# GPT-4.5 vs LLaMA 4 Herd: A Comparative Analysis

## ( जीपीटी - 4.5 वर्सस लामा 4 हर्ड : ए कम्पेरेटिव एनालिसिस )

A RESEARCH PROJECT / DISSERTATION SUBMITTED
FOR THE PARTIAL FULFILLMENT OF THE DEGREE OF

**B.Sc. DATA SCIENCE (HONOURS WITH RESEARCH)**
**UNDER FACULITY OF ENGINEERING AND TECHNOLOGY**


**Submitted by**

**URMEET PAWAN**
**Regn. No. 21090115870008**

**UNDER THE SUPERVISION OF**
**Vikram Singh**
Professor
Department of Computer Science & Engineering
Chaudhary Devi Lal University, Sirsa




**UNIVERSITY SCHOOL FOR GRADUATE STUDIES (USGS)**
**CHAUDHARY DEVI LAL UNIVERSITY**
**SIRSA-125055, HARYANA, INDIA**
**JUNE-2025**

# STUDENT'S DECLARATION

I, **Urmeet Pawan**, Regn. No. 21090115870007, hereby declare that the study embodied in the present Research Project / Dissertation entitled "**GPT-4.5 vs LLaMA 4 Herd: A Comparative Analysis**" is based on my original research work and carried out by myself. This work has been done under the supervision of Vikram Singh, Department of Computer Science and Engineering, Chaudhary Devi Lal University, Sirsa, Haryana (India). My indebtedness to other works has been duly acknowledged at the relevant places.

The matter embodied in the Research Project / Dissertation has not been submitted in part or full for any other degree or diploma of any University or Institute.

Dated:

**Urmeet Pawan**
**Regn. No. 21090115870008**
B.Sc. Data Science
(Honours with Research)
University School for Graduate Studies
Chaudhary Devi Lal University
Sirsa

# SUPERVISOR'S CERTIFICATE

This is to certify that the study embodied in this Research Project / Dissertation entitled **"GPT-4.5 vs LLaMA 4 Herd: A Comparative Analysis"** comprises the investigations carried out by **Mr. Urmeet Pawan,** Regn. No. **21090115870008** under my guidance and supervision for the award of the degree of the **B.Sc. Data Science (Honours with Research).** To the best of my knowledge, the contents of this report in full or part have not been submitted to any other Institute or University for the award of any degree or diploma. No extensive use has been made of the work of other investigators, and wherever it has been used references have been given in the text.

Dated:

**Vikram Singh**
**Supervisor**

# ACKNOWLEDGEMENTS

Here, I would like to take a moment to thank those who have helped and inspired me throughout this journey and without their guidance and support I might not have been able to complete this research work.

Many individuals have contributed to this project with their advice, interest, time and support. Their encouragement drove me to complete this project. I thank God for making everything possible. I would like to express my sincere gratitude to my worthy supervisor Vikram Singh, Department of Computer Science and Engineering, Chaudhary Devi Lal University, Sirsa, for their constant support, inspiration and guidance. I am truly grateful to my supervisor for being an excellent advisor. His attention to detail and completeness to every aspect of the work, including in research, presentation and technical writings has helped me tremendously to improve my skills. I truly appreciate this help and truly privileged to work under his supervision.

I also express my sincere thanks to all other teaching & non-teaching staff of University School for Graduate Studies, Chaudhary Devi Lal University, Sirsa for their support and encouragement.

**Urmeet**

**Regn. No. 21090115870008**

# ABSTRACT

The modern era is being more and more driven by AI and LLMs are playing a huge role in this revolution by being the foundational blocks of the AI chatbots like ChatGPT, LLaMA, Deepseek, Grok, Gemini etc. which are being used in a variety of domains like medicine, diagnosis, education, content generation, computer vision and NLP tasks like text summarization, Question Answering, Multilingual tasks, Code generation, Mathematics, Logical reasoning etc. This study provides a comparative analysis between two such models, the GPT-4.5 and LLaMA 4 Herd, on several aspects like technical, ethical and bias, functional and performance on various tasks across several categories like math, reasoning, multilingualism, code generation etc. The models used, GPT-4.5 and LLaMA 4 (Scout, Maverick and Behemoth), in the study are evaluated through a framework LLM-Lens, introduced later in the study, on various datasets like MMLU, GPQA science, MMMLU, LiveCodeBench, SWE-bench, Hallucination rates (HHEM-2.1), factuality rates, MATH-500, AIME'24 which are industry recognized datasets used for LLM evaluation on different benchmark platforms. The GPT-4.5 and the Maverick variant of LLaMA 4 are also tested on a custom made 50 prompts dataset and evaluated on various criteria independently. This work showcases the power and potential of LLMs in multiple domains. This project aims to serve as a step forward in the practical application of the human-like-text generation and hopes to contribute meaningfully in the domain of Large Language Models.

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1

# Introduction

LLMs have emerged as a transformative force in AI, driving unprecedented advancements in NLP tasks. These type of models leverage deep learning to train on massive text corpora, resulting in *"*remarkable capabilities in natural language processing tasks and beyond*"* (Naveed et al. 2023). LLMs typically contain billions of parameters and follow empirical *scaling laws* – as model size, data size and compute increase, performance improves predictably. For example, Kaplan et al. (2020) found that "Larger models are significantly more sample-efficient", allowing very large networks to be trained on modest data amounts for optimal performance. Architecturally, modern LLMs are built on multi-layered neural networks which are the building blocks of deep learning and especially on the Transformer type architecture (Vaswani et al. 2017). The chapter is divided into five parts. Section 1.1 describes LLMs. Section 1.2 describes LLM AI Chatbots. Section 1.3 describes GPT and its timeline. Section 1.4 describes LLaMA and its timeline. Section 1.5 describes Applications of LLMs. Section 1.6 contains the report organization.

## 1.1 LLMs

LLMs are a category of deep learning, models that are designed to understand, mimic and generate human like language. They are trained on a huge amount of text datasets (often web-scale) and have learned to predict the next word or token in a particular context. As Minaee et al. (2024) note, LLMs has "drawn a lot of attention due to their strong performance on a wide range of natural language tasks" since tools like ChatGPT demonstrated their power. These capabilities arise by training "billions of model's parameters on massive amounts of text data". In practice, LLMs are deep neural networks that use several layers of artificial neurons to form hierarchical representations. As LeCun et al. (2015) describe DL (deep learning) models "are composed of multiple processing layers to learn representations of data with multiple levels of abstraction." These methods have improved the state-of-the-art in various domains like text, speech and vision dramatically. We first review the underlying technologies that enable LLMs.

### 1.1.1 Deep Learning

Deep learning is a subset of AI where there are multi-layered neural network models. In these models, each layer transforms its input into increasingly abstract features, enabling the network to "discover intricate structure in large data sets". Training is done by the process called backpropagation, which "indicate[s] how a machine should change its internal parameters… to compute the representation in each layer from the representation in the previous layer". LeCun et al. (2015) highlight that deep learning has driven breakthroughs across fields: deep convolutional nets have revolutionized image, audio and video processing, while recurrent nets (and now Transformers) handle sequential data like language. Indeed, Krizhevsky et al. (2012) famously trained and developed a deep convolutional network on 60 million parameters on ImageNet, achieving vastly better accuracy than previous methods. This example illustrates how scaling up deep networks and data yields powerful models and this is the principle that carries over to LLMs. In the same spirit, today's LLMs are trained on trillions of parameters with trillions of tokens to capture complex language patterns.

### 1.1.2 Neural Networks

A Neural Network (NN) is the foundational component of deep learning which is inspired by the structure and function of a human's brain. They are the very basic building blocks of deep learning which consisting of several layers of interconnected artificial neurons (units) that apply linear transformations and nonlinear activations. By stacking many layers and adjusting millions (or billions) of weights via backpropagation, neural networks learn to map inputs to outputs. In the context of language, neural networks can learn word and sentence representations and complex relationships between words. As LeCun et al. (2015) explain, deep neural models learn data representations "with multiple levels of abstraction". Modern LLMs use variants of neural networks (e.g. Transformers) that are highly scalable. The depth and size of these networks, along with large training sets, underlie the dramatic gains in LLM capabilities.

### 1.1.3 Transformer Models

Transformer models are one of the deep learning architecture types, completely changing the way machines understand and generate language. They were introduced

by researchers at Google in 2017; now being the foundational block of many powerful LLMs, including BERT, GPT and others. The Transformer architecture is the core innovation enabling current LLMs. “new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely” (Vaswani et al. 2017). In contrast to recurrent models, Transformers process tokens in parallel and use self-attention mechanism to capture long-range dependencies. This design dramatically accelerates training and improves performance. The original Transformer model achieved top-tier results on machine translation, “improving over the existing best results” by leveraging attention (Vaswani et al. 2017). The Transformer’s scalability allows training of very large models – for example, OpenAI’s GPT-3 is an autoregressive Transformer trained on 175 billion parameters (Brown et al. 2020). In summary, by combining deep multi-layer networks with self-attention, Transformer-based LLMs can learn powerful generative language models that generalize across tasks

## 1.2 LLM AI Chatbots

LLMs have been widely deployed as conversational AI or **chatbots**, where users interact via text (or voice) dialogue. Chatbots built on LLMs can answer questions, assist with writing, or carry on open-ended conversation. The public emergence of ChatGPT in late 2022 marked the mainstream breakthrough of LLM chatbots (Minaee et al. 2024). Behind the scenes, OpenAI’s InstructGPT work demonstrated how to align LLMs to user intent: by fine-tuning GPT-3 on wit human feedbacks, InstructGPT produced outputs that users preferred over GPT-3’s outputs, despite having 100 times fewer parameters. In human evaluations, a 1.3B parameter InstructGPT model was “preferred to outputs from the 175B GPT-3” model (Ouyang et al. 2022). This fine-tuning (RLHF) dramatically improved helpfulness and truthfulness of chatbot responses. These advances – combining large pre-trained models with alignment tuning – underpin today’s LLM-powered assistants. As a result, chatbots like ChatGPT, Bard and open-source LLaMA-based interfaces have become general-purpose AI agents, capable of answering questions, drafting text and even managing tasks across domains. Their conversational flexibility and knowledge integration illustrate the practical impact of LLM research.

## 1.3 GPT

A GPT model is a Transformer type architecture network pre-trained on a huge corpus to predict the next token, making it generative. The models are "pre-trained" on general text data and after that optionally "fine-tuned" for specific tasks. Brown et al. (2020) introduced GPT-3 as an autoregressive language model type with 175 billion parameters – "10× more than any previous non-sparse language model" and showed that by simply scaling up the model size, major gains can be achieved. GPT models have driven much of the recent progress in LLM performance, evolving rapidly in both size and capability.

### 1.3.1 History and Timeline of GPT

**GPT (2018)** – The first GPT introduced generative pre-training for Transformers. It had tens of millions of parameters (117M) and established the viability of pre-trained LMs for NLP tasks.

**GPT-2 (2019)** – A much larger model (1.5B parameters) trained on web text. OpenAI initially withheld it due to concerns but later released it, showcasing that large-scale unsupervised learning could generate coherent paragraphs.

**GPT-3 (2020)** – Brown et al. (2020) trained GPT-3 on 175B parameters on internet corpora. They found that "scaling up language models greatly improves task-agnostic, few-shot performance". GPT-3 achieved great results on question-answering, translation and cloze tasks without fine-tuning, simply by prompting the model (Brown et al. 2020). It popularized the ability to perform tasks with very few or completely no examples (few-shot/zero-shot).

**GPT-3.5 / ChatGPT (2022) –** Building on GPT-3, OpenAI released intermediate models fine-tuned for dialogue (often called GPT-3.5). These powered the initial ChatGPT product, providing conversational responses with minimal latency. They benefited from the InstructGPT alignment work (Ouyang et al. 2022).

**InstructGPT (2022)** – Ouyang et al. (2022) presented GPT-3 which was fine-tuned on human feedback, yielding models better at following instructions. In evaluations, even a 1.3 billion parameter InstructGPT model outperformed the vanilla 175 billion parameters GPT-3 on human preference metrics (Ouyang et al. 2022). This proved that model alignment could beat sheer scale.

**GPT-4 (2023) –** OpenAI introduced GPT-4 as a large-scale, multimodal model (accepting textual as well as image inputs). The GPT-4 Technical Report (2023) highlights that GPT-4 achieved "human-level performance on various professional and academic benchmarks", for example, it scored in the top 10% on a simulated bar exam. GPT-4 remains a Transformer-based model with advanced instruction-following capabilities, refined via post-training alignment for safety and factuality (Achiam et al. 2023).

The latest iteration, GPT-4o (announced 2024), further extends GPT-4's capabilities. Known as "omni" for its multi-modality, GPT-4o accepts text, audio, images and video simultaneously. According to OpenAI, "GPT-4o… can reason across audio, vision and text in real time". It responds to voice inputs with low latency (hundreds of milliseconds) (OpenAI, 2024) and supports real-time multi-agent interactions. Though the full technical report is beyond our scope, GPT-4o represents the cutting edge of GPT developments, emphasizing real-time, multimodal AI assistants.

**GPT-4.5 (2025)**: The latest preview, dubbed GPT-4.5, further scales the GPT paradigm. According to OpenAI's announcement, GPT-4.5 improves upon GPT-4 by expanding training (unsupervised learning) and data, resulting in a "broader knowledge base" and more natural interactions. Early reports claim it feels more fluent with fewer hallucinations. While still under evaluation, GPT-4.5 represents the next step in the series, leveraging ever-larger pre-training to boost pattern recognition and creativity (OpenAI, 2025).

Throughout its evolution, ChatGPT has not only expanded its capabilities but also increased public and academic interest in responsible AI use, language model safety and ethical deployment in real-world applications.

## 1.4 LLaMA Herd

LLaMA which stands for Large Language Model Meta AI is Meta AI's family of open-source LLMs that are designed to democratize access to powerful language models. The original LLaMA models (2023) covered 7B–65B parameters and were trained entirely on public data. Touvron et al. 2023 report that LLaMA-13B outperformed OpenAI's GPT-3 (175B) on most of the benchmarks and the largest LLaMA-65B was on par with state-of-the-art commercial models. These results

showed that with efficient training, smaller open models could match much larger proprietary ones. LLaMA models use the Transformer backbone but also focus on efficiency and open access, releasing all models to the research community.

Meta continued with **LLaMA 2 (2023)**, a next-generation release of models up to 70B parameters. LLaMA 2 includes both foundational pre-trained models and specialized chat-tuned versions. According to the LLaMA 2 paper, the chat models (LLaMA 2-Chat) are optimized for dialogue like humans and "outperform open-source chat models on most benchmarks". Human evaluations indicated LLaMA 2-Chat is a viable alternative to closed-source chatbots in terms of safety and helpfulness. In short, LLaMA 2 built on the original LLaMA by increasing scale and adding alignment fine-tuning for conversational use (Touvron et al. 2023).

### 1.4.1 History and Timeline of LLaMA

**In February 2023**, the initial version, LLaMA 1, was launched with models having parameters from 7 billion to 65 billion. Trained exclusively on datasets that are available publicly encompassing roughly 1.4 trillion tokens, this model emphasized efficiency by demonstrating that smaller models could achieve results comparable to substantially larger counterparts. Notably, the 13-billion-parameter variant presented a competitive performance with OpenAI's GPT-3 (175B parameters). The 65-billion-parameter model performed on par with leading models such as Google's PaLM and DeepMind's Chinchilla. They were released under Meta's community license. (Touvron et al. 2023).

**In July 2023**, LLaMA 2 expanded on the original series, introducing larger models including a prominent 70-billion-parameter variant. This generation benefited from a 40% increase in training data volume, notably focusing on higher-quality datasets. LLaMA 2 marked a critical point in accessibility by releasing models under more permissive licenses, allowing broader commercial and research-oriented utilization. Meta also introduced Code LLaMA within this iteration, a variant specifically optimized for software coding tasks, enhancing its applicability across various technical domains (Touvron et al. 2023).

**From April to December 2024**, Meta continued to refine its language models with LLaMA 3, initially presented with models comprising 8 billion and 70 billion

parameters. Training expanded to approximately 15 trillion tokens from diversified public datasets. Additionally, over 10 million instruction-tuned examples were integrated for the purpose of enhancing the model's contextual understanding and responsiveness. Subsequent sub-versions—LLaMA 3.1, 3.2 and 3.3—introduced models with significantly larger scales, notably a 405-billion-parameter model. These variants offered extensive enhancements, including multilingual support, multimodal processing capabilities and extended context windows, reaching up to 128,000 tokens. This positioned the LLaMA series as a leading open-source alternative to proprietary language models (Grattafiori et al. 2024).

**In April 2025**, the release of LLaMA 4 in April 2025 marked a pivotal transformation with the adoption of mixture-of-experts (MoE) architectures. This approach enhanced model efficiency by activating subsets of parameters dynamically, facilitating scalable yet computationally efficient training. The LLaMA 4 herd introduced two primary models at launch: the "Scout", having 17 billion active parameters across 16 experts, which totals to 109 billion parameters and supporting context lengths up to 10 million tokens; and the "Maverick", also having 17 billion active parameters but across 128 experts, which totals to 400 billion parameters and extending context windows up to 1 million tokens. Furthermore, a significantly larger model, "Behemoth", comprising approximately two trillion total parameters (with 288 billion active parameters), was under active training as of this release. LLaMA 4 notably included native multimodal processing, adeptly handling text, images and videos and offered expansive multilingual functionality covering languages such as Hindi, French and Spanish. Its dataset comprised over 30 trillion tokens, combining publicly available data with proprietary sources from Meta (Hugging Face, 2025).

## 1.5 Applications of LLMs

LLMs have demonstrated substantial versatility across diverse fields due to their sophisticated natural language processing capabilities. The following are ten significant applications of LLMs:

**a) Text Generation and Summarization**

LLMs can automatically generate coherent and contextually relevant text. They excel in summarizing lengthy documents or articles, extracting critical information and

creating concise summaries, greatly assisting information retrieval processes (Brown et al. 2020).

**b) Machine Translation**

These models significantly improve translation accuracy across various languages, facilitating seamless global communication. Their advanced linguistic understanding allows translations to capture nuances, context and idiomatic expressions better than traditional translation tools (Vaswani et al. 2017).

**c) Sentiment Analysis and Opinion Mining**

LLMs analyze large volumes of text to identify and interpret subjective information, such as emotions and attitude of the opinions expressed in reviews, social media and customer feedback, which is valuable for market research and customer relationship management (Devlin et al. 2019).

**d) Question-Answering Systems**

LLMs underpin intelligent question-answering systems, enabling them to understand the context of the natural language queries and respond to them with precision. Applications include virtual assistants, customer service chatbots and educational platforms providing immediate, accurate responses (Brown et al. 2020).

**e) Code Generation and Software Development**

Models like Code LLaMA or GitHub Copilot, which leverage LLM technologies, aid developers by generating, debugging and optimizing code. They assist in reducing the time spent on routine coding tasks, thereby accelerating software development cycles.

**f) Educational Tools and Personalized Learning**

LLMs support personalized education by creating interactive learning environments. They generate customized study materials, quizzes and detailed feedback tailored to individual learning styles, significantly enhancing student engagement and comprehension.

**g) Medical Diagnostics and Clinical Decision Support**

In healthcare, LLMs analyze medical records, scientific literature and clinical notes to assist in diagnostic decision-making processes. They provide clinicians with evidence-based recommendations and interpret complex medical data, improving patient care and outcomes.

**h) Content Moderation and Safety**

LLMs effectively moderate online content by automatically detecting harmful or inappropriate text, images and videos. This capability helps social media platforms maintain safety standards, enforce guidelines and foster healthier online interactions.

**i) Creative Writing and Content Creation**

These models contribute to creative fields by drafting scripts, articles, stories and marketing content. Their ability to generate engaging, diverse and human-like content supports writers and marketers in producing high-quality materials more efficienty.

**j) Data Analysis and Insights Extraction**

LLMs analyze unstructured text data to identify patterns, trends and valuable insights. They streamline data-driven decision-making processes across sectors like finance, business analytics and market research, converting vast textual datasets into actionable intelligence (Devlin et al. 2019).

## 1.6 Report Organization

The present study has been divided into five chapters as under

### Chapter 1: Introduction

Introduces Large Language Models, Deep Learning, Neural Network, Transformer Models, LLM AI Chatbots, GPT and its history, LLaMA Herd and its history, Application of LLMs.

### Chapter 2: Literature review

Presents the review of literature related to the study under consideration. In this chapter review of the research paper containing the work related to the present study has been presented and discussed the research gaps.

### Chapter 3: Research, Objective and Methodology

Deals with the research methods used in this study. This chapter presents the problem statement, study objectives, the scope of the study, the research methods used in the study, the technique used in research, the tools used in research and the machine learning library used.

## Chapter 4: Proposed Framework and Results

The Chapter 4 contains the proposed framework, framework flow, Models used in the study and the Analysis on three core aspects.

## Chapter 5: Discussion, Conclusion and Future Scope

This chapter contains the results and the conclusion drawn from the present research work.

# Chapter 2

# Literature Review

The chapter is subdivided into three sections. Related work has been described in section 2.1 whereas Section 2.2 describes a comparative analysis of literature review and Section 2.3 outlines research gaps.

## 2.1 Related Work

**Brown et al. (2020)** introduced GPT-3 which was an autoregressive language model notable for its extensive size of 175 billion parameters, significantly larger than previous models. The primary contribution of their study was demonstrating that scaling language models substantially enhances their performance on diverse NLP tasks without the need of specific fine-tuning for specific tasks. Specifically, GPT-3 exhibits remarkable few-shot learning abilities, where it can effectively perform various tasks merely by being provided with a small number of examples or instructions. These tasks include question-answering, translation, cloze tests and certain tasks which require reasoning and domain adaptation, such as word unscrambling, arithmetic and novel word usage. The results indicated that, in many instances, GPT-3's performance with few-shot approach exceeded the existing state-of-the-art models' which were fine-tuned specifically for those tasks. However, Brown et al. also acknowledged certain limitations, including challenges with some reasoning tasks and methodological some issues related to training on internet-sourced datasets. This study underscored the potential of large-scale models in reducing dependence on extensive labeled data and highlighted broader implications regarding model capability and societal impacts.

**Vaswani et al. (2017)** presented Transformer architecture, as a novel neural network architecture which completely and solely relied on attention mechanisms, completely discarding recurrence and convolution in sequence transduction tasks. The authors proposed the use of self-attention and multi-head attention to model dependencies in input and output sequences more effectively, regardless of their distance. Their architecture consists of stacked encoder and decoder layers, each incorporating multi-head self-attention and feed-forward neural networks. One of the key contributions was the scaled dot-product attention mechanism, which allows the model to compute

attention weights efficiently. The Transformer achieved state-of-the-art results on machine translation benchmarks such as WMT 2014 English-to-German and English-to-French tasks, while requiring significantly less training time compared to traditional RNN-based models. Additionally, the model generalized well to syntactic tasks like English constituency parsing. This work laid the foundational framework for modern large language models (LLMs), including GPT and BERT, by demonstrating that attention alone is sufficient for high-performance sequence modeling.

**LeCun et al. (2015)** provided a landmark review of deep learning, outlining its theoretical foundations, key architectures and broad range of applications across multiple domains. The authors began by describing deep learning as a class of machine learning techniques that involve multi-layered neural networks capable of learning hierarchical feature representations directly from raw data. Unlike traditional machine learning methods that relied heavily on hand-crafted features, deep learning models—particularly deep neural networks and convolutional neural networks (CNNs)—automatically learn intermediate representations through successive layers of abstraction. A major focus of the paper is on the success of CNNs in image classification tasks, speech recognition and natural language processing. For instance, CNNs were shown to outperform prior models in ImageNet classification, speech transcription and biomedical data interpretation by leveraging local receptive fields, weight sharing and pooling mechanisms. The paper also covers the role of stochastic gradient descent and backpropagation in training deep networks efficiently. The authors emphasized the generalizability of these networks, citing their ability to scale with increased data and compute. Furthermore, the paper explores recurrent neural networks (RNNs) and memory-augmented networks for sequential data tasks such as machine translation and question answering. Ultimately, LeCun et al. projected that deep learning would continue to advance AI by enabling systems to perform intuitive, high-level reasoning tasks with minimal human supervision.

**Krizhevsky et al. (2012)** introduced a deep convolutional neural network (CNN), known as AlexNet, which achieved groundbreaking performance on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). Trained on 1.2 million high-resolution images across 1000 classes, the model achieved a top-5 error rate of 15.3%, outperforming the next-best entry by over 10%. The network architecture included

five convolutional layers and three fully connected layers, utilizing ReLU activation, dropout for regularization and local response normalization. The authors leveraged GPU acceleration and extensive data augmentation to handle overfitting and computational demands. This work marked a major milestone in computer vision and deep learning, initiating a shift toward deep CNNs for large-scale image classification.

**He et al. (2016)** introduced the deep residual learning framework (ResNet), addressing the degradation problem in very deep neural networks. Rather than learning direct mappings, residual networks learn residual functions with reference to the input, using identity-based shortcut connections. This reformulation makes it easier to optimize deep models and allows training of networks with over 100 layers. On ImageNet, ResNets achieved state-of-the-art performance, with a 152-layer model obtaining a top-5 error of 3.57%, winning ILSVRC 2015. The residual approach also generalized well to other tasks such as object detection and segmentation, demonstrating strong results on COCO and PASCAL VOC datasets. The paper's key contribution was showing that deeper networks could perform better when optimization is stabilized through residual design, marking a pivotal advancement in deep learning model architecture.

**Naveed et al. (2023)** present a comprehensive and structured survey of large language models (LLMs), offering an in-depth exploration of their evolution, architecture, training methodologies and real-world applications. The paper systematically covers foundational concepts such as pre-training, fine-tuning, instruction alignment and prompting and analyzes advanced topics like multimodal LLMs, augmented models and parameter-efficient tuning strategies. A key contribution is its categorization of LLM research into branches like architecture, efficient inference and evaluation, while highlighting emergent abilities such as reasoning, planning and zero-shot learning. The authors also discuss scalability, alignment challenges and optimization techniques, including mixture-of-experts, quantization and context length scaling. Their review extends to dozens of popular LLMs—such as GPT-3, T5, LLaMA, PaLM and Chinchilla—detailing architectural variations, tokenization strategies and performance benchmarks. This paper serves as both a reference and roadmap for researchers, synthesizing insights from hundreds of works to depict the state of LLM

research and its trajectory, particularly emphasizing the shift toward open-source, instruction-tuned and multimodal systems.

**Devlin et al. (2019)** presented BERT (Bidirectional Encoder Representations from Transformers), a novel language representation model that significantly advanced the benchmark in NLP. Unlike previous models that relied on unidirectional context (e.g., OpenAI GPT), BERT uses a deep bidirectional transformer architecture, allowing it to jointly condition on both left and right contexts across all layers. The authors proposed two unsupervised pre-training tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). These tasks enabled BERT to learn contextual relationships between words and sentences more effectively. Once pre-trained, BERT could be fine-tuned with minimal architectural changes for various downstream tasks such as question answering, natural language inference and named entity recognition. The results demonstrated that BERT outperformed existing models on 11 NLP tasks, including achieving an F1 score of 93.2 on SQuAD v1.1 and an accuracy of 86.7% on the MultiNLI benchmark. This work laid a foundational framework for subsequent transformer-based models and highlighted the effectiveness of bidirectional context in language understanding.

**Minaee et al. (2024)** delivered an extensive and structured survey of large language models (LLMs), focusing on their development, core technologies, capabilities and challenges. The authors reviewed major LLM families such as GPT, LLaMA and PaLM, detailing their architectural differences, parameter scaling and emergent abilities like in-context learning, multi-step reasoning and instruction following. The paper categorized LLMs by architecture—encoder-only (e.g., BERT), decoder-only (e.g., GPT) and encoder-decoder (e.g., T5)—and discussed pretraining objectives like masked language modeling and autoregressive prediction. It further examined key advancements including reinforcement learning from human feedback (RLHF), mixture-of-experts layers and efficient fine-tuning techniques. The survey also outlined critical components of building LLMs: data preparation, tokenization, alignment strategies and inference optimizations. Additionally, it analysed widely-used evaluation benchmarks, datasets and challenges such as bias, inefficiency and alignment. Concluding with future directions, the paper emphasized the growing importance of multimodal LLMs, open-source innovation and responsible deployment

practices, positioning itself as a foundational reference for both researchers and practitioners.

**Kaplan et al. (2020)** conducted a foundational study on scaling laws for neural language models, empirically analysing how performance varies with model size, dataset size and compute budget. Their experiments, primarily with Transformer architectures, revealed that language modeling loss follows a predictable power-law decrease as these three factors are scaled up. Surprisingly, architectural details like network depth or width had minimal impact compared to total scale. The study found that larger models are significantly more sample-efficient, needing less data to achieve similar performance and that compute-optimal training involves stopping before convergence with very large models. Importantly, the authors developed mathematical formulations for optimal compute allocation, showing that for best results, one should prioritize increasing model size over training time or dataset size. They also identified universal behaviors in overfitting and generalization, offering a reliable framework to guide future model training. This work became a critical reference for understanding the trade-offs in building ever-larger language models.

**Ouyang et al. (2022)** introduced InstructGPT, a family of models fine-tuned from GPT-3 using human feedback to better align outputs with user intentions. The authors developed a three-stage process involving supervised learning, reward modeling and reinforcement learning from human feedback (RLHF). They found that a 1.3B-parameter InstructGPT model was preferred over the much larger 175B GPT-3 in human evaluations, demonstrating that alignment can outweigh raw scale in improving helpfulness. InstructGPT also showed improved truthfulness and reduced toxicity—generating 25% fewer toxic outputs—although bias reduction remained limited. The paper reports a significant drop in hallucination rates, from 41% with GPT-3 to 21% with InstructGPT, on closed-domain tasks. Furthermore, the models generalized well to instructions not seen during training, including code summarization and non-English queries. Despite these advances, InstructGPT still made factual and logical errors and could follow harmful instructions if not properly guided. The authors concluded that RLHF is a promising method for aligning LLMs with human values while retaining task performance.

**Wei et al. (2022)** proposed Chain-of-Thought (CoT) prompting, a method that improves large language models' (LLMs) reasoning abilities by including intermediate natural language reasoning steps in prompts. The study found that CoT prompting significantly enhances performance on complex tasks like arithmetic, commonsense and symbolic reasoning, particularly for models with 100B+ parameters such as PaLM 540B and GPT-3. Experiments on benchmarks like GSM8K and SVAMP showed that PaLM 540B achieved state-of-the-art results with only eight CoT exemplars. The method outperformed standard few-shot prompting and demonstrated robustness across different annotators, prompt orders and tasks. Importantly, the authors observed that reasoning ability is an emergent property of model scale, with smaller models failing to benefit from CoT. Additionally, ablation studies confirmed that the improvements stem from step-by-step natural language reasoning, not just added computation or symbolic output. The work underscores CoT prompting as a simple yet powerful tool for unlocking reasoning in LLMs without fine-tuning**.**

**Achiam et al. (2023)** presented the technical report for GPT-4, a large multimodal transformer model capable of processing both image and text inputs and generating coherent text outputs. The report details how GPT-4 surpasses previous models (like GPT-3.5) in language understanding, multilingual reasoning and benchmark performance. It achieves human-level results on many standardized exams—scoring in the top 10% on the simulated bar exam and outperforming state-of-the-art models across tasks such as MMLU, ARC, HumanEval and GSM8K. GPT-4 was fine-tuned using Reinforcement Learning from Human Feedback (RLHF), improving factual accuracy and alignment with user intent. Notably, the model exhibits reduced hallucination and toxicity rates compared to its predecessors, as well as improved performance on safety and refusal metrics. The report also highlights advancements in predictable scaling, enabling capability estimation from smaller models and discusses safety pipelines that integrate model-assisted reward classifiers. Despite its strengths, GPT-4 still has limitations like occasional factual errors and overconfidence, necessitating cautious deployment and continued safety research.

**Poldrack et al. (2023)** investigated the capabilities of GPT-4 for AI-assisted coding through a series of interactive experiments. The study assessed GPT-4's ability to generate, refactor and test Python code, with a focus on usability for researchers

lacking prompt engineering expertise. Initial findings showed GPT-4 could solve 72% of tasks within a few prompts, but often required human correction due to outdated knowledge or mathematical hallucinations. Refactoring experiments revealed significant improvements in code quality, reducing linting errors and enhancing maintainability based on software metrics such as cyclomatic complexity and Halstead difficulty. However, automatic test generation, though yielding high coverage, suffered from frequent failures—demonstrating that test reliability remains a challenge. The authors concluded that while GPT-4 is a powerful assistant for improving and generating code, human oversight remains critical, especially for scientific and mathematically intensive applications. The results highlight both the promise and current limitations of large language models in software development.

**Yan et al. (2025)** proposed GPT-ImgEval, the first benchmark suite designed to evaluate the image generation capabilities of GPT-4o, OpenAI's multimodal model. The study assesses performance across three core tasks: text-to-image generation (GenEval), instruction-based image editing (Reason-Edit) and world knowledge-guided semantic synthesis (WISE). GPT-4o achieved state-of-the-art results, showing precise object rendering, accurate color/spatial reasoning and superior instruction-following in edits. It outperformed models like DALL·E 3 and Janus-Pro in quality and semantic alignment. The authors further investigated GPT-4o's internal architecture, hypothesizing a hybrid autoregressive-diffusion model, supported by classifier-based analysis. Despite its strengths, GPT-4o showed weaknesses including inconsistencies in high-res edits, warm color bias and difficulty with complex scenes or non-English text. For safety, GPT-4o's outputs remain detectable by AI forensic tools due to distinct upsampling artifacts. This work provides a rigorous foundation for evaluating generative multimodal models and offers architectural insights for future development.

**Touvron et al. (2023)** introduced LLaMA, a family of open-source foundation language models ranging from 7B to 65B parameters, trained solely on publicly available datasets. Unlike previous proprietary LLMs such as GPT-3 and PaLM, LLaMA models achieve competitive performance while being more efficient and accessible. The authors employed training strategies based on the Chinchilla scaling laws and incorporated architectural enhancements like SwiGLU activation, rotary embeddings and pre-normalization. LLaMA-13B surpasses GPT-3 across several

benchmarks, while the 65B model matches Chinchilla-70B and PaLM-540B. Evaluations cover reasoning (MMLU), math (GSM8K, MATH), code generation (HumanEval) and ethics (TruthfulQA, RealToxicityPrompts). LLaMA also demonstrates lower training carbon emissions and highlights the impact of data quality on bias and toxicity. The work emphasizes the potential of open models in democratizing access to advanced LLMs and encourages responsible AI development through transparent, reproducible training processes.

**Touvron et al. (2023)** proposed LLaMA 2, an open-weight family of pretrained and fine-tuned large language models ranging from 7B to 70B parameters, optimized for dialogue through the LLaMA 2-Chat variant. The models were trained on 2 trillion tokens of public data, with improvements like doubled context length, grouped-query attention and supervised fine-tuning (SFT) which is followed by RLHF (Reinforcement Learning with Human feedback). The team introduced innovations such as Ghost Attention for better multi-turn consistency and trained reward models for safety and helpfulness separately, collecting over 1.4 million human preference annotations. Evaluation results showed that LLaMA 2-Chat outperforms most open-source models and approaches closed-source alternatives like GPT-3.5 in tasks like MMLU and GSM8K, though it lags in code generation. Safety protocols included adversarial testing and red-teaming. The paper emphasizes transparency, energy reporting and responsible release, positioning LLaMA 2 as a scalable, open alternative for safe and high-quality conversational AI.

**Grattafiori et al. (2024)** introduced LLaMA 3, a suite of open-weight large language models designed to support multilinguality, coding, reasoning, tool use and long-context understanding. The models—ranging from 8B to 405B parameters—are trained on 15.6T high-quality tokens using a dense Transformer architecture and scale-optimized infrastructure. LLaMA 3.1, the latest version, integrates instruction tuning, Direct Preference Optimization (DPO) and post-training strategies for alignment, safety and capability enhancement. The flagship 405B model rivals GPT-4 in benchmarks like MMLU, GSM8K and HumanEval. Extensive safety work includes Llama Guard 3 for input/output filtering and a 6-round iterative alignment process. Additionally, LLaMA 3 explores multimodal extensions, incorporating vision and speech through adapter modules. Its models show strong performance in multilingual tasks and long-context benchmarks (up to 128K tokens). By prioritizing openness,

scalable training and architectural simplicity, the work positions LLaMA 3 as a robust, community-accessible alternative to proprietary LLMs like GPT-4.

**Huang et al. (2024)** conducted an empirical study to evaluate the quantization performance of LLaMA 3 models (8B and 70B) under low-bit compression techniques. The study focuses on post-training quantization (PTQ) and LoRA-based fine-tuning (LoRA-FT), assessing their impact on both language and multimodal tasks. Using benchmarks such as MMLU, HellaSwag and various visual-language datasets, the results show that 4-bit quantization maintains near-original performance, while 2–3 bit quantization leads to significant degradation—especially in visual-language settings. Methods like GPTQ, AWQ and SliM-LLM outperformed others in preserving accuracy at lower bit-widths. However, ultra-low bit-widths (≤2 bits) still lead to model collapse, particularly in MLLMs like LLaVA-Next. Notably, IR-QLoRA emerged as the most robust LoRA-FT method. The study concludes that while LLaMA 3 shows strong quantization tolerance, new techniques are needed for stability at extreme compression, especially in multimodal applications, guiding future low-resource deployment of LLMs.

**Fernandes et al. (2025)** evaluated 16 LLMs for their ability to generate Python code to solve LoRaWAN-based UAV deployment tasks using natural language prompts of increasing complexity. The study compared lightweight, locally executed models such as LLaMA-3.3 and Phi-4 against high-end proprietary models like GPT-4 and DeepSeek-V3, focusing on functional correctness, runtime behaviour and code reliability. Three zero-shot prompts were designed to progressively increase difficulty—ranging from index selection based on propagation loss to calculating exact received signal power. Code accuracy was scored on a 0–5 scale. Results showed that DeepSeek-V3, GPT-4, Phi-4 and LLaMA-3.3 consistently outperformed others across all prompts and temperature settings. Surprisingly, smaller models like Phi-4 (14B) matched or surpassed larger ones like Qwen-32B in stability and accuracy. The findings affirm the viability of smaller, open-access models for engineering applications, emphasizing the importance of model alignment, prompt clarity and temperature sensitivity in successful code generation for domain-specific tasks**.**

## 2.2 Comparative Analysis of Literature Review

A Comparative analysis of the reviewed literature is essential to highlight the strengths, limitations and methodological differences across existing works.

The following table summarizes key aspects of each study, enabling a clearer understanding of research trends and gaps.

**Table 2.1 Comparative Analysis of Literature Review**

| Research Articles | Focus / Application Area | LLM(s) Evaluated / Framework Used | Key Findings / Metrics |
|---|---|---|---|
| LeCun et al. (2015) | Deep learning architectures and their general applications | CNNs, RNNs | Established DL as a general-purpose tool; emphasized hierarchical feature learning and minimal supervision |
| Vaswani et al. (2017) | Sequence modeling with attention mechanisms | Transformer | Introduced self-attention; eliminated recurrence; foundational for LLMs like GPT and BERT |
| Krizhevsky et al. (2012) | Image classification with deep CNNs | AlexNet (CNN) | Top-5 error 15.3% on ImageNet; pioneered GPU-accelerated training for large-scale vision tasks |
| Devlin et al. | Bidirectional | BERT | Outperformed prior |

| | | | |
|---|---|---|---|
| (2019) | language understanding | | models on SQuAD and MultiNLI using MLM and NSP; bidirectional trans-former innovation |
| Brown et al. (2020) | Scaling laws and few-shot learning in NLP | GPT-3 | 175B parameters; few-shot capabilities rivaled fine-tuned models; limitations in reasoning and dataset bias |
| Kaplan et al. (2020) | Scaling laws for neural language models | Transformer-based LMs | Established scaling laws for model/data/compute; found predictable power-law trends in performance |
| Ouyang et al. (2022) | Alignment of LLMs with human intent | InstructGPT | Used RLHF; 1.3B InstructGPT preferred over 175B GPT-3; reduced hallucination and toxicity |
| Wei et al. (2022) | Reasoning enhancement via prompt design | GPT-3, PaLM 540B | Chain-of-Thought prompting improved reasoning on GSM8K, SVAMP; scale-dependent efficacy |
| Achiam et al. (2023) | Technical report on GPT-4's | GPT-4 | Multimodal; SOTA on MMLU, ARC, |

| | | | |
|---|---|---|---|
| | capabilities | | HumanEval; improved safety, reduced hallucinations |
| Poldrack et al. (2023) | Code generation and refactoring with GPT-4 | GPT-4 | 72% task success in code gen; improved code quality; needed human correction for hallucinations |
| Yan et al. (2025) | Image generation benchmarking for GPT-4o | GPT-4o | Outperformed DALL·E 3 and others; excelled in GenEval, Reason-Edit; weak in complex scenes, non-English text |
| Touvron et al. (2023) | Introduction of LLaMA family | LLaMA (13B-65B) | LLaMA-13B surpasses GPT-3 while LLaMA-65B matches Chinchilla-70B and PaLM-540B |
| Touvron et al. (2023) | Dialogue optimization and safety in LLaMA 2 | LLaMA 2 (7B–70B) | Trained on 2T tokens; RLHF + Ghost Attention; strong multi-turn performance |
| Grattafiori et al. (2024) | Multimodal and multilingual capabilities in LLaMA 3 | LLaMA 3.1 | 405B model rivaled GPT-4; supported vision/speech; Llama Guard 3 for alignment |

| | | | |
|---|---|---|---|
| Huang et al. (2024) | Quantization of LLaMA 3 for efficient inference | LLaMA 3 (8B, 70B) | 4-bit quantization preserved performance; 2–3 bit led to degradation; IR-QLoRA most robust |
| Fernandes et al. (2025) | Code generation using LLMs for UAV deployment | GPT-4, LLaMA-3.3, DeepSeek-V3, Phi-4 | DeepSeek and GPT-4 had highest code accuracy; smaller models viable for engineering tasks |
| Naveed et al. (2023) | Survey on LLMs' architecture, training and applications | GPT-3, T5, PaLM, LLaMA, Chinchilla | Categorized LLM design trends; highlighted efficient tuning, alignment and emergent abilities |
| Minaee et al. (2024) | Comprehensive survey of modern LLMs | GPT, PaLM, BERT, LLaMA | Surveyed LLM architectures, training methods and evaluation; emphasized open-source and multimodal trends |

## 2.3 Research Gaps

Despite the progress and various studies being introduced in the LLM AI Chatbot domain, there are still further research Gaps being generated continuously with the introduction of new techniques and methods in this domain. Some key research Gaps include:

1. While numerous studies comparing earlier versions of the GPT and LLaMA on various aspects, no studies provide a rigorous direct comparison between the newly introduced GPT-4.5 and LLaMA 4 currently representing the largest generation of LLMs.
2. Existing works focus on older LLaMA (e.g., 2/3) or GPT-4 but do not explore performance differences with the newest capabilities (like extended context windows or tool usage) in real-world scenarios.

The present research aims to fill the gap by independently validating and evaluating these new LLMs claiming to be substantially improved in use cases such as code generation, reasoning tasks, low-resource languages, or decision making.

# Chapter 3

# Research, Objective and Methodology

Research methodology is a blueprint which guides in solving the research problem in a systematic way. It is the step-by-step record of several steps generally practiced by a researcher in his/her journey of studying a research problem and the logic behind them. The chapter is subdivided into six parts. Section 3.1 contains the problem statement. Research objectives are in Section 3.2. Section 3.3 contains the research methodology that is used in the conductance of the current study. Section 3.4 is about scope of the study. Section 3.5 provides overview of the technique used followed by the Section 3.6 containing the tools used in the study.

## 3.1 Problem Statement

The problem is entitled as "GPT-4.5 vs The LLaMA 4 Herd: A Comparative Analysis". Both LLMs claiming to be the largest generation of LLMs, are being integrated into every possible field, IT or Non-IT, at a very rapid rate. There is a notable absence of rigorous, head-to-head analysis between the latest advancements i.e. GPT-4.5 and LLaMA 4 Herd, which are widely recognized for their enhanced performance, reasoning capabilities and user alignment. Critical dimensions such as instruction-following, hallucination mitigation, efficiency trade-offs and ethical alignment remain underexplored in current comparative studies.This lack of empirical benchmarking and transparent assessment limits end users' utility over these leading them to sluggish results that could have been better if known which one to utilize according to tasks (especially when models like GPT-4.5 are commercials) and stakeholders' ability to make informed decisions about model deployment in diverse application domains. Therefore, a comprehensive and systematic comparison of GPT-4.5 and LLaMA 4 Herd is essential to fill these gaps and guide responsible, effective use of modern LLMs.

## 3.2 Research Objective

1. To study LLMs as general.
2. To study the SOA LLMs, GPT-4.5 and LLaMA 4 Herd.
3. To design and implement a test framework to test both models on various aspects.

4. To perform independent validation and evaluation on the performance of both models.
5. To recommend the optimal LLM model for different tasks.

## 3.3 Research Methodology

This study adopts a comprehensive comparative experimental design to systematically evaluate the capabilities of GPT-4.5 and LLaMA 4 Herd with the help of the LLM-Lens framework, introduced in chapter 4 of this study, across a diverse set of categories and sub categories, reflecting both foundational competencies and real-world applicability. The evaluation focuses on three core aspects that are technical comparison, Ethical and Bias Handling and performance analysis via well-established Benchmarks and Prompt-Based Testing, each one performed for its relevance to contemporary use cases in a variety of applications across a number of fields. Informed by various prior, these tasks collectively test linguistic fluency, reasoning ability and contextual understanding. By synthesizing results from both structured benchmarks and expert judgment, the study aims to offer a transparent, reproducible and practically meaningful comparison between GPT-4.5 and LLaMA 4 Herd. The ultimate goal is to identify strengths, limitations and deployment trade-offs that inform future LLM development and adoption strategies.

## 3.4 Scope of the Study

This study focus to provide a comprehensive comparative analysis of two advanced LLM families which are OpenAI's GPT-4.5 and Meta AI's LLaMA 4 Herd (specifically the Scout, Maverick and Behemoth variants). The scope encompasses three major dimensions: technical architecture**,** functional performance and ethical and safety alignment**.** From a technical perspective, the research explores architectural foundations such as dense versus sparse (Mixture-of-Experts) designs, parameter scaling, context length, modality support and training regimes. On the performance front, the study evaluates both models using standard benchmark datasets (e.g., MMLU, GPQA, GSM8K, MATH-500, SWE-Bench) as well as custom prompt-based testing scored through a rubric across criteria like accuracy, relevance, clarity, reasoning and ethical soundness. Ethically, the study investigates hallucination rates, bias asymmetry, safety refusal behaviour and model transparency.

## 3.5 Techniques Used

To conduct a rigorous comparative analysis between GPT-4.5 and the LLaMA 4 Herd, this study employed a multi-pronged methodology combining prompt-based experimentation, benchmark reference analysis, qualitative feature comparison and ethical testing. These techniques were selected to ensure a holistic evaluation across technical, behavioral and practical dimensions.

### 3.5.1 Prompt-Based Testing

The core of this research involved prompt-based testing of both GPT-4.5 and the LLaMA 4 Herd. A custom-designed prompt dataset was created to evaluate key competencies across five categories: reasoning, code generation, creativity, multilingual understanding and ethical behavior. Each prompt was administered to both models under identical conditions using their respective platforms—OpenAI's ChatGPT web interface for GPT-4.5 and the Groq Console and Hugging Face API for LLaMA 4 variants. The generated outputs were then collected, analyzed and scored based on dimensions such as clarity, relevance, depth, fluency and safety. This hands-on evaluation enabled direct observation of each model's behavior and strengths in real-time usage scenarios.

### 3.5.2 Benchmark Reference Analysis

To complement the experimental prompt-based testing, this study incorporated quantitative performance data from publicly available LLM benchmarks. Standard datasets such as MMLU (Massive Multitask Language Understanding), GSM8K (mathematical problem-solving), HumanEval (code generation) and ARC/HELLASWAG (reasoning and commonsense tasks) were referenced. Performance metrics from these benchmarks were collected from sources such as the Open LLM Leaderboards, model cards and third-party evaluation reports. These benchmark results provided a reliable reference point for understanding how each model performs under standardized test conditions and helped contextualize the findings from custom testing.

### 3.5.3 Technical and Functional Analysis

To analyse the structural and operational distinctions between GPT-4.5 and LLaMA 4 models, a qualitative comparison was conducted based on publicly available technical

documentations, system cards, model cards and official blog releases. This included examining architectural aspects such as parameter count, activation strategies, memory mechanisms, attention types, context window capacity and support for modalities like text, vision and audio. The comparison also considered model deployment format (open-source vs. proprietary), API/interface access and tooling support. Functional attributes such as multilingual capabilities, long-context processing and tool-use readiness were evaluated using official feature descriptions and community testing insights. This technical and functional overview provides foundational context for understanding the design philosophies and implementation trade-offs behind each model.

### 3.5.4 Ethical Testing and Bias handling

Ethical safety and bias management were assessed through a qualitative review of each model's alignment strategies, refusal behaviour, hallucination tendencies, and ideological neutrality. GPT-4.5's alignment pipeline, which RLHF and safety training through red teaming, was compared against Meta's use of preference modeling, iterative safety fine-tuning, and safety classifiers like Llama Guard. Reported hallucination rates, toxic content filtering, and behaviour on sensitive prompts were referenced from system cards and third-party evaluations. Special attention was given to political asymmetry, factuality, and cultural inclusivity, using both self-reported metrics and leaderboard trends. This technique enabled a comparative lens into how each model approaches responsible AI behaviour and user safety.

## 3.6 Tools Used

To facilitate the design, execution and evaluation of the comparative study between GPT-4.5 and LLaMA 4 Herd, a range of tools and platforms are utilized. These tools support data preprocessing, model interaction, output evaluation and result visualization.

### 3.6.1 ChatGPT Interface

The ChatGPT interface (available at chat.openai.com) was used as the primary platform to interact with OpenAI's GPT-4.5 model. This web-based interface provides a conversational environment where users can engage with the model in a natural, dialogue-like format. It supports chat history, dynamic follow-up prompts and

contextual memory within a session, making it suitable for evaluating coherence, reasoning, ethical alignment and general user experience. The interface was chosen for its accessibility and realistic simulation of end-user interactions with the model.

### 3.6.2 Groq Console

The Groq Console was utilized to interact with the LLaMA 4 Herd models, specifically variants like LLaMA 4 Scout and Maverick. This interface provides high-speed inference by running models on Groq's custom-built Language Processing Units (LPUs), enabling real-time, low-latency responses. The Groq Console allows users to input prompts, view outputs instantly and test models in a controlled environment without requiring direct integration with APIs. It was particularly useful for evaluating the performance and responsiveness of LLaMA 4 models in interactive settings.

### 3.6.3 Hugging Face

The Hugging Face platform served as an additional environment for evaluating LLaMA 4 and other open-source models. It offers hosted inference endpoints and interactive model demos through a browser interface, allowing researchers to run prompts without local deployment. The platform also provides detailed model documentation, leaderboard rankings and benchmark comparisons, which were instrumental in cross-validating results and understanding model capabilities. Its accessibility and broad model availability made it a valuable tool for side-by-side evaluation and reproducibility.

### 3.6.4 Hugging Face LLM Leaderboards

LLM Leaderboards are online platforms that evaluate and rank LLMs based on standardized benchmarks and real-world tasks. These leaderboards, such as those hosted by Hugging Face, LMSYS, or HELM, allow researchers to compare model performance across areas like reasoning, code generation, multilingual understanding and ethical response handling. They provide objective insights using metrics like accuracy, helpfulness, toxicity and bias detection. In this research, LLM leaderboards were referenced as tools to supplement manual evaluation, verify model capabilities and validate performance claims from model providers like OpenAI and Meta.

### 3.6.5 Matplotlib

The popular data visualization library in Python, Matplotlib is one of Python libraries used for data visualization in a number of interactive formats. It provides an integrated and flexible set of plotting tools, allowing users to create a range of charts and plots. Matplotlib is widely used in domains like data analysis, data representation other fields for data exploration and visualization.

### 3.6.6 Rubric Score Sheet

A rubric score sheet is an evaluation tool that outlines specific criteria and performance levels to assess a task or product in a consistent and objective manner. It typically includes a list of evaluation dimensions, such as accuracy, clarity, creativity, or completeness, each with defined scoring scales (e.g., 1 to 5). Rubric score sheets are widely used in academic, research and professional settings to ensure transparent grading, structured feedback and comparative analysis. By breaking down a complex task into measurable components, a rubric helps evaluators make fair and replicable judgments, especially when assessing subjective or qualitative outcomes.

# Chapter 4

# LLM-Lens

This chapter introduces the LLM-Lens framework that is used to evaluate the LLMs from several different angles. The chapter has been partitioned into four different sections. Section 4.1 contains the proposed framework and its flow. LLM Models used in the study are described in the Section 4.2. Section 4.3 presents the analysis and results on various different aspects.

## 4.1 LLM-Lens

The LLM-Lens is a multi-dimensional comparison framework which focuses on comparing the two largest LLMs claiming to be the State-of-Art largest generational models, on a variety of tasks and on a variety of different aspects, from technical to ethical and architectural to performance. The LLM-Lens aims to put out a fair evaluation environment between the two models to get to know the strong and weak fields of each model over several domains and the ethical and bias handling nature of both models. The comparison data is prepared from official and credible data sources like OpenAI research papers, LLaMA official documentation, well establish LLM leaderboards for different datasets like MMLU, GSM8k, SAT, HarmlessEval etc. that tests the ability of the model over various aspects from technical to performance based. The models are also evaluated on a small custom made prompt dataset. The proposed framework's Flow:

1. Data on technical, functional and ethical/bias handling is collected from official papers.
2. Benchmarks results are collected from model cards, leaderboards and third-party analyses (LMSYS, Hugging Face, official blogs) and are grouped under different competencies (reasoning, math, coding, ethics).
3. Each model's performance tabulated and analyzed within and across categories.
4. Identifying the strengths and limitations of both models across categories.
5. Paragraph analysis interpreting other aspects like ethical/bias handling architecture difference, functionality difference.
6. Best model in different categories are than stated.
7. Prepared a custom prompt dataset for evaluation on small scale.

8. Evaluated the scores in different criteria.
9. Results, observations, key findings are documented for the selected best model.

## 4.2 LLMs used in the study

Two different models have been used to perform the analysis, Gpt-4.5 and LLaMA 4 herd. These model's strength the limitations vary from domain to domain, are even on par in some. Each model is deeply studied to get the facts as reliable and credible as possible.

### 4.2.1 GPT-4.5

GPT-4.5 (released as research preview) is an advanced LLM which is developed by OpenAI as an bridging model between GPT-4 and future models like GPT-5, sometimes referred as GPT-4 turbo or Orion. GPT-4.5 was released in February 2025 and is built upon the strengths of its predecessor model GPT-4, offering improved factual accuracy, reduced hallucinations and more natural, human alike, conversational ability. The specific architectural details have not been disclosed by the OpenAI but it is believed that it is also a dense transformer type architecture model which is trained on diverse datasets with different techniques like supervised learning and RLHF. It is only available to ChatGPT pro users and was released as a research preview for feedback purposes for further developments.

### 4.2.2 LLaMA 4 herd

The LLaMA 4 herd is the family of the three newly released LLaMA models by Meta AI in April 2025. These models represent a significant leap in evolution of the Meta's open-weight LLM offerings by leveraging the advanced MoE (Mixture-of-Experts) architecture for scalability and efficiency. The three models named as Scout, Maverick, released and upcoming Behemoth share the same architecture with different parameter counts. The LLaMA herd models are natively multimodal, being capable of processing text, images and videos and were trained on more than 30 trillion token across 200+ languages. These models are released under Meta's community license and are open source for use and testing, supporting long-context processing and balanced responses.

## 4.3 Analysis and Results

This sub section contains all the comparisons performed in this study. This sub section is divided into three sub-sub-sections. Section 4.3.1 represents the technical analysis. Section 4.3.2 describes functional and ethical analysis. Section 4.3.3 has the performance analysis, while section 4.3.4 presents the prompt-based testing and its observations.

### 4.3.1 Technical Analysis

The technical analysis highlights the differences between architecture type, parameter count, Tokenizing and vocabulary, context window length and instruction tuning and alignment.

**a) Architecture Type**

**GPT-4.5:** The exact design is proprietary, but the evidences available suggest that it continues OpenAI's use of dense transformer models. Some experts estimates that GPT-4.5 may be a multi-trillion- parameter sparse model with ~5-7 trillion parameters with ~600 billion active parameters just like a MoE architecture type. The official OpenAI description emphasizes "scalable techniques" and heavy compute for unsupervised learning being consistent with a dense or sparse Transformer. Trained on various alignment training techniques like Unsupervised pretraining, supervised fine tuning, RLHF and instruction hierarchy said to be trained upon a variety of data such as public, licensed and synthetic data, including data from smaller models too (OpenAI, 2025).

**LLaMA 4 (Scout, Maverick and Behemoth):** All the three variants are built upon Mixture-of-Experts (MoE) Transformer architecture. In MoE, a single token only activates a small fraction of the total parameters, as a result, while all the parameters are stored in memory, only a subset of the total parameters are activated while serving these models. With the improving of inference efficiency it also lowers the model seving cost and latency. MoE architectures are more compute efficient for training, inference and deliver higher quality compared to a dense model. Scout and maverick are explicitly described as MoE models that are based on a larger MoE teacher Behemoth which is yet to be released. This is because both Scout and Maverick is distilled from the upcoming LLaMA 4 Behemoth. The LLaMA 4 Scout uses a full

MoE structure where LLaMA 4 Maverick can switch between MoE and dense transformer (Meta AI, 2025).

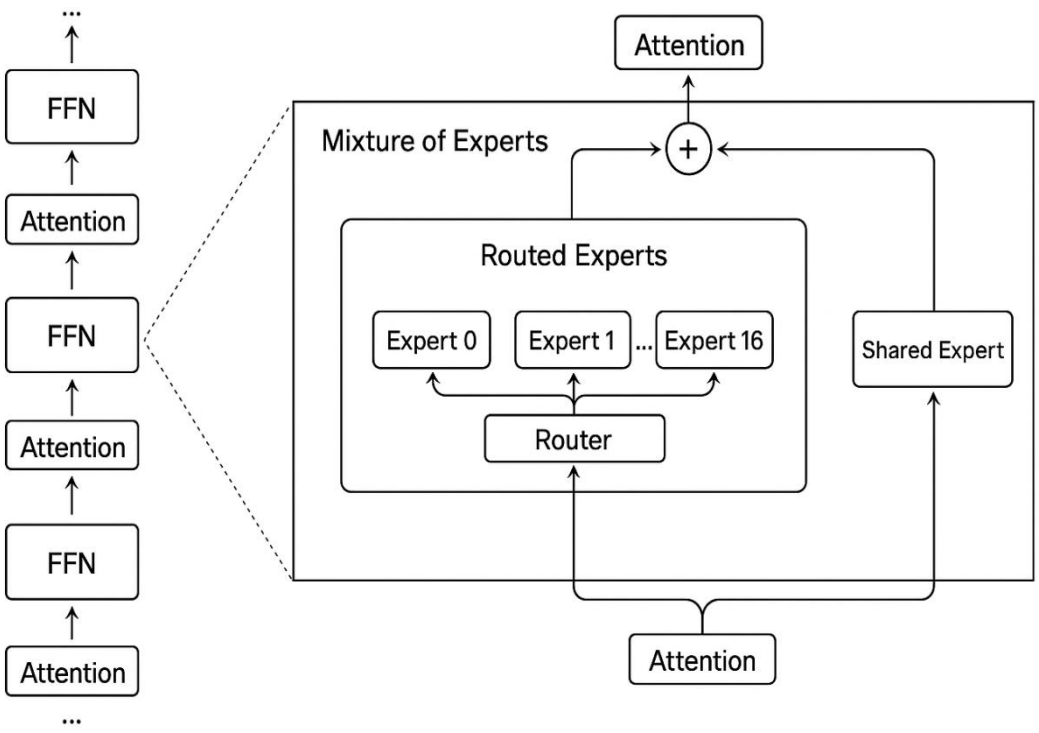


**Fig 4.1 Mixture-of-Experts Architecture** (Meta AI, 2025)

**b) Parameters Count**

**GPT-4.5:** Official numbers haven't made public but the estimated number of parameters sumps to ~7-10 trillion.

**LLaMA 4 Scout:** It has 109 Billion parameters with 17 billion as active. It has 16 experts and is designed to fit on a single GPU. This variant uses a sparse MoE to keep active size low.

**LLaMA 4 Maverick:** It has ≈400 Billion parameters with 17 billion as active. It has 128 experts. Only ~17 billion parameters are used per forward pass (via MoE routing) which makes it efficient relative to total scale.

**LLaMA 4 Behemoth:** It has ≈2 Billion parameters with 17 billion as active. It has 16 experts. This model is still in training, serving mainly as a high capacity teacher for now (Meta AI, 2025).

**c) Tokenizer and Vocabulary**

**GPT-4.5:** OpenAI's GPT series usually uses subword tokenization (byte Pair Encoding) where as GPT-4 and GPT-3.5 uses the "cl100k_base" encoding for the API. GPT-4.5 is liked to be using the similar BPE tokenizing (the same 100k-size cl100k or maybe an updates version), though the numbers have not been made public. GPT-4.5 lacks the native-multimodality while recent previous versions like GPT-4o have performed amazingly on tasks like text-to-image generation (Yan et al.2025).

**LLaMA 4:** LLaMA 4 follows Meta's LLaMA family convention. Earlier LLaMA models used a 32k BPE vocabulary. Meta's official documents mention "native multimodality" , they haven't stated the vocab size explicitly but LLaMA 4 likely uses a SentencePiece BPE tokenizer with a similar vocabulary size because Meta has b=not made any announcement about aby change. Importantly because LLaMA 4 is multimodal, its tokenizer also includes image and video token vocabularies for raw visual inputs.

**d) Context window length**

**GPR-4.5**: Though it has not been officially specified but it is likely that GPT-4.5's preview also supports 32k tokens at default and goes up to 128k tokens, similar to the GPT-4o variant. It suggests being very capable but has not stated a larger context, though the GPT-4.5's introductions highlights "context and knowledge" improvements not concrete limit was given (OpenAI, 2025).

**LLaMA 4 Scout:** Scout has 10 million tokens which are instantiated after instruct-tuning. This is an industry-record context window, enabled by the new "iRope" technique. RoPE or Rotatory Positional Encoding is an elegant approach to handle the positional information in transformer models. Instead of telling the model "this is position 5" of the token, it is a way to encode "where" a token is in sequence. In practice, Scout's 10 million token limit can be reached under heavy quantization on a single GPU like H100 (Hugging Face, 2025). The LLaMA 3 (8B, 70B) also provided efficient inference on 4-bit quantization while leading to degrading on 2-bit (Huang et al. 2024).

**LLaMA 4 Maverick:** 1,000,000 tokens (post tuning). Without the tuning, both Scout and Maverick utilizes a 256k- token window, but the fine tuning leads to an extension in the context window length (Hugging Face, 2025).

**LLaMA 4 Behemoth:** The pretraining use 256k context window. Similar or slightly higher context is expected after tuning, likely up to 1 million or so.

**e) Instruction Tuning & Alignment**

**GPT-4.5:** It continues using the OpenAI's alignment process. After the unsupervised pertaining, it is instruction –tuned by using RLHF (Reinforcement Learning from Human Feedback) or similar methods as done for previous models. The release notes highlight improvements like reduced hallucination and emotional intelligence, which is likely to be come from post-training alignment. It follows the same pattern of supervised fine-tuning + preference based RL to optimize responses as implemented in the previous models.

**LLaMA 4:** Meta has released instruction-tuned variants like "Llama -4-Scout-17B-16e-instruct' on Hugging face. The model card notes extensive safety fine tuning by the usage of human and synthetic data to reduce harmful outputs, refine refusal style and improve prompt steering which include training with human preference and toxicity filters, made more steerable by refining system prompt behavior. In shorts words, LLaMA 4 has in-built alignment and offers chat variants for developing purposes.

## Table 4.1 Technical Architecture and Training Comparison

| **Aspect** | **GPT-4.5 (Orion)** | **LLaMA 4 (Scout, Mavrick, Behemoth)** |
|---|---|---|
| **Developer** | OpenAI | Meta AI |
| **Release Date** | February 27, 2025 | April 5, 2025 (Scout & Maverick); Behemoth pending |
| **Model Type** | Dense Transformer | Mixture-of-Experts (MoE) Transformer |
| **Architecture** | Dense Transformer | MoE with 16–128 experts per layer |
| **Parameter Count** | Estimated 12.8 trillion | Scout: 109B total / 17B active<br>Maverick: 400B total / 17B active<br>Behemoth: ~2T total / 288B active |
| **Active Parameters** | Estimated ~600B per token | Scout & Maverick: 17B per token<br>Behemoth: 288B per token |

| | | |
|---|---|---|
| **Context Length** | 128,000 tokens | Scout: 10 million tokens Maverick: 1 million tokens |
| **Tokenizer** | Proprietary (likely BPE) | Likely SentencePiece or BPE (not explicitly disclosed) |
| **Multimodal Support** | Yes (text, image, file uploads) | Yes (native support for text, image, video) |
| **Training Data** | Public, licensed and synthetic data; includes data from smaller models | 200 languages with over 100 languages having 1 billion tokens each |
| **Training Techniques** | Unsupervised pretraining, supervised fine-tuning, RLHF, instruction hierarchy | Pretraining with MoE architecture; early fusion for multimodality; iRoPE for context scaling; FP8 training efficiency |

### 4.3.2 Ethical, Safety & Bias Handling

This section analyzes both models on aspects like bias mitigation, Toxic content avoidance, Fairness & inclusivity, Sensitive/controversial topics.

**a) Bias Mitigation**

**GPT-4.5:** It is trained with supervised fine-tuning and RLHF to align outputs which is similar to GPT-4o. OpenAI has not reported any significance increase in safety or bias risk relative to GPT-4o. In fairness/bias tests (BBQ benchmark), Gpt-4.5 performs on par with GPT-4o (OpenAI, 2025).

**LLaMA 4:** Meta has applied safety-focused training by using their "GOAT" tester to spot and reduce biases. They emphasize neutrality by declining the polarized prompts more evenly and articulating both sides of issues. The LLaMA 4's political lean is half of LLaMA 3.3's which indicated a reduction in ideological bias (Business Insider, 2025).

**b) Toxic Content Avoidance**

**GPT-4.5:** OpenAI's safety evaluations show GPT-4.5 refusing disallowed (toxic/illegal) requests at rates comparable to or slightly better than GPT-4. For example, it passes standard refusal tests ~99% of the time. (OpenAI continues to use content filtering and RLHF to avoid toxic outputs) (OpenAI, 2025).

**LLaMA 4:** Meta has not published detailed toxic-content metrics. The LLaMA 4 models are open and rely on Meta's community license and Acceptable Use Policy for downstream safety. However, Meta did use the GOAT framework during training to uncover and mitigate harmful outputs. (Open release means users must apply their own content safeguards in deployment) (Meta AI, 2025).

**c) Fairness & inclusivity**

**GPT-4.5:** When evaluated on benchmarks like BBQ (Bias Benchmark for QA), GPT-4.5 gave correct unbiased answers about as often as GPT-4o. OpenAI designs ChatGPT interface to treat users consistently and avoid protected biases, though detailed inclusivity training strategies have not been made public (OpenAI, 2025).

**LLaMA 4:** Meta specifically tuned LLaMA 4 to be more inclusive of viewpoints. They report only ~1% of test questions had asymmetric answers (one side answered, the other refused). LLaMA 4 was trained to present factual information "without judgment" and covers multiple perspectives, aiming to reduce systemic biases known in earlier models.

**d) Sensitive/controversial topics**

**GPT-4.5:** GPT-4.5 is intended to be helpful yet safe on sensitive topics. OpenAI states it has "stronger alignment with user intent" and "fewer hallucinations", aiming for factual, balanced replies (e.g. it "excels at factual accuracy". System cards indicate GPT-4.5 underwent rigorous safety/red-team testing before release (Business Insider, 2025).

**LLaMA 4:** LLaMA 4 was explicitly tuned to handle contentious subjects more often. According to Meta's public statements, LLaMA 4 is more willing to address controversial content and is more balanced than earlier Llama versions. Meta says LLaMA 4 now *refuses* controversial political/social prompts only ~2% of the time (versus ~7% under Llama 3). The models were tested on "debated" questions and are designed to answer both sides equally. It has been noted that LLaMA 4 now "provides helpful, factual responses … and doesn't favour some views over others" (Meta AI, 2025) (Business Insider, 2025).

### 4.3.3 Performance Analysis

This section represents the performance analysis of both models on different categorical aspects like Reasoning, Mathematics, Coding, Multilingualism, Factuality metrics and hallucination rates.

**a) Reasoning**

General reasoning is tested upon datasets like Massive Multitask Language Understanding (MMLU), which is a dataset that measures general knowledge across 57 different languages and ideal for AI systems that are multifaceted and require extensive world knowledge and problem solving ability and used to assess the LLM's understanding and reasoning in a wide range of subject) and GPQA (Good Practices in Quality assurance, which is a multiple-choice Q&A dataset of tough questions that are written and validated by experts in chemistry, physics and biology, containing 448 multiple-choice questions.

GPT-4.5 scored 85.1% on the multilingual MMLU benchmark (zero-shot accuracy) and 71.4% on GPQA (a challenging science QA set). In Meta's internal tests, Llama 4 Scout (17B) scored 74.3% on MMLU and 57.2% on GPQA, while Maverick (17B w/ MoE) scored 80.5% on MMLU and 69.8% on GPQA. The largest Llama 4 (Behemoth, 402B active MoE) achieved 82.2% on MMLU Pro and 73.7% on GPQA.

**Table 4.2 GPT-4.5 and LLaMA 4 MMLU and GPQA Scores**

| Model | MMLU (Accuracy) | GPQA (Science QA) |
|---|---|---|
| GPT-4.5 | 85% | 71.4% |
| Llama 4 Scout | 74.3% | 57.2% |

| | | |
|---|---|---|
| Llama 4 Maverick | 80.5% | 69.8% |
| Llama 4 Behemoth | 82.2% | 73.7% |

As it can be seen that GPT-4.5 leads on these broad knowledge benchmarks, but Llama 4 variants are also not far as Maverick and Behemoth narrows the gap. GPT-4.5's advantage is clear in zero-shot reasoning. All Llama 4 models exceed previous open models. For instance, Llama 3.3 at 70B was at ~73% on MMLU-pro. Though it can be seen clearly that Llama 4 behemoth is surely approaching the GPT-4.5's level on knowledge tasks (OpenAI, 2025) (Meta AI, 2025).

**b) Mathematics**

Benchmarks: Math reasoning is tested by datasets like GSM8K (grade-school math), SAT Math, AIME (U.S. Math Olympiad) and the new MATH-500 benchmark. OpenAI reports GPT-4.5 scored 36.7% on the very hard AIME '24 exam (OpenAI, 2025). In contrast, Meta reports Llama 4 Behemoth scored 95.0% on a proprietary "MATH-500" benchmark (a 500-question advanced math set)(Meta AI, 2025), significantly ahead of competitors. No public GSM8K or SAT scores are available yet for Llama 4. In summary, GPT-4.5 made only modest gains in advanced math (36.7% AIME), whereas the gigantic Behemoth model excels on STEM/math tests (95.0% on MATH-500). No direct comparisons exist for Scout/Maverick on these math benchmarks (Meta AI, 2025) (OpenAI, 2025).

**Table 4.3 GPT-4.5 and LLaMA 4 AIME'24 and MATH-500 Scores**

| **Model** | **AIME'24** | **MATH-500** |
|---|---|---|
| GPT-4.5 | 36.7% | - |
| Llama 4 Scout, Maverick, Behemoth | - | -, -, 95.0% |

For reference, GPT-4 (2023) scored around 58–60% on the MATH dataset (AMC/AIME problems) with chain-of-thought prompting, far above GPT-3.5 (Bito.ai, 2024). (GPT-4.5's lower AIME performance suggests it traded some systematic math skill for other capabilities.) Llama 4's outstanding MATH-500 score

suggests that at extreme scale, its mixture-of-experts design yields very strong STEM reasoning.

**c) Code Generation**

Code Generation has been tested on the LiveCodeBench Benchmark platform is an evaluation platform of LLMs for code which is a holistic and containment free and continuously collects new problems over time. Not only code generation it also focuses on broader capabilities such as code execution, text output prediction and self-repair.

GPT-4.5 internal tests show moderate coding gains: on one coding contest ("SWE-Lancer Diamond"), GPT-4.5 achieved 32.6% pass rate vs 23.3% for GPT-4o. The GPT-4 also dominated the LoRaWAN tasks on lower temperatures (Fernandes et al. 2025), though we don't have any official temperature value info and testing but on another coding suite like "SWE-Bench", GPT-4.5 scored 38.0% vs 30.7%. These indicate improvements; although GPT-4.5 sometimes trails specialized code models (GPT-3.5 had 61.0% on SWE-Bench vs GPT-4.5's 38.0%). Meta's benchmarks use its own "LiveCodeBench" (live coding tasks). Llama 4 Scout scored 32.8% on LiveCodeBench (nearly equal to GPT-4o), Maverick scored 43.4% and Behemoth 49.4%. All these exceed GPT-4o's ~32%. A comparison table is below. (No official HumanEval/MBPP scores have been released yet for Llama 4 or GPT-4.5, so we rely on these reported contest-style results)(Meta AI, 2025) (OpenAI, 2025).

**Table 4.4 GPT-4.5 and LLaMA 4 LiveCodeBench Scores**

| Model | LiveCodeBench (coding pass rate) | SWE-Lancer Diamond | SWE-Bench |
|---|---|---|---|
| GPT-4.5 | - | 32.6% | 38.0% |
| Llama 4 Scout | 32.8% | - | - |
| Llama 4 Maverick | 43.4% | - | - |
| Llama 4 Behemoth | 49.4% | - | - |

In summary, GPT-4.5's code performance is improved but somewhat uneven (it unexpectedly underperformed GPT-3.5 on one contest). The Llama 4 models scale better: even the small Scout matches GPT-4o, while Maverick/Behemoth outscore it substantially. Previous GPT-4 scored ~73% on HumanEval with chain-of-thought

prompting (Bito.ai, 2024), so GPT-4.5 is likely similar or better. The bottom line: on real-world coding and STEM problem solving, Llama 4 (especially Maverick and Behemoth) appears very strong, rivalling GPT-4.5, though direct head-to-head data is still emerging.

**d) Multilingualism**

Being Multilingual of a LLM is becoming more and more important as they are being used all over the world for different tasks in different regions and ethnicity. Multilingualism is tested upon the MMMLU dataset which is multilingual version of the MMLU dataset. The GPT-4.5 scores a high 85.1% which beats the Llama 4 Maverick (having 84.6% in MMMLU) but Llama 4 Behemoth has a slight upper hand here as it wins the aspect by getting 85.8% (Meta AI, 2025) (OpenAI, 2025).

**Table 4.5 GPT-4.5 and LLaMA 4 MMMLU Scores**

| **Model** | **MMMLU** |
|---|---|
| GPT-4.5 | 85.1% |
| Llama 4 Scout | - |
| Llama 4 Maverick | 84.6% |
| Llama 4 Behemoth | 85.8% |

Llama 4 Behemoth surely won here by a minimal distant but GPT-4.5 still gets ahead of Llama 4 Maverick. There are no official Llama 4 Scout MMMLU test scores for yet.

e) **Factuality Metrics and Hallucination rates**

Hallucination, in context of LLMs, refers to times where the model produces/generates text that is incorrect factually, nonsensical or inconsistent with the input data. Hallucination rates are the measures of how often a model produces false or made-up information. OpenAI reports that GPT-4.5 halved its factual error rate on knowledge questions (down to 37.1% from GPT-4o's ~61.8%). On the SimpleQA factual knowledge test, GPT-4.5 scored 62.5%, far above GPT-4o. This confirms OpenAI's claim that GPT-4.5 "excels at factual accuracy". Meta has not published comparable hallucination rates for Llama 4 (OpenAI, 2025).

Though we don't have any official numbers or scores, the two models have been tested on the Vectara's HHEM-2.1 (Hughes Hallucination Evaluation Model) dataset. We have the results in the table below.

**Table 4.6 GPT-4.5 and LLaMA 4's factuality and Hallucination rates**

| **Model** | **Hallucination Rate (lower is better)** | **Factual Consistency rate (higher is better)** | **Answer rate (higher is better)** | **Average Summary Length (words)** |
|---|---|---|---|---|
| GPT-4.5 | 1.2% | 98.8% | 100.0% | 77.0 |
| Llama 4 Scout | 4.7% | 95.3% | 100.0% | 80.7 |
| Llama 4 Maverick | 4.6% | 95.4% | 100.0% | 84.8 |

There are results for Meta Llama 4 Behemoth because it is still in training. The Scores surely shows that GPT-4.5 is better in giving factual data when asked a query. The Llama 4 models are falling behind than GPT-4.5 in both hallucination rate and Factual Consistency rate by a significant margin.

**f) Emotional Intelligence and Social Reasoning**

Benchmarks: Tasks like EQ-Bench and SocialIQa evaluate empathy, social reasoning and emotional awareness. OpenAI claims GPT-4.5 has greater social "EQ" than previous models, making it better at pacing conversations and responding to emotions. In practice, GPT-4.5 performs well on social cues (the demo prompts show it inviting further conversation appropriately). Meta has not released Llama 4 scores on EQ-Bench or SocialIQa, but crowdsourced comparisons suggest Llama 4 is strong. In the LMSYS Chatbot Arena (which implicitly tests social fluency), Llama 4 Maverick achieved an Elo rating only slightly below GPT-4 Turbo, indicating very competent conversational-social behavior. (By comparison, prior Llama 3.3 70B was far lower). In summary, both models are highly capable socially: GPT-4.5 emphasizes emotional intelligence by design, while fine-tuned Llama 4 conversational variants are already beating most competitors in open chat tests. Formal benchmark scores are pending (OpenAI, 2025)(Medium, 2024).

### 4.3.4 Prompt-Based Testing

In this section, we performed Prompt-Based testing on a small scale. We prepared a custom made small dataset of 50 prompts where each aspect like Logical reasoning, Code generation, Creativity and Writing, Ethical/ bias handling and Multilingualism had 10 prompts each (JODurmeeet, 2025). Further the models were evaluated on five different criteria like Accuracy, Relevance, Clarity, Chain-of-thoughts process, Ethical soundness with the help of a Rubric-Score sheet and were scored on a scale of 1 to 5 to maintain the easiness during the scoring and evaluation. As LLaMA 4 Behemoth is still in training and LLaMA 4 scout is already outperformed by LLaMA 4 Maverick, we compared the GPT-4.5 with LLaMA 4 Maverick. The Flow goes like:

1. Dataset was prepared using 50 custom made prompts for five different aspects
2. ChatGPT interface for GPT-4.5 and Groq Console for LLaMA 4 Maverick was used to get responses on the dataset.
3. Further the responses were recorded into the rubric-score sheet.
4. The responses were evaluated on various criteria and scored on a scale of 1 to 5.
5. The resulted scores were visualized using Matplotlib python library.

The results on different criteria are:

**a) Accuracy**

Accuracy was used as a criterion to check whether the facts and the responses generated by the models were correct or off from the actual ones or from what they should be instead.

The GPT-4.5 shows a slight upper hand in the accuracy criteria.

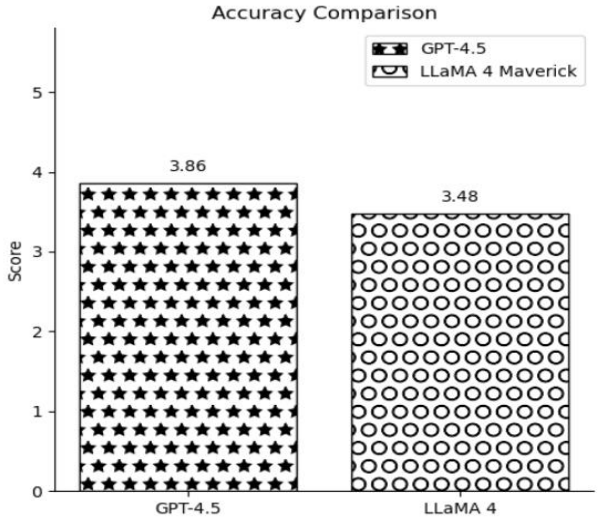


**Fig 4.2 Accuracy score of GPT-4.5 and LLaMA 4 Maverick**

**b) Relevance**

Relevance was used as a criterion to evaluate whether the responses generated by both models are relevant in regarding to the respective prompts. The GPT-4.5 here beats the LLaMA 4 Maverick in this criterion too.

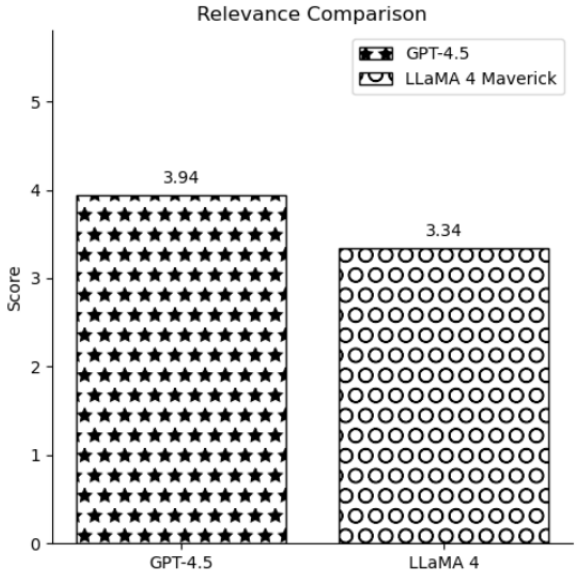


**Fig 4.3 Relevance score of GPT-4.5 and LLaMA 4 Maverick**

**c) Chain-of-Thoughts**

Chain-of-thoughts (CoT) was chosen as a criterion to see whether the models are giving a step-by-step explained response or not. Both models were on par GPT-4.5 leading by a negligible margin. CoT is a method that increases the reasoning abilities of LLMs (Wei et al. 2022).

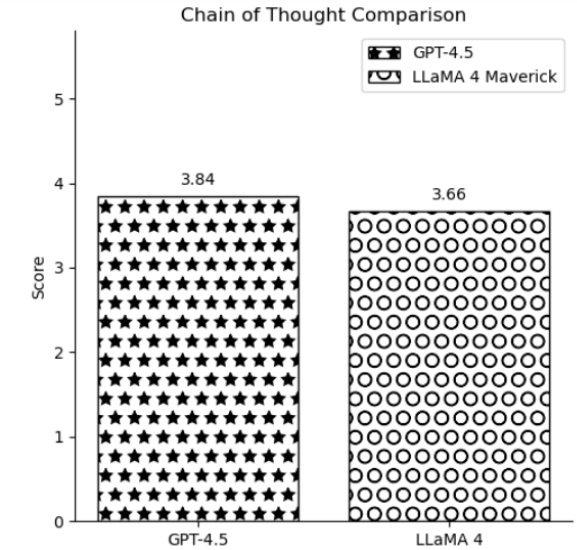


**Fig 4.4 Chain-of-Thoughts score of GPT-4.5 and LLaMA 4 Maverick**

**d) Ethical/Bias Handling**

The responses were further evaluated in the basis on ethical and bias scores. These scores are important to evaluate whether the model's responses are ethical and socially acceptable or are biased.

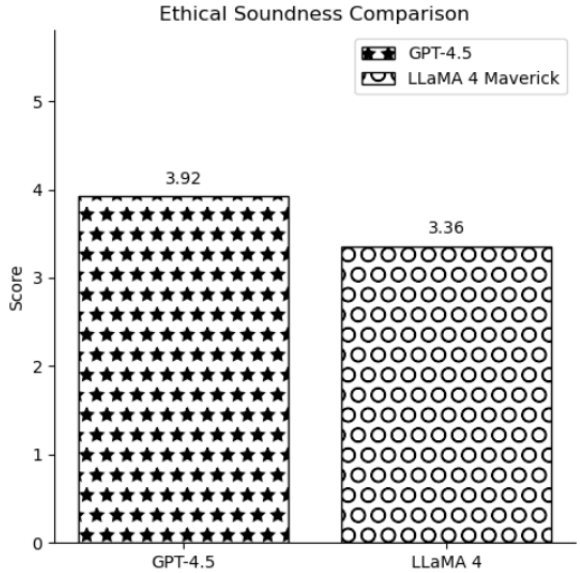


**Fig 4.5 Ethical/Bias score of GPT-4.5 and LLaMA 4 Maverick**

**e) Clarity of Response**

This criterion is responsible to evaluate either the answers to the prompts produced by the models are clear and easy to understand, is something missing or wrong. The LLaMA 4 Maverick is beaten by GPT-4.5 by a considerable margin.

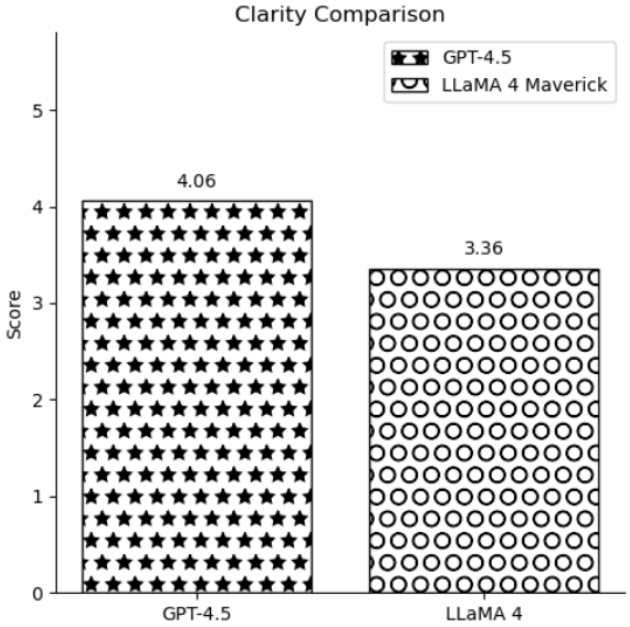


**Fig 4.6 Clarity score of GPT-4.5 and LLaMA 4 Maverick**

The differences between these scores were justified when we looked at some of the prompts responses. For example when both models were asked to translate “Thank you for your help” in Japanese, the responses were:


**Fig 4.7 Response of LLaMA 4 Maverick and GPT-4.5**

As it can be seen that LLaMA 4 Maverick provided an explanation but didn’t translated the text into the native Japanese language while GPT-4.5 gave a concise response providing the native translation also but no explanation regarding the use of the phrase.

These results are taken form the evaluation of a small dataset and surely don’t aims to depict the superior model from such a small scale evaluation, the models would have performed differently and more well if provided a dataset with enough number of prompts to assess every aspect deeply.

# Chapter 5

# Results and Conclusion

The chapter describes the suggested research discussion, conclusion and future scope. Three sections make up this chapter. Section 5.1 contains the results of the proposal. The discussion of the study is in Section 5.2. Section 5.3conatains the outcome/conclusion of the study. The future of this study, which will improve this area even more, is presented in section 5.4.

## 5.1 Results

The prompt-based testing showed some amusing outcomes and results. The LLaMA 4 (Maverick variant) was behind then GPT-4.5 by a significant margin. Though LLaMA claims several improvements and advancements in various aspects like political prompt articulation and multilingualism, the Small scale variants than Behemoth still need either more fine-tuning or more training. As the Behemoth is still in training a concrete decisions on the overall review can be put to hold but The LLaMA 4 Scout and Maverick still needs to be refined if compared to the Gigantic boss GPT-4.5.

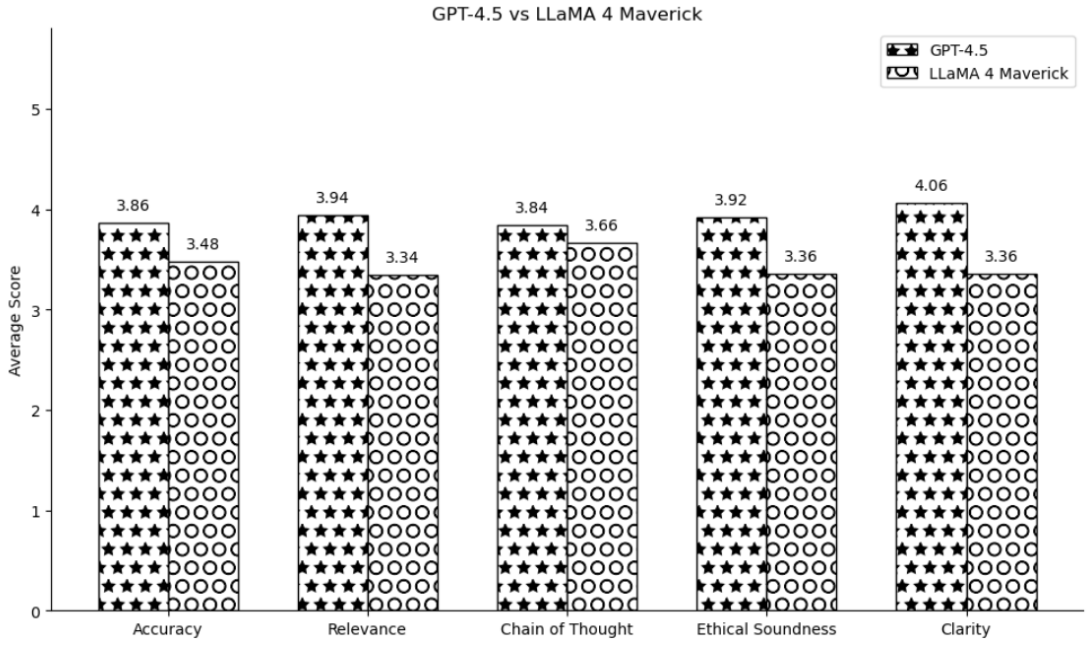


**Fig 5.1 GPT-4.5 vs LLaMA 4 Overall Comparison**

As it can be seen in the above figure, the LLaMA 4 Maverick variant may come close to GPT-4.5 in chain-of-thoughts processing and accuracy but is still outperformed in Clarity, Relevance and Ethical soundness. These results are from a small scale dataset with only 50 prompts in it. The performances may change when encountered with large scale data with a huge load of varieties. Below are some of the Key findings that came into sight from the different analysis performed in section 4.3 of chapter 4:

## Key Findings

This section contains the key findings that are found in the above performed analysis. These findings are distilled from the research and benchmark based comparative analysis of GPT-4.5 and LLaMA 4 (Scout, Maverick and Behemoth) highlighting the strengths, weaknesses and architectural distinctions:

1. GPT-4.5 shows superiority in factual accuracy by demonstrating a significant low in hallucination rates and high in factual precision. It achieves 62.5% on SimpleQA and reduced the factual error to 37.1% which makes it the most reliable model for the knowledge-based tasks.
2. LLaMA 4 Behemoth excels in advanced mathematical reasoning which is shown by its high score of 95.0% on MATH-500 which suppresses the GPT-4.5's 36.7% score on AIME'24 highlighting its strength in high-level math reasoning.
3. GPT-4.5 maintains its leadership in general reasoning with an surpassing score of 85.1% on MMLU and a strong performance on GPQA. GPT-4.5 remains the most capable model in zero-shot, multi-domain reasoning.
4. LLaMA models implement a sparse Mixture-of-Experts (MoE) architecture. It activates as few as 17 billion parameters per inference step which results in high efficiency while preserving performance, especially in Scout and Maverick.
5. LLaMA supports an unprecedented 10 million token window that enables effective processing of extremely long documents also enabling a longer conversation without losing the context.
6. GPT-4.5 is designed for human-like interactions and it excels in emotional intelligence, emotional dialogue, pacing and empathetic responses which gives it an edge in human-AI conversation.
7. GPT-4.5 is accessible only through OpenAI commercial platforms like ChatGPT Plus and OpenAI Playground APIs, limiting its research flexibility. On the other

hand, LLaMA modles are openly released under the Meta's community license, enabling local deployment and fine-tuning.

8. LLaMA Scout and Maverick are trained natively on text, images and video which offers multimodal understanding without the additional adapters but GPT-4.5 shows absence in these capabilities (but these were present in GPT-4o).
9. Both the models show almost on par performance to each other in multilingualism, Behemoth beating GPT-4.5 by a slight difference. It would be safe to say that both models are great for multilingual tasks.
10. Meta reports significant reductions in bias asymmetry. The Scout and Maverick now refuses politically polarized prompts equally and presents multiple viewpoints. This shows improvement upon the previous LLaMA 3 and matches the GPT-4-level neutrality.
11. The LLaMA 4 Scout can achieve up to 10 million token context window length with heavy quantization on a single GPU, normally it stays up to 256k token length.
12. The prompt-based testing shows that the Maverick variant still underperforms when compared to GPT-4.5.
13. Another key finding is that the response generation speed of GPT-4.5 is very slow, on the other hand the LLaMA 4 variants provides very fast inference speed.

## 5.2 Discussion

In this research, a framework called LLM-Lens was introduced for the analysis of two largest generation LLMs, OpenAI's GPT-4.5 and Meta AI's LLaMA 4 Herd (Scout, Maverick and Behemoth), was conducted to identify the model which is most effective for a variety of tasks across different categories. Both the models were evaluated on the basis on several different aspects including technical differences, functional differences, ethical and bias handling capabilities, toxic content avoidance, sensitive topic articulation, performance on different industry recognized datasets for different tasks across categories like math, reasoning, creativity and writing, code generation etc. which led to some amusing results and findings.

From the analysis, it is clearly visible that GPT-4.5 maintains an edge over LLaMA 4 in factual accuracy, conversational coherence, emotional intelligence and safety alignment. These attributes makes it highly reliable for general purpose AI tasks and applications. However its proprietary nature and limited accessibility restricts its

research and development usage. In contrast, LLaMA 4 models shows remarkable architectural innovation through their use of sparse Mixture-of-Experts (MoE), enabling efficiency at scale and support of long-context processing up to 10 million tokens which is quite a leap in conversational AI domain. While both AI families show improvements and advancements on the previous short-comings, the ecosystems they serve are different. GPT-4.5 prefers and provides robustness and safety within a controlled environment, whereas LLaMA 4 prefers and provides openness, scalability and user-directed customization.

The Prompt-based evaluation shows that the GPT-4.5 still is far ahead of the Meta's current LLaMA 4 family, though the picture can be different after the release of the LLaMA 4 Behemoth the pre-release scores surely shows that it is on par with GPT-4.5 even beating it at some aspects but for now GPT-4.5 is taking the crown.

## 5.3 Conclusion

This research project presented the study in a comprehensive way by the introduction and application of the LLM-Lens framework on the two massive scale LLMs said to be the most superior ones in their domain. The primary aim was to use LLM-Lens and identify each of the two model's strength and weakness across a variety of tasks among several different categories ranging from technical to emotional and functional to ethical. Both the models highlighted distinct strengths, design philosophies and usage implications for modern large language models. GPT-4.5 emerges as a highly aligned and refined model. The performance across reasoning tasks, emotional intelligence, safety refusal benchmarks and excellence in factual accuracy, ethical behavior and conversational depths shows the reaffirmation of the OpenAI's focus on user trust and high-stake deployment. In contrast, Meta's LLaMA 4 models brings in innovation through their efficient MoE architecture, massive context window lengths (10 million tokens) and open accessibility. The prompt-based testing performed in this study clearly shows at small scale how the models perform differently, in fact the small models like scout and Maverick are outperformed by GPT-4.5 easily, though it can also be due the difference in parameters count. The LLaMA 4 Behemoth further demonstrates the power of large scale by outperforming GPT-4.5 in advanced mathematical and coding benchmarks. Considering the results and findings, it is not wrong to say that the findings suggest that while GPT-4.5 is ideal for high-quality,

controlled applications, the LLaMA 4 Herd offers powerful capabilities for researchers, developers and organizations seeking custom, transparent and scalable AI systems. The study also aims to reinforce the view that no single model is superior universally, every model, when trained and tuned thoughtfully, excels within its own designs and constraints.

## 5.4 Future Scope

While this research successfully identified the best LLM in different categories based on several aspects like technical, functional and performance, the LLM-Lens framework can be used in various areas where further exploration can be done to enhance and expand the study.

1. Analysis can be further expanded for the performance comparison on tasks involving image and video inputs, especially when LLaMA 4 supports native multimodality and GPT-4.5 not.
2. To gauge the practicality of deploying each model in production environment, analyze latency, response time, memory usage and energy efficiency during live inference.
3. Investigation of the fine-tuning potential of LLaMA 4 variants for domain-specific application like medical or physical sciences domain can be done.
4. The LLM-Lens can be used to evaluate other LLMs like Gemini, Qwen, Grok, Mistral etc.
5. The dataset used to perform prompt-based testing can be turned into a large scale dataset and can be used to perform the prompt-based testing on a large scale and the model's score evaluation can be done on the obtained scores.

Thus, this research opens up a number of exciting possibilities for future work that can make LLM functions and applications for powerful, efficient, accurate and implementable.

# REFRENCES


[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[2] Bito.ai. (2024, July 26). *Gemini 1.5 Pro vs GPT-4 Turbo Benchmarks*. https://bito.ai/blog/gemini-1-5-pro-vs-gpt-4-turbo-benchmarks

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

[4] Business Insider. (2025, April 7). *Meta Llama 4 AI model Contentious Questions Woke 2025*. https://www.businessinsider.com/meta-llama-4-ai-model-contentious-questions-woke-2025-4

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

[6] Fernandes, D., Matos-Carvalho, J. P., Fernandes, C. M., & Fachada, N. (2025). DeepSeek-V3, GPT-4, Phi-4 and LLaMA-3.3 generate correct code for LoRaWAN-related engineering tasks. *arXiv preprint arXiv:2502.14926*.

[7] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... & Vasic, P. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[9] Huang, W., Zheng, X., Ma, X., Qin, H., Lv, C., Chen, H., ... & Magno, M. (2024). An empirical study of llama3 quantization: From llms to mllms. *Visual Intelligence*, *2*(1), 36.

[10] Hugging Face. (2025, April 5). *Welcome Llama 4 Maverick & Scout on Hugging Face*. https://huggingface.co/blog/llama4-release#welcome-llama-4-maverick--scout-on-hugging-face

[11] JODurmeeet. (2025, June 13). *Prompt-dataset-50* [GitHub repository]. GitHub. https://github.com/JODurmeeet/Prompt-dataset-50

[12] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems, 25*.

[14] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[15] Medium. (2025, April 7). *Meta's Llama 4 Maverick ranks #2 globally in Chatbot Arena, challenging GPT-4*. Medium. https://medium.com/artificial-synapse-media/metas-llama-4-maverick-ranks-2-globally-in-chatbot-arena-challenging-gpt-4

[16] Meta AI. (2025). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*. Meta. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[17] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J.(2024) ( Large language models: A survey. arXiv 2024. *arXiv preprint arXiv:2402.06196*.

[18] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

[19] OpenAI. (2024, May 13). *Hello GPT-4o*. OpenAI. https://openai.com/index/hello-gpt-4o/

[20] OpenAI. (2025, February 27). *Introducing GPT-4.5*. https://openai.com/index/introducing-gpt-4-5/

[21] OpenAI. (2025, February 27). *OpenAI GPT-4.5 System Card* [PDF]. https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf

[22] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

[23] Poldrack, R. A., Lu, T., & Beguš, G. (2023). AI-assisted coding: Experiments with GPT-4. *arXiv preprint arXiv:2304.13187*.

[24] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[25] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[27] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

[28] Yan, Z., Ye, J., Li, W., Huang, Z., Yuan, S., He, X., ... & Yuan, L. (2025). Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*.