM: How to Compress Gradients in Federated Fune-tuning for Large Language Models. *arXiv preprint arXiv:2405.13746* **2024**.
68. Rasley, J.; Rajbhandari, S.; Ruwase, O.; Yang, S.; Zhang, Y.; He, Y. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *arXiv preprint arXiv:2007.00072* **2020**.
69. Zniyed, Y.; Nguyen, T.P.; et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
70. Laurençon, H.; Tronchon, L.; Cord, M.; Sanh, V. What matters when building vision-language models? *Advances in Neural Information Processing Systems* **2024**, *37*, 87874–87907.
71. Denton, E.L.; Zaremba, W.; Bruna, J.; LeCun, Y.; Fergus, R. Exploiting linear structure within convolutional networks for efficient evaluation. *Advances in neural information processing systems* **2014**, *27*.
72. Huang, J.; Zhang, Z.; Zheng, S.; Qin, F.; Wang, Y. DISTMM: Accelerating Distributed Multimodal Model Training. In Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24), 2024, pp. 1157–1171.
73. Bai, Y.; Zhang, Y.; Yang, J.; Liu, J.; Tang, J.; Wu, J.; Gao, J.; Wang, J. BinaryBERT: Pushing the limit of BERT quantization. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4334–4343.
74. Zheng, J.; Zhang, H.; Wang, L.; Qiu, W.; Zheng, H.; Zheng, Z. Safely Learning with Private Data: A Federated Learning Framework for Large Language Model. *arXiv preprint arXiv:2406.14898* **2024**.

75. Liu, D.; Yang, M.; Qu, X.; Zhou, P.; Cheng, Y.; Hu, W. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403* **2024**.
76. Jelinek, F. *Statistical methods for speech recognition*; MIT press, 1998.
77. Zhu, K.; Li, S.; Zhang, X.; Wang, J.; Xie, C.; Wu, F.; Xie, R. An Energy-Efficient Dynamic Offloading Algorithm for Edge Computing Based on Deep Reinforcement Learning. *IEEE Access* **2024**.
78. Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; Kiela, D. Flava: A foundational language and vision alignment model. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15638–15650.
79. Tan, A.Z.; Yu, H.; Cui, L.; Yang, Q. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems* **2022**, *34*, 9587–9603.
80. Popov, V.; Kudinov, M.; Piontkovskaya, I.; Vytovtov, P.; Nevidomsky, A. Distributed fine-tuning of language models on private data. In Proceedings of the International Conference on Learning Representations, 2018.
81. Youyang, Q.; Jinwen, Z.; Qi, C. Federated Learning driven Large Language Models for Swarm Intelligence: A Survey. *arXiv preprint arXiv:2406.09831* **2024**.
82. Zhang, Z.; Cai, D.; Zhang, Y.; Xu, M.; Wang, S.; Zhou, A. FedRDMA: Communication-Efficient Cross-Silo Federated LLM via Chunked RDMA Transmission. In Proceedings of the Proceedings of the 4th Workshop on Machine Learning and Systems, 2024, pp. 126–133.
83. Fan, T.; Kang, Y.; Ma, G.; Chen, W.; Wei, W.; Fan, L.; Yang, Q. Fate-LLM: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049* **2023**.
84. Shenghui, L.; Wei, L.; Lin, W. Synergizing Foundation Models and Federated Learning: A Survey. *arXiv preprint arXiv:2406.12844* **2024**.
85. Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; Zhou, D. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984* **2020**.
86. Nguyen, C.V.; Shen, X.; Aponte, R.; Xia, Y.; et al. A Survey of Small Language Models. *arXiv preprint arXiv:2410.20011* **2024**.
87. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
88. Li, D.; Shao, R.; Xie, A.; Xing, E.P.; Ma, X.; Stoica, I.; Gonzalez, J.E.; Zhang, H. DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training. In Proceedings of the First Conference on Language Modeling, 2024.
89. Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* **2020**, *1*.
90. Rui, Y.; Jingyi, C.; Dihan, L.; Wenhao, W.; Yaxin, D.; Yanfeng, W.; Siheng, C. FedLLM-Bench: Realistic Benchmarks for Federated Learning of Large Language Models. *arXiv preprint arXiv:2406.04845* **2024**.
91. Markov, I.; Vladu, A.; Guo, Q.; Alistarh, D. Quantized distributed training of large models with convergence guarantees. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 24020–24044.
92. Zhang, M.; Arora, S.; Chalamala, R.; Wu, A.; Spector, B.; Singhal, A.; Ramesh, K.; Ré, C. LoLCATs: On Low-Rank Linearizing of Large Language Models. *arXiv preprint arXiv:2410.10254* **2024**.
93. Borzunov, A.; Baranchuk, D.; Dettmers, T.; Ryabinin, M.; Belkada, Y.; Chumachenko, A.; Samygin, P.; Raffel, C. Petals: Collaborative inference and fine-tuning of large models. *arXiv preprint arXiv:2209.01188* **2022**.
94. Hu, B.; Li, J.; Xu, L.; Lee, M.; Jajoo, A.; Kim, G.W.; Xu, H.; Akella, A. Blockllm: Multi-tenant finer-grained serving for large language models. *arXiv preprint arXiv:2404.18322* **2024**.
95. Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* **2024**.
96. Fu, T.; Huang, H.; Ning, X.; Zhang, G.; Chen, B.; Wu, T.; Wang, H.; Huang, Z.; Li, S.; Yan, S.; et al. Moa: Mixture of sparse attention for automatic large language model compression. *arXiv preprint arXiv:2406.14909* **2024**.
97. Yang, Z.; Yang, Y.; Zhao, C.; Guo, Q.; He, W.; Ji, W. PerLLM: Personalized Inference Scheduling with Edge-Cloud Collaboration for Diverse LLM Services. *arXiv preprint arXiv:2405.14636* **2024**.
98. Li, Y.; Li, M.; Zhang, X.; Xu, G.; Chen, F.; Yuan, Y.; Zou, Y.; Zhao, M.; Lu, J.; Yu, D. Unity is Power: Semi-Asynchronous Collaborative Training of Large-Scale Models with Structured Pruning in Resource-Limited Clients. *arXiv preprint arXiv:2410.08457* **2024**.

99. Ali, S.S.; Ali, M.; Bhatti, D.M.S.; Choi, B.J. dy-TACFL: Dynamic Temporal Adaptive Clustered Federated Learning for Heterogeneous Clients. *Electronics* **2025**, *14*, 152.
100. Albanie, S.; Nagrani, A.; Vedaldi, A.; Zisserman, A. Emotion recognition in speech using cross-modal transfer in the wild. In Proceedings of the Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 292–301.
101. Bai, J.; Chen, D.; Qian, B.; Yao, L.; Li, Y. Federated fine-tuning of large language models under heterogeneous language tasks and client resources. *arXiv e-prints* **2024**, pp. arXiv–2402.
102. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* **2021**.
103. Lu, P.; Gong, R.; Jiang, S.; Qiu, L.; Huang, S.; Liang, X.; Zhu, S.C. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165* **2021**.
104. Xu, J.; Li, Z.; Chen, W.; Wang, Q.; Gao, X.; Cai, Q.; Ling, Z. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088* **2024**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.