Article

Not peer-reviewed version

A Comparative Survey of Large Language Models: Foundation, Instruction-Tuned, and Multimodal Variants

Owen Graham and Jim Balford *

Posted Date: 13 June 2025

doi: 10.20944/preprints202506.1134.v1

Keywords: LLM

 doi:10.20944/preprints202506.1134.v1

Article

A Comparative Survey of Large Language Models: Foundation, Instruction-Tuned, and Multimodal Variants

Owen Graham and Jim Balford *

* Correspondence: balfordjim@mail.org

Abstract: The rapid evolution of large language models (LLMs) has transformed natural language processing, enabling machines to perform complex language understanding, generation, and reasoning tasks with unprecedented fluency and adaptability. This survey presents a comprehensive comparative analysis of three major classes of LLMs: foundation models, instruction-tuned models, and multimodal variants. We first define and contextualize each category—foundation models as the general-purpose pretrained backbones, instruction-tuned models as task-optimized derivatives guided by human or synthetic instructions, and multimodal models as those extending language understanding to vision, audio, and other modalities. The paper examines architectural innovations, training methodologies, benchmark performances, and real-world applications across these model types. Through systematic comparison, we highlight the trade-offs in generality, alignment, efficiency, and modality integration. We further discuss deployment trends, ethical considerations, and emerging challenges, offering insights into the future trajectory of unified, scalable, and human-aligned language models. This survey aims to serve researchers and practitioners by clarifying the landscape and guiding informed decisions in the design and application of LLMs.

Keywords: LLM

1. Introduction

In recent years, Large Language Models (LLMs) have emerged as a cornerstone of modern artificial intelligence, significantly advancing capabilities in natural language understanding, generation, translation, summarization, and more. Models such as GPT, BERT, and T5 have demonstrated an unprecedented ability to learn from vast corpora of text and generalize across a wide variety of language tasks. As these models grow in scale and complexity, so too does the diversity in their design, training strategies, and application domains.

This growing ecosystem of LLMs can be broadly categorized into three major types: **foundation models**, **instruction-tuned models**, and **multimodal models**. **Foundation models** are large-scale pretrained models trained on massive text datasets using unsupervised or self-supervised learning objectives. These models serve as general-purpose engines capable of performing downstream tasks with minimal fine-tuning. Building on these, **instruction-tuned models** are adapted using human-written or machine-generated instructions to better align with human intent and improve task-specific performance. More recently, **multimodal language models** have been introduced, capable of processing and integrating multiple input modalities such as text, images, audio, and video, thus extending the capabilities of LLMs beyond pure language tasks.

While each of these classes has demonstrated remarkable progress, they differ significantly in terms of architecture, training paradigms, scalability, performance, and real-world applicability. As such, there is a growing need for a systematic and comparative analysis that examines the strengths, limitations, and evolution of these variants. Understanding these distinctions is essential for researchers aiming to advance the field, practitioners seeking to deploy models in specific domains, and policymakers grappling with the implications of increasingly powerful AI systems.

 doi:10.20944/preprints202506.1134.v1

This survey provides a comprehensive comparative overview of foundation, instruction-tuned, and multimodal LLMs. We explore their design principles, core functionalities, training strategies, and benchmark performance. We also discuss critical deployment issues including alignment, efficiency, accessibility, and ethical challenges. By offering a structured taxonomy and in-depth comparison, this paper aims to clarify the current landscape and illuminate promising directions for future research and development.

The remainder of this paper is organized as follows: Section 2 introduces a taxonomy for categorizing LLMs. Section 3 reviews foundation models. Section 4 covers instruction-tuned models, and Section 5 delves into multimodal models. Section 6 presents a comparative analysis, followed by discussions on deployment trends in Section 7. Section 8 outlines challenges and open questions, and Section 9 explores future directions. We conclude in Section 10 with final reflections on the evolving role of LLMs.

1.1. Background on Large Language Models (LLMs)

Large Language Models (LLMs) represent a major leap in the field of artificial intelligence, particularly within natural language processing (NLP). Built on the transformer architecture introduced by Vaswani et al. (2017), LLMs leverage self-attention mechanisms to capture complex patterns and long-range dependencies in text data. These models are trained on massive corpora, often comprising hundreds of billions to trillions of tokens, enabling them to acquire rich linguistic, factual, and even reasoning capabilities through scale and generalization.

Early milestones in LLM development, such as **BERT** (Devlin et al., 2018) and **GPT-2** (Radford et al., 2019), showcased the efficacy of pretraining on large text datasets followed by fine-tuning on specific tasks. This paradigm shift led to models that could generalize across a wide range of NLP tasks with minimal task-specific architecture changes. Subsequent models like **GPT-3**, **T5**, **PaLM**, and **Chinchilla** pushed the limits of model size, training data, and performance, fueling a wave of research and commercial adoption.

At their core, LLMs are statistical pattern learners that estimate the probability distribution over sequences of tokens, enabling them to generate coherent and contextually appropriate outputs. They are trained using objectives such as next-token prediction (causal language modeling) or masked token prediction (masked language modeling), depending on the architecture. As a result, LLMs have become foundational engines for a variety of downstream applications, including chatbots, code generation tools, search engines, medical assistants, and creative writing systems.

However, the raw capabilities of LLMs are not without limitations. Foundation models, despite their generality, often struggle with alignment, hallucination, and following user intent. These issues have motivated the development of instruction-tuned models and alignment techniques such as supervised fine-tuning and reinforcement learning with human feedback (RLHF). In parallel, the scope of language modeling has expanded to incorporate multimodal inputs, enabling models to understand and reason over both text and other data types such as images, audio, and video.

Thus, LLMs today are no longer a monolithic concept but rather a family of evolving architectures and techniques that reflect diverse goals—from general-purpose understanding to task-specific instruction following and multimodal reasoning. This complexity underpins the need for a comparative survey that maps out the distinctions and interconnections among the foundational, instruction-tuned, and multimodal LLM paradigms.

1.2. Motivation for Comparative Analysis

The proliferation of large language models (LLMs) over the past few years has led to the emergence of a wide array of architectures, training paradigms, and application domains. While early LLMs primarily served as general-purpose pretrained models, the landscape has since diversified into specialized categories such as **instruction-tuned** and **multimodal** variants. Each of these model classes embodies different design choices and optimization goals, resulting in varying performance characteristics, strengths, and trade-offs.

 doi:10.20944/preprints202506.1134.v1

Despite their shared foundation in transformer-based architectures, **foundation models**, **instruction-tuned models**, and **multimodal models** differ substantially in their interaction with end users, their alignment with human intent, and their applicability to complex real-world tasks. For example, foundation models like GPT-3 or PaLM demonstrate strong generative capabilities but may struggle with task specificity or safe alignment. Instruction-tuned models, by contrast, are explicitly trained to follow human directives, improving usability and control. Multimodal models further extend capabilities by integrating visual and auditory data, unlocking new domains such as image captioning, visual question answering, and cross-modal retrieval.

With the increasing availability of both open-source and proprietary models, researchers, developers, and stakeholders face a growing need to **understand the comparative merits and limitations** of each type of LLM. Without a structured framework or comprehensive comparison, model selection becomes ad hoc, and reproducibility or scalability across use cases becomes challenging. Moreover, the rapid pace of innovation in this space has outpaced the consolidation of best practices and clear evaluation standards.

A comparative analysis is therefore essential for multiple reasons:

- **Clarity and Taxonomy**: To provide a structured understanding of how different classes of LLMs are defined, trained, and deployed.
- **Informed Decision-Making**: To guide researchers and practitioners in choosing the right model class for specific applications, domains, and constraints.
- **Performance and Trade-Offs**: To systematically assess how foundation, instruction-tuned, and multimodal models differ in terms of performance, generalization, interpretability, alignment, and cost.
- **Bridging Gaps**: To identify where models can be improved by integrating insights across paradigms—such as combining the generality of foundation models with the usability of instruction-tuned systems and the perceptual capabilities of multimodal models.
- **Foresight for Development**: To anticipate trends and guide the next generation of models toward unified, context-aware, and human-aligned architectures.

By conducting a focused comparative survey, this paper aims to map the evolving terrain of LLM development, offering both a reference framework and a forward-looking perspective on the trajectories shaping this rapidly advancing field.

2. Taxonomy of Large Language Models

Large Language Models (LLMs) can no longer be viewed as a monolithic class of architectures or capabilities. As the field has matured, LLMs have diversified significantly in terms of their structure, training objectives, application domains, and interaction modalities. To facilitate systematic analysis and comparison, this section presents a taxonomy that categorizes LLMs along three key dimensions: architectural design, purpose of deployment, and evolutionary progression. This taxonomy serves as the foundation for understanding the distinctions among foundation models, instruction-tuned models, and multimodal models.

2.1. Classification by Architecture

LLMs differ fundamentally in how they are architected. The primary distinction arises from the configuration of the Transformer architecture, which includes three major types:

- **Decoder-Only Models**:
  These models predict the next token in a sequence, making them ideal for generative tasks. Examples include GPT-2, GPT-3, and GPT-4.
- **Encoder-Only Models**:

 doi:10.20944/preprints202506.1134.v1

These models encode the entire input sequence for tasks like classification or embedding generation. BERT and RoBERTa are canonical encoder-only models.

- **Encoder-Decoder (Seq2Seq) Models**:
  These models encode an input sequence and decode an output sequence, making them suitable for translation and summarization. Examples include T5, BART, and FLAN-T5.

Hybrid architectures (e.g., Perceiver, RETRO) and **efficient transformers** (e.g., Longformer, Performer) are also emerging to address scalability and long-context understanding.

2.2. Classification by Purpose

This axis categorizes LLMs based on their intended role and functional tuning:

- **Foundation Models**:
  These are large pretrained models designed to serve as general-purpose backbones across a range of downstream tasks. They are typically trained on vast, diverse datasets using unsupervised or self-supervised learning. Examples: GPT-3, PaLM, Chinchilla, LLaMA.
- **Instruction-Tuned Models**:
  These are derived from foundation models via additional training on datasets consisting of prompts and expected responses. They aim to improve alignment with user intent and are better at following natural language instructions. Examples: InstructGPT, FLAN-T5, OpenAssistant, Mistral-Instruct.
- **Multimodal Models**:
  These models extend language understanding to other modalities such as vision, audio, and video. They can accept multiple input types and generate text or multimodal outputs. Examples: GPT-4V, Gemini, CLIP, Flamingo, Kosmos-2.

2.3. Evolution of LLMs Over Time

The development of LLMs can be viewed as progressing through three overlapping phases:

- **Phase 1: General-Purpose Pretraining**
  Emphasis on scale and generality. Key contributions: GPT-2, BERT, T5.
- **Phase 2: Task Alignment and Instruction Following**
  Models are refined to better align with human intent through supervised fine-tuning, reinforcement learning, and curated datasets. Key contributions: InstructGPT, FLAN, Alpaca.
- **Phase 3: Multimodal and Unified Models**
  Language models are extended to process other modalities, enabling capabilities like image generation, audio captioning, and video analysis. Key contributions: Flamingo, GPT-4V, Gemini, MM-ReAct.

This evolutionary trajectory reflects the field's shift from pure scale to **capability**, **alignment**, and **multimodal understanding**.

3. Foundation Language Models

Foundation language models form the base layer of the large language model ecosystem. These models are pretrained on large-scale unlabeled datasets using self-supervised learning objectives and are designed to serve as general-purpose models that can be adapted for a wide range of downstream

 doi:10.20944/preprints202506.1134.v1

tasks. Their strength lies in their ability to acquire broad linguistic, factual, and contextual knowledge from massive corpora without task-specific supervision.

3.1. Definition and Characteristics

Foundation models are defined by their role as **pretrained general-purpose models** that serve as the "foundations" upon which more specialized or fine-tuned models are built. Their characteristics include:

- **Massive Pretraining Datasets**: Typically trained on hundreds of billions of tokens sourced from books, web data, Wikipedia, code, and other heterogeneous sources.
- **Self-Supervised Objectives**: Common objectives include next-token prediction (causal language modeling) and masked language modeling (MLM).
- **Zero-Shot and Few-Shot Learning Capabilities**: Demonstrated ability to generalize to unseen tasks with little or no additional fine-tuning.
- **Scalability**: Performance improves significantly with model and dataset scale, as demonstrated by scaling laws (Kaplan et al., 2020).

3.2. Major Examples

Model	Developer	Architecture	Parameters	Training Data Size	Training Objective
GPT-3	OpenAI	Decoder-only	175B	300B tokens	Causal LM
PaLM	Google	Decoder-only	540B	780B tokens	Causal LM
Chinchilla	DeepMind	Decoder-only	70B	1.4T tokens	Causal LM
BERT	Google	Encoder-only	340M	Wikipedia + Books	MLM
T5	Google	Encoder-decoder	11B	C4 (Colossal Clean Crawled Corpus)	Text-to-text

These models are typically not aligned with user instructions by default and may require additional fine-tuning for safe and controllable outputs.

3.3. Training Data and Objectives

Foundation models rely heavily on the quality and diversity of their training data. The training corpora often include:

- **Open-access web data** (e.g., Common Crawl)
- **Digitized books and academic content**
- **Code repositories**
- **Conversational and forum data**

The two dominant training objectives are:

- **Causal Language Modeling** (CLM): Predicting the next token based on previous context (e.g., GPT-style models).
- **Masked Language Modeling** (MLM): Predicting randomly masked tokens within a sequence (e.g., BERT, RoBERTa).

Recent approaches have also explored denoising objectives (T5), permutation-based models (XLNet), and retrieval-augmented training (RETRO).

 doi:10.20944/preprints202506.1134.v1

3.4. Strengths and Limitations

Strengths:

- Broad generalization across language tasks
- High performance in zero/few-shot settings
- Scalable and reusable across domains
- Strong latent knowledge representation

Limitations:

- Lack of alignment with human intent or task-specific goals
- Tendency to produce hallucinations or factually incorrect content
- Computationally expensive to train and deploy
- Inability to handle multimodal input or grounded reasoning without adaptation

Suggested Table:

Comparison of Major Foundation Models

Model	Year	Parameters	Training Corpus	Strengths	Weaknesses
GPT-3	2020	175B	Web Text, Books, etc.	Strong few-shot ability	Expensive inference, hallucination
BERT	2018	340M	Books Corpus, Wiki	Bidirectional understanding	Not generative
PaLM	2022	540B	Diverse, filtered web	Very high performance	High energy/resource cost
Chinchilla	2022	70B	1.4T tokens	Efficient scaling	Limited instruction tuning

4. Instruction-Tuned Language Models

Instruction-tuned language models represent a significant advancement in aligning large language models (LLMs) with human intent. Building upon general-purpose foundation models, instruction-tuned models are refined through additional training on datasets that pair natural language instructions with desired outputs. This additional phase helps the model better understand and follow human prompts, improving controllability, safety, and performance on specific tasks, especially in few-shot or zero-shot scenarios.

4.1. Definition and Purpose

Instruction-tuned models are derived from pretrained foundation models and fine-tuned on curated datasets containing instructions and corresponding outputs. The primary purpose is to bridge the gap between raw generative capability and user-aligned behavior by teaching models to generalize from explicit directives.

Key goals include:

- **Improving alignment with user instructions**
- **Enhancing usability without task-specific fine-tuning**
- **Mitigating hallucinations and unsafe outputs**
- **Facilitating natural human-computer interaction**

4.2. Training Methods and Datasets

 doi:10.20944/preprints202506.1134.v1

The instruction-tuning pipeline typically includes:

1. **Supervised Fine-Tuning (SFT)**:
    - Models are fine-tuned on large datasets of <instruction, input, output> triplets.
    - Example datasets: **FLAN**, **Super-Natural Instructions**, **OpenAssistant**, **Alpaca**, **ShareGPT**.
2. **Reinforcement Learning with Human Feedback (RLHF)**:
    - Models are trained to prefer outputs that align with human preferences.
    - Steps include reward modeling and policy optimization (e.g., PPO).
    - Used in models like **InstructGPT**, **ChatGPT**, **Claude**.
3. **Self-Instruct and Synthetic Generation**:
    - Models bootstrap additional instruction data from themselves or other LLMs.
    - Example: **Self-Instruct** (Wang et al., 2022).

4.3. Notable Instruction-Tuned Models

Model	Base Model	Developer	Tuning Method	Key Features
InstructGPT	GPT-3	OpenAI	SFT + RLHF	First widely adopted instruction-tuned model
ChatGPT	GPT-3.5 / GPT-4	OpenAI	SFT + RLHF + Conversation	Dialogue-optimized, real-time responsiveness
FLAN-T5	T5	Google	SFT on diverse tasks	Strong zero-shot and generalization ability
Alpaca	LLaMA	Stanford	SFT on GPT-generated data	Lightweight, open-source instructional tuning
Open Assistant	LLaMA	LAION	Community-sourced SFT	Open RLHF pipeline
Claude	Proprietary	Anthropic	Constitutional AI + RLHF	Focus on safe, steerable behavior

4.4. Capabilities and Advantages

Instruction-tuned models offer several advantages over their foundation counterparts:

- **Better Prompt Following**: Clear understanding and execution of user directives.
- **Improved Generalization**: Effective zero- and few-shot performance across unseen tasks.
- **Enhanced Safety and Usefulness**: RLHF and constitutional training reduce harmful or biased outputs.
- **User-Friendly Interaction**: More suitable for deployment in assistants, tutors, search engines, and chatbots.

4.5. Limitations and Challenges

Despite their benefits, instruction-tuned models face key limitations:

- **Instruction Sensitivity**: Performance varies significantly depending on prompt phrasing.
- **Bias Amplification**: Human preferences and training data biases can propagate.

 doi:10.20944/preprints202506.1134.v1

- **High Resource Requirements**: Tuning with human feedback is costly and time-consuming.
- **Lack of Grounded Knowledge**: Still prone to hallucinations without retrieval mechanisms or external tools.

5. Multimodal Language Models

Multimodal language models (MLLMs) extend the capabilities of large language models beyond text, enabling them to process and generate content across multiple modalities such as **images**, **audio**, **video**, and **structured data**. These models represent a critical step toward building general-purpose AI systems capable of **perception**, **reasoning**, and **interaction** in real-world environments where information is inherently multimodal.

5.1. Definition and Scope

Multimodal language models integrate inputs from more than one modality (e.g., vision + language) and produce outputs in one or multiple modalities. They are built upon the same transformer-based foundations as text-only models but are architecturally augmented to handle visual embeddings, speech signals, or other structured inputs.

Scope includes:

- **Vision-Language Models (VLMs)**: Image captioning, visual question answering, and image-grounded dialogue (e.g., GPT-4V, Flamingo).
- **Speech-Language Models**: Automatic speech recognition (ASR), speech synthesis, and spoken question answering (e.g., Whisper, AudioLM).
- **Cross-Modal Reasoning**: Tasks like referring expression comprehension, video-language alignment, and embodied AI.

5.2. Architectural Approaches

There are several common design strategies for integrating multimodal information into LLMs:

- **Dual-Encoder Models**:
  Separate encoders process each modality, and embeddings are aligned in a joint representation space. Example: CLIP (Contrastive Language–Image Pretraining).
- **Fusion Models**:
  Modalities are combined in a shared transformer architecture using early, mid, or late fusion. Example: Flamingo, Kosmos-2.
- **Adapter-Based Integration**:
  Lightweight adapters are added to pretrained models to process specific modalities without retraining the entire network. Example: PaLI-X, LLaVA.
- **Multimodal Prompting**:
  Inputs from non-text modalities are transformed into "prompts" or embeddings that can be consumed by text-based models (e.g., visual tokens or audio spectrogram embeddings).

5.3. Prominent Multimodal LLMs

Model	Modalities	Developer	Architecture Type	Core Capabilities
GPT-4V (Vision)	Text + Image	OpenAI	Unified Transformer	Image understanding, document reasoning

 doi:10.20944/preprints202506.1134.v1

Model	Modalities	Developer	Architecture Type	Core Capabilities
Gemini 1.5	Text + Image + Audio	Google DeepMind	Multimodal transformer	State-of-the-art multi-input reasoning
CLIP	Image + Text	OpenAI	Dual Encoder	Visual search, zero-shot classification
Flamingo	Text + Image	DeepMind	Fusion Transformer	Image-grounded dialogue, captioning
Kosmos-2	Text + Image + OCR	Microsoft	Multimodal Transformer	Vision-language grounding and reasoning
LLaVA	Text + Image	UC Berkeley	Adapter + Vision Encoder	Open-ended VQA, visual chat

5.4. Use Cases and Applications

Multimodal LLMs are rapidly finding application across domains:

- **Healthcare**: Radiology report generation from medical images, cross-modal diagnostics
- **Education**: Visual tutoring, diagram explanation, language-learning with image/audio context
- **Search and Retrieval**: Multimodal search engines (e.g., Google Lens + Gemini)
- **Accessibility**: Image and scene description for visually impaired users
- **Creative Tools**: Image captioning, visual storytelling, audio-based content generation

5.5. Strengths and Limitations

Strengths:

- Rich, grounded understanding of the physical world
- Improved performance on real-world tasks involving diverse inputs
- Enables more human-like interaction (e.g., talking about images or sounds)

Limitations:

- Training data for multimodal inputs is less abundant and less standardized
- Computationally more expensive to train and fine-tune
- Still prone to hallucination or false cross-modal associations
- Challenges in aligning and synchronizing multiple input types

Suggested Diagram:
"Architecture of a Multimodal LLM" –
Illustrate a model that:

- Accepts image and text inputs
- Encodes both with respective encoders
- Performs attention fusion in a joint transformer
- Outputs text (e.g., answer to visual question)

6. Comparative Analysis

To understand the evolution and specialization of large language models (LLMs), it is essential to compare **foundation models**, **instruction-tuned models**, and **multimodal models** across several

 doi:10.20944/preprints202506.1134.v1

core dimensions. This comparative analysis highlights their respective architectures, training paradigms, capabilities, performance benchmarks, and application suitability, providing insight into their trade-offs and complementary roles in the LLM ecosystem.

6.1. Comparison Dimensions

This comparison is structured around the following axes:

- **Training Objective & Data**
- **Architecture & Modalities**
- **Task Generalization**
- **Instruction Following**
- **Performance on Benchmarks**
- **Practical Applications**
- **Limitations**

6.2. Comparative Table

Dimension	Foundation Models	Instruction-Tuned Models	Multimodal Models
Training Objective	Self-supervised (e.g., MLM, CLM)	Supervised fine-tuning + RLHF	Multimodal fusion + optionally RLHF
Input Modalities	Text only	Text only	Text + images/audio/video
Output Modality	Text	Text	Text or multimodal outputs
Architecture	Encoder / Decoder / Encoder-Decoder	Based on foundation models	Unified or dual encoders; adapter-based
Instruction Following	Weak (zero-shot prompting)	Strong (fine-tuned on instructions)	Strong (for image-grounded tasks)
Few-shot Learning	Emerging capability	Highly effective	Limited (depends on task)
Zero-shot Performance	Moderate to strong	Strong (especially on unseen tasks)	Variable (strong in retrieval & classification)
Example Models	GPT-3, PaLM, BERT, T5	InstructGPT, ChatGPT, FLAN-T5, Claude	GPT-4V, Gemini, Flamingo, Kosmos-2
Primary Use Cases	Pretraining backbone, embeddings	Chatbots, assistants, instruction interfaces	VQA, captioning, multimodal agents
Limitations	Weak task alignment, hallucinations	Prompt sensitivity, bias amplification	Data scarcity, high compute costs

6.3. Performance on Benchmarks

 doi:10.20944/preprints202506.1134.v1

Benchmark	Foundation Models	Instruction-Tuned Models	Multimodal Models
MMLU (text tasks)	Good (GPT-3, PaLM)	Excellent (FLAN-T5, ChatGPT)	Variable
HELM / BIG-Bench	Limited	Strong generalization	Not applicable
VQAv2 / COCO	Not applicable	Not applicable	Excellent (GPT-4V, Flamingo)
GSM8K (math)	Moderate	Good (Instruct GPT, Claude)	Dependent on task design

6.4. Analysis of Trade-Offs

- **Foundation Models** provide general-purpose capabilities and a scalable base for downstream use but require fine-tuning for task alignment and safety.
- **Instruction-Tuned Models** deliver much better user alignment and usability but inherit biases and errors from the foundation models and the tuning datasets.
- **Multimodal Models** unlock perception and cross-modal reasoning but face greater engineering complexity, limited data availability, and slower inference.

Each model type is optimal for different **use cases**:

- Foundation models excel in **low-resource generalization**.
- Instruction-tuned models are ideal for **interactive applications**.
- Multimodal models are indispensable for **real-world AI agents and grounded understanding**.

6.5. Suggested Diagram

"**LLM Variant Landscape**" – A triangular diagram placing:

- Foundation models at the base (general scope),
- Instruction-tuned models at one corner (aligned behavior),
- Multimodal models at another corner (expanded input modalities),

...showing the trade-offs in scope, control, and complexity.

7. Deployment and Ecosystem Trends

As large language models (LLMs) mature, their deployment strategies and the supporting ecosystems are undergoing rapid transformation. From cloud-based APIs to edge-optimized models, the trend reflects a shift from research prototypes to production-scale systems. This section analyzes deployment architectures, industry adoption patterns, toolchain developments, and emerging practices around the responsible and scalable integration of LLMs.

7.1. Deployment Architectures

Deployment choices depend on model size, latency needs, privacy constraints, and application context. Common strategies include:

- **Cloud-based APIs**:

  Hosted models accessed via RESTful endpoints (e.g., OpenAI API, Gemini API, Claude).

 doi:10.20944/preprints202506.1134.v1

Pros: scalability, model freshness, minimal infrastructure burden.

Cons: dependency, latency, data governance risks.

- **On-premise/Private Hosting**:
  Useful for privacy-critical industries (e.g., legal, healthcare).
  Example: LLaMA or Mistral deployed internally with quantization or inference optimization.
- **Edge Deployment**:
  Lightweight models (e.g., Phi-2, DistilGPT, MobileBERT) adapted for mobile, IoT, and embedded systems.
  Often optimized via pruning, quantization, or knowledge distillation.
- **Hybrid Architectures**:
  Combine local inference with cloud augmentation (e.g., retrieval-augmented generation or tool use).

7.2. Ecosystem Tools and Frameworks

A wide array of tools supports the training, fine-tuning, deployment, and monitoring of LLMs:

Category	**Tools**
Training & Fine-tuning	Hugging Face Transformers, Deep Speed, LoRA, PEFT, Axolotl
Serving & Inference	vLLM, TGI (Text Generation Inference), ONNX Runtime, NVIDIA TensorRT
Evaluation	HELM, BIG-bench, TruthfulQA, RAGAS
RLHF & Alignment	TRL (Transformers Reinforcement Learning), Open Feedback, Reinforcement Studio
Multimodal Frameworks	Open Flamingo, MiniGPT-4, LLaVA, Hugging Face's Diffusers + Transformers
Monitoring & Governance	Prompt Layer, Lang fuse, Arize Phoenix, Weights & Biases

7.3. Industry Adoption Trends

Industry-wide interest in LLMs has catalyzed innovation and adoption across diverse sectors:

- **Customer Support**: AI agents (e.g., ChatGPT-based interfaces, Ada, Intercom) are used for ticket triaging, live assistance, and content summarization.
- **Enterprise Productivity**: Copilots (e.g., GitHub Copilot, Microsoft 365 Copilot) are increasingly integrated into office tools and IDEs.
- **Healthcare**: LLMs support clinical summarization, drug discovery, and medical Q&A (e.g., Med-PaLM, Glass AI).
- **Education**: Personalized tutoring, content creation, and assessment automation.
- **Creative Industries**: Storyboarding, video scripting, audio narration using multimodal variants (e.g., Sora, DALL·E, ElevenLabs).

7.4. Emerging Deployment Trends

 doi:10.20944/preprints202506.1134.v1

Several meta-trends are shaping how LLMs are built, released, and scaled:

- **Open Weight Releases**:
  Increasing community momentum around open-access models (e.g., Mistral, Mixtral, Phi-3, LLaMA 3) for reproducibility and customization.
- **Model Distillation & Quantization**:
  Techniques like 4-bit quantization (e.g., GPTQ) and distillation (e.g., DistilBERT) allow deployment on constrained hardware.
- **Retrieval-Augmented Generation (RAG)**:
  Combines LLMs with search or vector databases to ground outputs in external knowledge (e.g., LangChain, Haystack, LlamaIndex).
- **Agentic Systems**:
  Tools like AutoGen, CrewAI, LangGraph enable orchestration of multiple LLMs or tools in a goal-driven, persistent context.
- **Synthetic Data & Feedback Loops**:
  Self-generated data is used for continuous learning, evaluation, or instruction tuning (e.g., Self-Instruct, Feedback-Augmented Training).

7.5. Challenges in Deployment

Despite progress, deployment at scale faces notable hurdles:

- **Cost Efficiency**:
  LLM inference is resource-intensive. Token compression, batching, and caching are critical optimizations.
- **Latency and Interactivity**:
  Key for real-time applications like voice agents or interactive chatbots.
- **Governance and Safety**:
  Preventing misuse, managing model drift, and ensuring transparency remain unresolved in many commercial settings.
- **Evaluation at Scale**:
  Robust automated metrics (beyond BLEU/ROUGE) are still evolving, especially for subjective and open-ended tasks.

Suggested Diagram:
"**Ecosystem Landscape for LLM Deployment**" –A layered diagram showing:

- Foundation and fine-tuned models at the core
- Toolchains for tuning, serving, and evaluation in the middle layer
- Real-world applications and industry verticals in the outermost layer

8. Challenges and Open Questions

Despite the impressive advancements in large language models (LLMs), significant technical, ethical, and operational challenges remain. These limitations raise important open questions around **scalability**, **alignment**, **generalization**, **multimodality**, and **responsibility**, all of which must be addressed to ensure safe and effective deployment across domains.

8.1. Model Alignment and Safety

 doi:10.20944/preprints202506.1134.v1

Challenge:

Ensuring that LLMs reliably follow human intent, avoid harmful outputs, and remain controllable in dynamic environments.

Open Questions:

- How can alignment be maintained across increasingly capable models?
- What are scalable alternatives to RLHF and instruction tuning?
- Can models be made self-monitoring and self-correcting in real time?

8.2. Hallucination and Factual Consistency

Challenge:

LLMs, especially generative ones, frequently produce plausible-sounding but factually incorrect statements (hallucinations), even in high-stakes domains like healthcare or law.

Open Questions:

- How can we quantify and minimize hallucination across modalities?
- Can retrieval-augmented generation (RAG) fully solve this issue?
- How should factual grounding be incorporated during training?

8.3. Multimodal Integration Complexity

Challenge:

Integrating multiple modalities introduces additional alignment, fusion, and representation challenges. Synchronizing time-based data (e.g., audio + video + text) remains particularly difficult.

Open Questions:

- What are the best architectural paradigms for robust multimodal fusion?
- How do we benchmark cross-modal reasoning and transfer learning?
- How do we address bias and representation issues across modalities?

8.4. Generalization vs Specialization Trade-Off

Challenge:

Highly specialized models excel at specific tasks but often lack generalization, while foundation models are broad but may underperform on domain-specific tasks.

Open Questions:

- How can models balance generality with domain-specific expertise?
- Is modular composition (e.g., agents or tool-use) more efficient than training monolithic models?
- Can adapters or mixture-of-experts architectures improve task-specific efficiency?

8.5. Data Efficiency and Scaling Laws

Challenge:

LLMs require massive datasets and compute resources, raising concerns about sustainability, accessibility, and diminishing returns from scale.

Open Questions:

- What are the limits of current scaling laws?
- How can small models match large models through better training strategies (e.g., curriculum learning, active learning)?
- What role can synthetic or semi-supervised data play?

 doi:10.20944/preprints202506.1134.v1

8.6. Evaluation, Robustness, and Benchmarking

Challenge:

Standardized evaluation remains inconsistent across tasks and modalities. Many benchmarks fail to capture real-world performance, robustness to adversarial inputs, or ethical behavior.

Open Questions:

- What comprehensive and reliable metrics can be used beyond BLEU, ROUGE, or accuracy?
- How can we measure social biases, robustness, calibration, and uncertainty?
- How do we design benchmarks that evolve with models?

8.7. Legal, Ethical, and Societal Implications

Challenge:

Issues related to IP infringement, misinformation, surveillance, and bias become magnified as LLMs are deployed at scale.

Open Questions:

- Who is responsible for harms caused by autonomous LLM agents?
- How do we ensure transparent auditing of black-box foundation models?
- What governance structures are needed for open-weight vs proprietary models?

8.8. Model Interpretability and Trust

Challenge:

The inner workings of LLMs remain largely opaque. Users and developers struggle to interpret model behavior or predict failure modes.

Open Questions:

- Can interpretability techniques scale with model size and complexity?
- Are attention maps or feature attribution useful in multimodal settings?
- How do we build user trust in models that remain probabilistic and non-deterministic?

Suggested Diagram:

"**Landscape of Challenges in LLMs**" –

A radial chart or spider diagram visualizing key challenges (alignment, hallucination, multimodality, generalization, scaling, evaluation, ethics, interpretability) plotted by perceived **impact** vs **maturity of solutions**.

9. Future Directions

As large language models (LLMs) continue to evolve, the research community is shifting focus from scaling raw capabilities to **enhancing trustworthiness, efficiency, adaptability, and grounded reasoning**. This section outlines promising future directions that could shape the next generation of foundation, instruction-tuned, and multimodal language models.

9.1. Toward Unified Multimodal Intelligence

Future LLMs are expected to support seamless reasoning across **text, image, audio, video, and 3D modalities**, enabling richer forms of understanding and interaction. Innovations may include:

- **Truly universal encoders** that process arbitrary modality combinations.
- **Temporal reasoning frameworks** for video understanding and narration.
- **Cross-modal agent architectures** for real-world interaction (e.g., robotics, AR/VR).

9.2. Modular and Composable Architectures

 doi:10.20944/preprints202506.1134.v1

Rather than scaling monolithic models indefinitely, future systems may become **modular and composable**:

- **Mixture-of-Experts (MoE)** systems that dynamically activate subnetworks.
- **Adapter-based fine-tuning** for task-specific customization.
- **Agent-based orchestration** where smaller models or tools are composed for goal-driven behavior.

This shift can improve efficiency, interpretability, and resource sharing across domains.

9.3. Continual and Lifelong Learning

Static LLMs rapidly become outdated. Future models must:

- **Learn continuously** from new data and feedback in real-world environments.
- **Adapt on-device** or in federated settings, without catastrophic forgetting.
- Use **meta-learning** to quickly generalize to new domains with few examples.

9.4. Enhanced Alignment and Human-AI Interaction

Research is increasingly focused on **safe, grounded, and cooperative AI**. Emerging techniques include:

- **Constitutional AI** and **ethical scaffolding** for value alignment.
- **Human-in-the-loop interaction** for iterative improvement and personalized behavior.
- **Causal reasoning** and **epistemic uncertainty estimation** to improve robustness and trust.

9.5. Efficient and Sustainable Model Design

Reducing the environmental and economic footprint of LLMs is a priority:

- **Sparse and low-rank modeling**, quantization, and pruning for energy-efficient deployment.
- **Distilled small models** competitive with large models in constrained settings.
- **Hardware-aware training** and inference optimizations.

9.6. Model Governance and Open Research Infrastructure

To ensure responsible innovation, future directions must include:

- **Transparent benchmarking** and **standardized evaluation protocols**.
- **Auditable model cards**, datasheets, and usage logs.
- **Open ecosystems** with community-curated datasets and decentralized model stewardship (e.g., OpenLLM, EleutherAI, Hugging Face Hub).

9.7. Grounded and Tool-Augmented Models

Next-gen LLMs will integrate tightly with external tools and knowledge sources:

- **Retrieval-Augmented Generation (RAG)** pipelines that dynamically access structured and unstructured data.
- **Tool use APIs** for computation, database access, web browsing, and more.
- **Multistep planning and memory systems** enabling agent-like behavior.

9.8. Societal Integration and Human-Centric Design

LLMs will increasingly permeate human-centric domains such as:

 doi:10.20944/preprints202506.1134.v1

- **Education** (e.g., intelligent tutors with curriculum-awareness),
- **Healthcare** (e.g., patient-tailored dialogue agents),
- **Creative Workflows** (e.g., story generation, design tools, scientific assistants).

Future systems must be **inclusive**, **transparent**, and **collaborative**, reinforcing human agency rather than replacing it.

Suggested Diagram:

"**Roadmap for the Future of LLMs**" –A multi-phase roadmap diagram with time horizons (short-term, mid-term, long-term), each showing key advances in:

- Architecture
- Safety and alignment
- Deployment and tools
- Societal integration

10. Conclusions

Large Language Models (LLMs) have rapidly transformed from academic curiosities into foundational technologies across industry, research, and society. This survey has provided a structured comparative analysis of three key classes of LLMs—**foundation models**, **instruction-tuned models**, and **multimodal variants**—highlighting their respective architectures, capabilities, and deployment paradigms.

Foundation models demonstrate the power of scale and generalization but often lack alignment with user intent. Instruction-tuned models build upon these by refining model behavior through task-specific prompting and reinforcement learning from human feedback (RLHF). Multimodal LLMs represent the frontier, aiming to unify understanding across vision, language, and audio through increasingly complex architectures and data strategies.

Across all categories, we observe a convergence toward more **interactive**, **adaptable**, and **agentic** systems that blend multiple modalities, real-time feedback, external tools, and personalized behaviors. Yet this progress is accompanied by unresolved challenges in **factuality**, **safety**, **interpretability**, **bias mitigation**, and **sustainability**—areas that demand sustained research and cross-disciplinary collaboration.

As the field moves forward, the development of **modular**, **efficient**, and **ethically grounded** LLMs will be central to scaling innovation while safeguarding societal values. Comparative evaluations, open benchmarks, and transparent ecosystems will serve as the foundation for this progress.

In summary, the evolution of LLMs is far from complete. The integration of language, reasoning, perception, and interaction continues to blur traditional boundaries between disciplines and systems. This survey has laid the groundwork for understanding current distinctions and convergences in LLM design, and we anticipate that the next generation of models will be shaped not just by data and computation—but by purpose, responsibility, and inclusivity.

References

1. Pahune, S., & Chandrasekharan, M. (2023). Several categories of large language models (llms): A short survey. *arXiv preprint arXiv:2307.10188*.
2. Nokhwal, S., Chilakalapudi, P., Donekal, P., Nokhwal, S., Pahune, S., & Chaudhary, A. (2024, April). Accelerating neural network training: A brief review. In *Proceedings of the 2024 8th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence* (pp. 31-35).
3. Nokhwal, S., Pahune, S., & Chaudhary, A. (2023, April). Embau: A novel technique to embed audio data using shuffled frog leaping algorithm. In *proceedings of the 2023 7th international conference on intelligent systems, metaheuristics & swarm intelligence* (pp. 79-86).

 doi:10.20944/preprints202506.1134.v1

4. Nokhwal, S., Nokhwal, S., Pahune, S., & Chaudhary, A. (2024, April). Quantum generative adversarial networks: Bridging classical and quantum realms. In *Proceedings of the 2024 8th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence* (pp. 105-109).
5. Pahune, S., & Rewatkar, N. (2024). Large language models and generative ai's expanding role in healthcare.
6. Pahune, S. A. (2024). A brief overview of how ai enables healthcare sector rural development.
7. Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025). The Importance of AI Data Governance in Large Language Models. *Big Data and Cognitive Computing*, *9*(6), 147.
8. Pahune, S., & Akhtar, Z. (2025). Transitioning from mlops to llmops: Navigating the unique challenges of large language models. *Information*, *16*(2), 87.
9. Pahune, S. A., & Rewatkar, N. (2024). Investigating the application of quantum-enhanced generative adversarial networks in optimizing supply chain processes. *International Research Journal of Engineering and Technology (IRJET)*, *11*, 446.
10. Pahune, S., & Rewatkar, N. (2024). Cognitive automation in the supply chain: Unleashing the power of rpa vs. gen ai.
11. Gali, M., & Mahamkali, A. (2022). A Distributed Deep Meta Learning based Task Offloading Framework for Smart City Internet of Things with Edge-Cloud Computing. *J. Internet Serv. Inf. Secur.*, *12*(4), 224-237.
12. Mahamkali, A., Gali, M., Muniyandy, E., & Sundaram, A. (2023, October). IoT-Empowered Drones: Smart Cyber security Framework with Machine Learning Perspective. In *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)* (Vol. 1, pp. 1-9). IEEE.
13. Mahamkali, A. (2022). Health Care Internet of Things (IOT) During Pandemic–A Review. *Journal of Pharmaceutical Negative Results*, *13*.
14. Pande, S. D., & Khamparia, A. (Eds.). (2024). *Networks Attack Detection on 5G Networks Using Data Mining Techniques*. CRC Press.
15. Sharma, A., Gali, M., Mahamkali, A., Prasad, K. R., Singh, P. P., & Mittal, A. (2023, August). IoT-enabled Secure Service-Oriented Architecture (IOT-SOA) through Blockchain. In *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)* (pp. 264-268). IEEE.
16. Kumar, A., Keshta, I., Bhola, J., Wasim Bhatt, M., AlQahtani, S. A., & Gali, M. (2024). Application of Artificial Neural Network Unified with Fuzzy Logic for Systematic Stock Market Prediction. *Fluctuation and Noise Letters*, *23*(02), 2440001.
17. Kulkarni, C., Seifeddine, Z. D. M., Gali, M., & Degadwala, S. (2024). Mining intelligence hierarchical feature for malware detection over 5G network. In *Networks Attack Detection on 5G Networks using Data Mining Techniques* (pp. 64-82). CRC Press.
18. Gali, M., & Mahamkali, A. Health Care Internet of Things (IOT) During Pandemic–A Review.(2022). *Journal of Pharmaceutical Negative Results*, 572-574.
19. Veluguri, S. P. (2025, March). ConvAttRecurNet: An Attention-based Hybrid Model for Suicidal Thoughts Detection. In *2025 3rd International Conference on Disruptive Technologies (ICDT)* (pp. 860-865). IEEE.
20. Veluguri, S. P. (2025, January). Deep PPG: Improving Heart Rate Estimates with Activity Prediction. In *2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET)* (pp. 1-6). IEEE.