Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model

Konstantinos I. Roumeliotis * , Nikolaos D. Tselikas , Dimitrios K. Nasiopoulos

Posted Date: 1 August 2023

doi: 10.20944/preprints202307.2142.v1

Keywords: llama 2; llama2; llama 2 projects; llama 2 model architecture; llama 2 fine-tuning

Article

Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model

Konstantinos I. Roumeliotis [1,*], Nikolaos D. Tselikas [1] and Dimitrios K. Nasiopoulos [2]

[1] Department of Informatics and Telecommunications, University of Peloponnese, Akadimaikou G. K. Vla-chou Street, 22 131 Tripoli, Greece; ntsel@uop.gr

[2] Department of Agribusiness and Supply Chain Management, School of Applied Economics and Social Sciences, Agricultural University of Athens, 118 55 Athens, Greece; dimnas@aua.gr

* Correspondence: k.roumeliotis@uop.gr; Tel.: +30-271-037-2216

Abstract: The rapidly evolving field of artificial intelligence (AI) continues to witness the introduction of innovative open-source pre-trained models, fostering advancements in various applications. One such model is Llama 2, an open-source pre-trained model released by Meta, which has garnered significant attention among early adopters. In addition to exploring the foundational elements of the Llama v2 model, this paper investigates how these early adopters leverage the capabilities of Llama 2 in their AI projects. Through a qualitative study, we delve into the perspectives, experiences, and strategies employed by early adopters to leverage Llama 2's capabilities. The findings shed light on the model's strengths, weaknesses, and areas of improvement, offering valuable insights for the AI community and Meta to enhance future model iterations. Additionally, we discuss the implications of Llama 2's adoption on the broader open-source AI landscape, addressing challenges and opportunities for developers and researchers in the pursuit of cutting-edge AI solutions. The present study constitutes an early exploration of the Llama 2 pre-trained model, holding promise as a foundational basis for forthcoming research investigations.

Keywords: Llama 2; Llama2; Llama 2 projects; Llama 2 model architecture; Llama 2 fine-tuning

1. Introduction

The rapid advancements in artificial intelligence (AI) have ushered in an era of groundbreaking open-source pre-trained models, empowering researchers and practitioners across diverse domains. One such model that has recently gained substantial attention is Llama 2, introduced by Meta, an industry leader in AI technology [1]. Llama 2 represents a compelling fusion of innovation and accessibility in the domain of Large Language Models (LLMs), rendering it an appealing option for pioneering users seeking to harness its potential in their artificial intelligence endeavors [2].

As AI technology continues to evolve, understanding the perspectives and experiences of early adopters in utilizing novel pre-trained models like Llama 2 becomes crucial. Early adopters play a pivotal role in the adoption and dissemination of these cutting-edge technologies, driving innovation and providing valuable insights for further model enhancements. Exploring their utilization patterns, challenges, and strategies can offer crucial guidance for both researchers and developers, fostering the effective application and development of AI solutions.

In this context, the present paper presents an early investigation into how early adopters are utilizing Meta's new open-source pre-trained model, Llama 2. Through qualitative research methods, we aim to capture the essence of their experiences and perceptions while employing Llama 2 in their AI projects. By delving into the practical application of Llama 2, we endeavor to identify its strengths, weaknesses, and potential areas for improvement.

Despite being publicly announced only 10 days ago, Llama 2 has attracted considerable attention from early adopters. In the short span between August 18, 2023, and August 28, 2023, these adopters

 doi:10.20944/preprints202307.2142.v1

have demonstrated successful implementation of various tasks, such as model deployment, chatbot development, fine-tuning in different languages, domain-specific chatbot creation (medical domain), parameter customization for CPU and GPU, and runtime efficiency optimization with limited resources.

This study contributes to the broader landscape of AI research by shedding light on the practical implications of adopting Llama 2 and the value it brings to AI applications. Furthermore, by highlighting the early adopters' perspectives, we seek to foster a better understanding of the model's impact on the development of AI technologies and the potential challenges faced in its utilization.

Four pivotal hypotheses are introduced, which will be substantiated through the out-comes of the experiments:

- Hypothesis 1. Research Objective: Research Objective: The research aims to assess the audience's response to Llama 2 and verify Meta's expectations that an open-source model will experience faster development compared to closed-source models [2].
- Hypothesis 2. Research Objective: The research aims to assess the challenges encountered by early adopters in deploying the Llama 2 model.
- Hypothesis 3. Research Objective: The research aims to assess the challenges encountered by early adopters in fine-tuning the Llama 2 model.
- Hypothesis 4. Research Objective: The research seeks to unveil that the medical domain consistently ranks among the primary domains that early adopters engage with, undertaking fine-tuning of models.

Section 2 provides a concise background and context concerning the evolution of language models, natural language processing (NLP), the transformer architecture, and supervised fine-tuning. In Section 3, the Llama 2 models and licensing are presented. Section 4 delves into the training process of Llama 2. In Section 5, we showcase the case studies and projects conducted by the early adopters of the Llama 2 model. Subsequently, we highlight the key findings derived from the analysis of early adopters' experiences, followed by a comprehensive discussion of the implications arising from the utilization of Llama 2. The paper concludes with recommendations for future research endeavors, emphasizing the significance of ongoing efforts to refine and optimize pretrained models like Llama 2 for the betterment of the AI community and humanity.

2. Background and Context

2.1. Evolution of Language Models

The journey of Language Models' evolution has been nothing short of extraordinary, surpassing all expectations and redefining the boundaries of artificial intelligence. Starting from simple rule-based systems, these models have undergone a remarkable transformation, culminating in the emergence of colossal models like GPT-3, boasting billions of parameters and demonstrating human-like fluency and creativity [3]. The relentless pursuit of innovation and research in natural language processing has ushered in a new era, where machines can comprehend and generate text with unparalleled accuracy and context-awareness. This evolution of language models has transformed language from a barrier into a bridge, connecting humans and machines in ways once deemed science fiction. As language models continue to progress, the line between human and artificial intelligence continues to blur, paving the way for a future where machines play an integral role in shaping how we communicate, learn, and interact with information [4].

The Evolution of Language Models marks a groundbreaking revolution, fundamentally altering how we perceive and interact with machines. What originated as simple statistical models has rapidly evolved into remarkably sophisticated neural architectures, capable of grasping the intricacies of human language and generating coherent responses rivaling human speech. The introduction of transformer-based models and innovative pre-training techniques has propelled language models to unprecedented heights, surpassing previous milestones and reshaping the landscape of AI applications [5]. Today, they power a multitude of real-world applications, from chatbots and virtual assistants that enhance customer experiences to language translation systems that facilitate global

 doi:10.20944/preprints202307.2142.v1

communication like never before [6]. As the evolution of language models continues, fueled by robust research and increased computational power, we are witnessing a future where machines transcend being mere tools and become companions, collaborators, and confidants in our quest for knowledge and progress.

2.2. Natural Language Processing (NLP)

Natural Language Processing (NLP) represents a groundbreaking advancement in the field of computer science, enabling machines to comprehend and engage with human language with remarkable precision and fluency [7]. This revolutionary technology has transformed numerous industries, spanning customer service, healthcare, finance, and education, transcending the limits of human-machine communication [8]. NLP's capacity to decipher language intricacies, encompassing context, semantics, and sentiment, empowers businesses to extract invaluable insights from vast unstructured data, driving them towards unparalleled efficiency and innovation [9]. As NLP continues to evolve and integrate into cutting-edge applications such as virtual assistants, sentiment analysis, and language translation, it solidifies its indispensable role, shaping the future of human-computer interaction and propelling society into an era of limitless possibilities [10].

In the ever-changing realm of Artificial Intelligence, Natural Language Processing (NLP) stands out as an innovative field that has shattered the boundaries of what machines can achieve. Leveraging advanced algorithms, deep learning techniques, and sophisticated linguistic models, NLP has triumphed over the once daunting task of understanding human language, ushering in a new era of seamless collaboration between humans and machines [11]. Its applications go beyond traditional text analysis, encompassing speech recognition, language generation, and sentiment-aware content creation [12]. NLP has evolved from being a mere novelty to becoming a foundational element of modern technology, reshaping how businesses operate, how individuals interact with their devices, and how society harnesses the power of information [13]. As NLP continues to advance, unlocking new realms of communication, its potential for further innovation and societal transformation seems limitless, establishing itself as one of the most potent and awe-inspiring branches of AI.

2.3. Transformer Architecture

The emergence of the Transformer Architecture marks a groundbreaking advancement in the realm of deep learning, forever reshaping natural language processing and the AI landscape [14]. Its innovative self-attention mechanism, free from conventional recurrent or convolutional structures, shattered the constraints of sequential processing, enabling unprecedented parallelization and scalability [5]. The Transformer's capacity to model long-range dependencies and capture intricate linguistic patterns gave rise to a new generation of language models such as BERT [15] and GPT [3], achieving previously unimaginable levels of contextual comprehension and context-aware language generation. This architectural marvel has become the cornerstone of state-of-the-art applications, spanning machine translation, sentiment analysis, chatbots, and voice assistants, setting the gold standard for language-related tasks and unlocking the true potential of AI-driven language processing [12]. As researchers and engineers continue to refine and push the boundaries of the Transformer paradigm, its profound influence on artificial intelligence will only expand, cementing its status as one of the most influential and transformative developments in the history of modern deep learning.

The Transformer Architecture stands as an unstoppable catalyst, reshaping not only the landscape of natural language processing but also extending its influence into various domains within the vast expanse of deep learning [16]. Its attention-based approach, facilitating efficient and contextually sensitive information flow across extensive sequences, has transcended the boundaries of language tasks, leading to its seamless integration into computer vision, speech recognition, and even music generation [17]. This groundbreaking architecture has shattered records in model size and performance, establishing new benchmarks for AI and ushering in an era of cutting-edge models [18]. With its remarkable adaptability to diverse problem domains and data modalities, the Transformer has become an indispensable and versatile tool embraced by researchers and

 doi:10.20944/preprints202307.2142.v1

practitioners alike. As the Transformer continues to evolve and inspires the next wave of architectural innovations, its profound legacy will be eternally etched in the history of AI, commemorating its status as the driving force behind unprecedented progress and achievements in the field of deep learning.

2.4. Supervised fine-tuning

Supervised fine-tuning stands as an incredibly potent strategy that revitalizes pre-trained neural networks, unlocking their full potential by adapting them to specific tasks with unmatched precision [19]. By harnessing the extensive knowledge embedded in pre-trained models, supervised fine-tuning facilitates rapid and efficient training on task-specific datasets, significantly reducing the reliance on large amounts of labeled data [20]. This empowering approach results in models that showcase exceptional performance, even when faced with limited training samples, making it a game-changer in scenarios where data is scarce or expensive to obtain. Supervised fine-tuning has ignited a revolution across diverse domains, spanning from computer vision and natural language processing to audio analysis, empowering researchers and practitioners to leverage state-of-the-art models without starting from scratch [21]. With its ability to transfer knowledge, enhance performance, and democratize access to sophisticated AI solutions, supervised fine-tuning solidifies its position as an indispensable technique, breaking barriers and propelling the AI community towards unprecedented efficiency and innovation.

Supervised fine-tuning emerges as a formidable ally in the quest for exceptional AI models, wielding the potential to transform generic pre-trained architectures into specialized powerhouses tailored to specific tasks [22]. Its capacity to adapt and specialize neural networks with limited labeled data circumvents the resource-intensive process of training models from scratch, accelerating progress across the artificial intelligence landscape [23]. Empowering practitioners to achieve unprecedented levels of accuracy and performance, supervised fine-tuning revolutionizes various applications, from image classification and sentiment analysis to machine translation and speech recognition. By fine-tuning on specific tasks, these models acquire domain expertise, showcasing a refined understanding of intricate patterns and nuances within the target data [24]. As this technique continues to evolve, pushing the boundaries of AI capabilities, supervised fine-tuning solidifies its position as a transformative force, ushering in an era where potent machine-learning solutions are within reach for a wide range of practical challenges [25].

3. Llama 2 Models and Licensing

The LLaMA and LLaMA 2 models are instances of Generative Pretrained Transformer (GPT) models, built upon the original Transformers architecture [1]. The LLaMA models employ GPT-3-like pre-normalization, utilizing the RMS Norm normalizing function at the input of each transformer sub-layer [26]. This approach enhances training stability by rescaling the invariance property and implicit learning rate adaptation ability. Additionally, LLaMA benefits from the SwiGLU activation function, replacing the conventional ReLU non-linearity activation function, leading to improved training performance [1].

Incorporating insights from the GPT-Neo-X project, LLaMA incorporates rotary positional embeddings (RoPE) at each layer, contributing to its overall performance [1]. Notably, LLaMA 2 introduces essential architectural differences, as detailed in the corresponding paper's appendix. These differences include an increased context length, doubling the context window size from 2048 to 4096 tokens [26]. This extension enables the model to handle more extensive information, proving beneficial for tasks involving long documents, chat histories, and summarization.

Furthermore, LLaMA 2 implements a grouped-query attention (GQA) format with eight key-value projections, addressing the complexity concerns associated with the original Multi-Head attention baseline [1]. This modification proves effective in managing the increased context windows or batch sizes.

As a result of these updates, LLaMA demonstrates significantly improved performance across various tasks, surpassing or closely matching other specialized GPT models such as Falcon and MPT

 doi:10.20944/preprints202307.2142.v1

[27]. The model's promising performance paves the way for further research, anticipating future comparisons with prominent closed-source models like GPT-4 and Bard.

3.1. Accessibility and Licensing

Llama 2 is an open-source project, rendering its source code accessible to the general public, thereby enabling unfettered scrutiny, utilization, modification, and distribution. The project adheres to a permissive licensing model, affording users considerable freedoms to employ the model for diverse purposes with minimal constraints [2].

From the perspective of the research community, this permissive licensing confers notable advantages by fostering unrestrained access to the Llama 2 model [28]. Researchers can seamlessly integrate the model into their scholarly inquiries, experiments, and academic pursuits without encountering legal impediments. Such open access facilitates collaborative efforts, encourages innovation, and drives progress in the realm of natural language processing [29].

Likewise, the business community also benefits significantly from the implications of this liberal licensing approach. By virtue of this framework, companies and startups can seamlessly incorporate the Llama 2 model into their products and services, sidestepping the necessity for intricate licensing arrangements or substantial financial commitments. This unimpeded access empowers businesses to conduct extensive experimentation with the model, catalyzing the development of novel applications and innovative solutions, and harnessing Llama 2's capabilities [29].

The permissive licensing strategy employed for Llama 2 is widely regarded as a propitious development for both the research and business communities [29]. It propels widespread experimentation, fosters robust development, and facilitates broad adoption of the model, thereby potentially engendering transformative advancements in the domain of natural language processing and its associated fields.

It is imperative to highlight that the utilization of Llama Materials for enhancing any other sizable language model, with the exception of Llama 2 or its derivatives, is subject to restrictions. Moreover, if the products or services offered by the Licensee manage to amass a user base exceeding 700 million monthly active users, a distinct license request to Meta becomes mandatory [30].

3.2. Llama 2 Models and Versions

Meta has developed and publicly released the Llama 2 family of large language models (LLMs), comprising a set of pre-trained and fine-tuned generative text models with parameter sizes spanning from 7 billion to 70 billion [1]. Among these, the fine-tuned LLMs, specifically named Llama-2-Chat, have been tailored to optimize performance in dialogue-based use cases. Through comprehensive benchmark assessments, Llama-2-Chat models have demonstrated superior capabilities compared to open-source chat models. Additionally, in human evaluations focused on assessing helpfulness and safety, Llama-2-Chat models have shown comparable performance to some prominent closed-source models, including ChatGPT and PaLM [31,32].

Llama 2 encompasses various parameter sizes, such as 7B, 13B, and 70B, and offers both pre-trained and fine-tuned variants. The parameter size plays a crucial role in determining model accuracy, with larger parameter sizes signifying extensive training with vast corpora, leading to more precise and dependable responses [31]. In addition to the various size variants, it is pertinent to mention the availability of a fine-tuned version of the model tailored specifically for chat applications, denoted as Llama 2-Chat [32].

In order for a user to download the pre-trained models, they are required to request access through the official Meta website, agreeing to the specified terms and conditions. Upon approval, the user will receive an email containing a custom unique URL, which grants access to download the models. Within a new Python project, the user can utilize the provided Git URL on GitHub to clone the Llama 2 repository [31]. By executing the download.sh script, the user will be prompted to enter the custom unique URL and choose the desired models for download. The available options include 7B, 13B, 70B, 7B-chat, 13B-chat, and 70B-chat [1]. It is essential to note that the pre-trained model files

 doi:10.20944/preprints202307.2142.v1

are quite large, hence the user must have sufficient storage space, processing power, GPU and RAM to handle these models, especially if they intend to perform fine-tuning [26,31].

4. Training Process

4.1. Pretraining Data

The training corpus of Llama 2 comprises a novel blend of publicly accessible data sources, excluding any data originating from Meta's products or services. During the data selection process, diligent measures were taken to exclude data from websites known to contain substantial volumes of personal information about private individuals. The model underwent training on an extensive dataset comprising 2 trillion tokens, exhibiting twice the context length of its predecessor, Llama 1 [28]. This design choice strikes a balance between performance and computational cost, with a deliberate emphasis on up-sampling the most factual sources to enhance knowledge while mitigating potential hallucination issues [1].

The developers of Llama 2 retained much of the pretraining settings and model architecture employed in Llama 1. The model adheres to the standard transformer architecture proposed by Vaswani [33], utilizing pre-normalization with RMSNorm [34] and the SwiGLU activation function [35]. Furthermore, it integrates rotary positional embeddings (RoPE) [1,36].

Key differences between Llama 1 and Llama 2 lie in the augmentation of context length and the adoption of grouped-query attention (GQA). These architectural modifications contribute to the improved capabilities of Llama 2 and its ability to handle more extensive contextual information during language generation tasks [1].

4.2. Llama 2 Fine-tuning

Llama 2 is pre-trained using publicly available online data [27]. An initial version of Llama-2-chat is then created through the use of supervised fine-tuning. Next, Llama-2-chat is iteratively refined using Reinforcement Learning from Human Feedback (RLHF), which includes rejection sampling and proximal policy optimization (PPO) [1].

In the pursuit of optimizing the performance of Language Model Models (LLMs) for dialogue-style instructions, the key aspects revolve around the quality and diversity of third-party Source-Free Tuning (SFT) data [36]. Although numerous sources provide such data, their limited diversity and quality led the focus to prioritize the collection of high-quality SFT examples, resulting in significant improvement. Meta's study also found that a limited set of clean instruction-tuning data could yield satisfactory outcomes, and approximately tens of thousands of SFT annotations were sufficient for achieving high-quality results [1]. Notably, the annotation platform and vendor choices influenced the downstream model performance, emphasizing the significance of thorough data checks. The validation process confirmed the high quality of outputs from the SFT model, suggesting a potential to reallocate annotation efforts towards preference-based annotation for Reinforcement Learning from Human Feedback (RLHF). The investigation encompassed 27,540 annotations, excluding Meta user data, and drew parallels with related research highlighting the effectiveness of focusing on quality over quantity in instruction-tuning endeavors [1].

The Reinforcement Learning with Human Feedback (RLHF) is a model training approach used to further align the behavior of a fine-tuned language model with human preferences and instructions [37]. Human preference data is collected, where annotators select their preferred choice between two model outputs, aiding in the training of a reward model that automates preference decisions [38]. The collection procedure involves a binary comparison protocol to maximize prompt diversity, with annotators rating the degree of preference for their chosen response. Safety and helpfulness aspects are specifically focused on, allowing the application of distinct guidelines to each. The reward modeling data, referred to as "Meta reward modeling data," comprises over one million binary comparisons, surpassing existing open-source datasets in terms of conversation turns and average length [1]. Continuous updates to the reward model are essential to adapt to the evolving Llama 2-Chat iterations and maintain accurate rewards for the latest model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 August 2023 doi:10.20944/preprints202307.2142.v1

The reward model plays a crucial role in Reinforcement Learning with Human Feedback (RLHF), where it takes a model response and its corresponding prompt, and outputs a scalar score indicating the quality in terms of helpfulness and safety [38]. By leveraging these response scores as rewards, the RLHF process optimizes Llama 2-Chat to align better with human preferences and enhance helpfulness and safety [1]. To address the trade-off between helpfulness and safety, two separate reward models are trained - one for each aspect. The reward models are initialized from pretrained chat model checkpoints to ensure knowledge transfer and prevent information mismatch. Training objectives involve converting collected human preference data into binary ranking labels, with a margin component to handle different preference ratings [1,39]. Additionally, the reward models are trained on a combination of newly collected data and existing open-source datasets to improve generalization and prevent reward hacking. The results demonstrate the superiority of the proposed reward models over other baselines.

4.3. Llama 2 Eco-consciousness

In alignment with Meta's corporate eco-conscious responsibility, the developers meticulously recorded the Carbon Footprint associated with the GPU hours of computation on hardware [32]. The Carbon Footprint resulting from the pretraining phase involved a cumulative usage of 3.3 million GPU hours on A100-80GB hardware, with a power consumption range of 350-400W. The estimated total carbon emissions during this process amounted to 539 metric tons of CO2 equivalent (tCO2eq). Notably, Meta's sustainability program successfully offset 100% of these emissions [1].

Specifically, for each Llama 2 model variant, the corresponding GPU hours, power consumption, and carbon emissions were recorded as follows:

- Llama 2 7B: 184,320 GPU hours, 400W power consumption, and 31.22 tCO2eq carbon emissions.
- Llama 2 13B: 368,640 GPU hours, 400W power consumption, and 62.44 tCO2eq carbon emissions.
- Llama 2 70B: 1,720,320 GPU hours, 400W power consumption, and 291.42 tCO2eq carbon emissions.

In total, the combined emissions for all Llama 2 models amounted to 3311,616 GPU hours and 539.00 tCO2eq carbon emissions. It is worth noting that the entire carbon emissions were effectively offset by Meta's sustainability program. Additionally, since the models are being openly released, there is no need for others to incur pre-training costs [32].

5. Llama 2: Early Adopters' Case Studies and Projects

Having presented the foundational components of the Llama 2 model, this chapter proceeds to showcase the case studies and projects undertaken by the early adopters.

5.1. Official Llama2 Recipes Repository

The 'llama-recipes' repository serves as an accompanying resource to the Llama 2 model, aimed at facilitating swift and practical implementation of fine-tuning for domain adaptation and model inference for the fine-tuned versions [40]. With the primary objective of user-friendliness, the repository provides illustrative examples employing Hugging Face converted variants of the models. For seamless access to the converted models, a step-by-step guide on model conversion is provided for reference.

While Llama 2 holds tremendous potential for various applications, its novelty necessitates careful consideration of potential risks associated with its deployment. As extensive testing is inherently limited in covering all conceivable scenarios, developers are encouraged to adopt responsible AI practices in order to address these risks. To aid developers in this endeavor, a dedicated Responsible Use Guide has been formulated, offering comprehensive insights and guidelines. Additional details on this subject can be explored further in the associated research paper.

To obtain the models, developers can readily follow the instructions outlined in the Llama 2 repository, thus ensuring a seamless process of model retrieval. By providing a supportive ecosystem

 doi:10.20944/preprints202307.2142.v1

of examples, guidelines, and downloads, the 'llama-recipes' repository is poised to empower developers in harnessing the capabilities of Llama 2 while adhering to best practices for responsible AI deployment.

5.2. Llama2.c by @karpathykarpathy

The provided Llama project repository offers a comprehensive solution for training the Llama 2 LLM architecture from scratch in PyTorch [41]. The repository facilitates the export of model weights to a binary file, which can then be loaded into a concise 500-line C file (run.c) for efficient model inference. Additionally, the repository enables users to load, finetune, and infer Meta's Llama 2, though this aspect is continuously evolving. The project stands as a "fullstack" solution, encompassing both training and inference functionalities for Llama 2 LLM, prioritizing minimalism and simplicity.

Contrary to the notion that large parameter LLMs are necessary for practical utility, the repository advocates for the effectiveness of comparatively smaller LLMs in specialized domains. By narrowing the domain appropriately, even compact LLMs can demonstrate remarkable performance.

It is important to note that this project originated from the nanoGPT model, which was later fine-tuned to accommodate the Llama-2 architecture in place of GPT-2. Notably, the core aspect of the project involved crafting the C inference engine in run.c. Given its recent inception, the project is in its nascent stages and is undergoing rapid progress. Credit is acknowledged to the inspiring llama.cpp project that prompted the development of this repository. In pursuit of minimalism, the choice was made to hard-code the Llama 2 architecture, adhere to fp32, and generate a pure C inference file without any dependencies, enhancing ease of implementation and accessibility.

5.3. Llama2-Chinese by @FlagAlpha

The Llama2 Chinese Community stands as a dynamic and innovative hub, dedicated to optimizing and advancing the Llama2 model specifically for Chinese language applications [42]. Leveraging vast Chinese language data, the community undertakes continuous iterative upgrades of the Llama2 model, bolstering its prowess in handling Chinese text. Comprised of a team of skilled NLP high-level engineers, the community provides robust technical support, guiding members towards professional development and achievement. Moreover, the community fosters an atmosphere of knowledge exchange, organizing regular online events, technical discussions, and project showcases, encouraging collaborative learning and innovative breakthroughs. As a global platform, the community welcomes developers and researchers from diverse backgrounds to unite in their shared interest and passion for large language models (LLMs) and collectively explore the boundless potential of Llama2 in the realm of Chinese NLP technology. With a commitment to open sharing and responsible AI practices, the Llama2 Chinese Community stands poised to shape the future of Chinese language processing and inspire cutting-edge solutions in this thriving field.

5.4. Llama2-chatbot by @a16z-infra

The LLaMA 2 Chatbot App is an experimental Streamlit application specifically designed for LLaMA 2 (or any other large language model - LLM) [43]. The app showcases an interactive chatbot interface with session chat history, allowing users to engage in dynamic conversations. It offers the flexibility to select from multiple LLaMA 2 API endpoints on Replicate, including 7B, 13B, and 70B options, with 70B set as the default. Additionally, users can configure model hyperparameters from the sidebar, such as Temperature, Top P, and Max Sequence Length, to tailor the chatbot's responses. The app also features "User:" and "Assistant:" prompts, distinctly distinguishing the conversation participants. Furthermore, models, 7B, 13B, and 70B, are deployed on Replicate, utilizing A100 40Gb and A100 80Gb resources. A Docker image is thoughtfully provided, enabling easy deployment of the app in Fly.io. For an immersive experience, users can access the live demo at LLaMA2.ai, while retaining their session chat history throughout each interaction, although refreshing the page will reset the history. The LLaMA 2 Chatbot App offers an exciting exploration of language models'

 doi:10.20944/preprints202307.2142.v1

capabilities, empowering users to engage in interactive conversations while experimenting with different model configurations.

5.5. Llama2-webui by @liltom-eth

The llama2-webui repository offers a user-friendly solution to run Llama 2 models with a gradio web UI, facilitating GPU or CPU inference from any operating system (Linux/Windows/Mac) [44]. It extends support to all Llama 2 models, including 7B, 13B, 70B, GPTQ, and GGML, with 8-bit and 4-bit modes. Users can leverage GPU inference with a minimum of 6 GB VRAM, or opt for CPU inference, providing flexibility in choosing the hardware configuration. Supported model backends include Nvidia GPU with transformers, bitsandbytes (for 8-bit inference), and AutoGPTQ (for 4-bit inference), making use of GPUs with at least 6 GB VRAM. Additionally, CPU and Mac/AMD GPU inference are enabled through llama.cpp, and a demonstration of CPU inference on Macbook Air is showcased. The web UI interface utilizes gradio, ensuring an intuitive and interactive experience for users, regardless of their proficiency level. This repository empowers users to effortlessly experiment with various Llama 2 models and configurations, all via a seamless web-based interface, promoting accessibility and ease of use across diverse environments.

5.6. Llama-2-Open-Source-LLM-CPU-Inference by @kennethleungty

The repository presents a comprehensive and well-explained guide for running quantized open-source Large Language Models (LLMs) on CPUs [45]. The guide covers the usage of LLama 2, C Transformers, GGML, and LangChain, offering step-by-step instructions to facilitate smooth deployment. With a focus on document question-and-answer (Q&A) scenarios, the guide provides practical insights for implementing self-managed or private model deployment. By hosting open-source LLMs locally on-premise or in the cloud, teams can address data privacy and residency concerns, reducing reliance on third-party commercial LLM providers like OpenAI's GPT4. While GPU instances are commonly preferred for compute capacity, this project showcases how quantized versions of open-source LLMs can be efficiently run on local CPUs, mitigating excessive costs and enabling budget-friendly deployments. Through this valuable resource, users can explore the potential of open-source LLMs and empower themselves to tailor LLM applications to their specific needs, effectively expanding the range of options available for model deployment. The step-by-step guide is accessible on TowardsDataScience, providing a well-rounded understanding of running quantized open-source LLM applications on CPUs for document Q&A tasks.

5.7. Docker-llama2-chat by @soulteary

The Docker LLaMA2 Chat repository offers an efficient and user-friendly approach to experience the capabilities of LLaMA2 models in various configurations [46]. The repository provides a step-by-step guide for fast and straightforward local deployment of official LLaMA2 models, such as 7B or 13B, as well as the Chinese version of LLaMA2. The deployment process is made easy through Docker, allowing users to run quantized versions of the LLaMA2 models on CPUs with minimal resource requirements. The repository includes comprehensive blog tutorials, enabling users to explore different types of LLaMA2 models, each tailored for specific use cases. With detailed instructions and commands, users can quickly set up the LLaMA2 models and initiate them through a web interface, gaining immediate access to interactive chat applications. The repository showcases a variety of LLaMA2 models, including INT4 quantization and CPU inference support, widening the range of available options for model deployment. Moreover, users can leverage the provided Docker image to run GGML models and experiment with diverse LLaMA2 applications. The repository's user-friendly approach and extensive documentation make it an excellent resource for individuals seeking to delve into LLaMA2's capabilities and integrate it into their projects effortlessly.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 August 2023 doi:10.20944/preprints202307.2142.v1

5.8. Llama2 by @dataprofessor

The Llama 2 Chat repository presents a user-friendly chatbot application built using the open-source Llama 2 Large Language Model (LLM) from Meta [47]. Specifically, the app utilizes the Llama2-7B model, which is deployed by the Andreessen Horowitz (a16z) team and hosted on the Replicate platform. The application has been refactored to be lightweight, ensuring easy deployment to the Streamlit Community Cloud. To run the app, users need to obtain their own Replicate API token after signing up on the Replicate platform. With the provided API token, users can interact with the chatbot and explore its capabilities. The chatbot's implementation allows users to try other Llama 2 models, such as Llama2-13B and Llama2-70B, with varying parameter sizes. This chatbot app serves as an accessible and engaging platform for users to experience the power and versatility of Llama 2, making it an ideal choice for those interested in exploring large language models and their applications.

5.9. Llama-2-jax by @ayaka14732

The JAX Implementation of Llama 2 is an ambitious and advanced project focused on implementing the Llama 2 model using JAX, enabling efficient training and inference on Google Cloud TPU [48]. The repository aims to develop a high-quality codebase exemplifying the Transformer model's implementation using JAX while providing valuable insights for the NLP community by facilitating the identification of common errors and inconsistencies across various transformer models. This implementation supports various features, including parameter conversion to and from Hugging Face, model architecture components, cross-entropy loss, logits processing, generation methods like beam search and sampling, optimization, data loading, inference, and training, among others. The environment setup involves specific versions of Python, JAX, PyTorch, and Transformers. Users can create a virtual environment, install the necessary dependencies, and download LLaMA weights for the implementation. Special configurations are provided for TPU pods and multi-host environments. This repository serves as a valuable resource for anyone interested in JAX-based implementations of transformer models, particularly the Llama 2 model, with a focus on performance and efficiency using Google Cloud TPUs.

5.10. LLaMA2-Accessory by @Alpha-VLLM

LLaMA2-Accessory is an exceptional open-source toolkit designed for the pre-training, fine-tuning, and deployment of Large Language Models (LLMs) and multimodal LLMs [49]. This repository builds upon the foundation of LLaMA-Adapter, introducing more advanced features to enhance LLM development. The toolkit offers support for a wide range of datasets and tasks, including pre-training with RefinedWeb and StarCoder, single-modal fine-tuning with Alpaca, ShareGPT, LIMA, UltraChat, and MOSS, and multi-modal fine-tuning with various image-text pairs, interleaved image-text data, and visual instruction data. Efficient optimization and deployment techniques are also included, such as parameter-efficient fine-tuning, fully sharded data parallelism, and advanced visual encoders and LLMs like CLIP, Q-Former, ImageBind, LLaMA, and LLaMA2. The core contributors have put tremendous effort into making this toolkit a powerful resource for LLM development.

5.11. Llama2-Medical-Chatbot by @AIAnytime

Llama2-Medical-Chatbot is an innovative medical chatbot that harnesses the capabilities of Llama2 and Sentence Transformers [50]. This powerful combination allows the bot to deliver intelligent and context-aware responses to medical queries. The chatbot is further enhanced by the integration of Langchain and Chainlit, providing advanced language processing and understanding capabilities. Operating on a robust CPU machine with at least 16GB of RAM, the chatbot ensures smooth and efficient performance, making it an invaluable tool for medical professionals and individuals seeking reliable medical information and assistance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 August 2023 doi:10.20944/preprints202307.2142.v1

5.12. Llama2-haystack by @anakin87

The "llama2-haystack" repository contains experimental work where Llama2 is used with Haystack, an NLP/LLM framework [51]. The notebook showcases various hacky experiments aiming to load and utilize Llama2 within the Haystack environment. While it's not an official or fully refined implementation, it may serve as a useful resource for others who are also experimenting with Llama2 and Haystack. The notebook highlights the installation of Transformers and other necessary libraries, the loading of Llama-2-13b-chat-hf with 4-bit quantization, and the process of handling Tensor Parallelism issues. Additionally, a minimal version of Haystack is installed, and a creative approach is taken to load the model into Haystack's PromptNode.

6. Results

In the previous chapters, we thoroughly examined the architecture of Llama 2, its models, the methods employed for model training, the technologies utilized, and the fine-tuning process. Furthermore, in Section 5, we presented the most significant Case Studies and Projects undertaken by Early Adopters. Despite being at an early stage of research, Llama 2 has garnered significant interest from early adopters, who have promptly engaged in fine-tuning the model for various domains, including the medical domain. A similar trend was observed with the public release of GPT-3 in early 2023, where early adopters swiftly embraced the technology for diverse medical applications. As we are still in the nascent stages of development, the only certainty is that pre-trained models like Llama 2 will gradually replace older models in various domains.

The following **Error! Reference source not found.** succinctly summarizes the case studies and their corresponding areas of focus. Through this table, we can extract valuable insights into the emerging trends and tendencies for domains that will be early adopters of this new technology. Analyzing the patterns and preferences of these adopters will provide valuable information to anticipate the adoption trends across various domains.

Table 1. Llama 2: Early Adopters' Projects and their Areas of Focus.

Llama 2: Early Adopters' Projects	Areas of Focus
Recipes Repository [40]	fine-tuning
Llama2.c by [41]	model deployment / PyTorch
Llama2-Chinese [42]	fine-tuning / language / Chinese
Llama2-chatbot [43]	chatbot
Llama2-webui [44]	model deployment / model optimization / CPU / GPU
Llama-2-Open-Source-LLM-CPU-Inference [45]	model deployment / model optimization / CPU
Docker-llama2-chat [46]	chatbot / model deployment / docker / CPU
Llama2 [47]	model deployment / chatbot
Llama-2-jax [48]	model deployment / JAX / PyTorch / Google Cloud TPUs
LLaMA2-Accessory [49]	model deployment / fine-tuning
Llama2-Medical-Chatbot [50]	chatbot / medical / fine-tuning / model deployment / CPU
Llama2-haystack [51]	model deployment / Haystack

Despite being publicly announced only 10 days ago, Llama 2 has garnered significant interest from early adopters. Within this brief timeframe (August 18, 2023, to August 28, 2023), these adopters have successfully accomplished various tasks, including model deployment, chatbot development,

 doi:10.20944/preprints202307.2142.v1

fine-tuning in multiple languages, domain-specific chatbot creation (medical domain), parameter customization for CPU and GPU, and optimization to enhance runtime efficiency with minimal resources. It is essential to acknowledge that the sample size used for these activities is relatively small, given the short duration since the model's launch. Consequently, drawing definitive conclusions may be premature; nonetheless, certain assumptions presented in the Introduction chapter could be reasonably supported based on the observed early adoption trends.

Hypothesis 1

Research Objective: The research aims to assess the audience's response to Llama 2 and verify Meta's expectations that an open-source model will experience faster development compared to closed models [2].

- Null Hypothesis (H0): There is no significant difference in the audience's response to Llama 2 between the open-source model and closed models.
- Alternative Hypothesis (H1): There is a significant difference in the audience's response to Llama 2, with the open-source model experiencing faster development compared to closed models, as expected by Meta.

Certainly, the measurement of timing is a challenging aspect to quantify; nevertheless, the immense enthusiasm displayed by the media and the prompt response from early adopters may serve as indications that the open-source nature of Llama 2 has attracted such a response. Therefore, with some reservation, the null hypothesis could be rejected.

Hypothesis 2

Research Objective: The research aims to assess the challenges encountered by early adopters in deploying the Llama 2 model.

- Null Hypothesis (H0): There is no significant difference in the challenges encountered by early adopters in deploying the Llama 2 model.
- Alternative Hypothesis (H1): There is a significant difference in the challenges encountered by early adopters in deploying the Llama 2 model.

Based on the data presented in **Error! Reference source not found.**, it can be observed that early adopters did not encounter any challenges in deploying the Llama 2 model. This finding emphasizes the significant level of preparation that a product must undergo before its public launch. Notably, the Meta company had evidently prepared the ground thoroughly prior to making the model available to the public. Consequently, the null hypothesis can be rejected with confidence.

Hypothesis 3

Research Objective: The research aims to assess the challenges encountered by early adopters in fine-tuning the Llama 2 model.

- Null Hypothesis (H0): There is no significant difference in the challenges encountered by early adopters in fine-tuning the Llama 2 model.
- Alternative Hypothesis (H1): Early adopters encounter significant challenges in fine-tuning the Llama 2 model.

The findings of this study indicated that early adopters promptly engaged in fine-tuning the Llama 2 model for both specific and non-specific domain projects. This prompt adoption could be attributed to the clear instructions provided by Meta simultaneously with the launch of Llama 2 [18]. Therefore, the null hypothesis cannot be rejected, suggesting that there is no significant difference in the timing of fine-tuning.

Hypothesis 4

Research Objective: The research seeks to unveil that the medical domain consistently ranks among the primary domains that early adopters engage with, undertaking fine-tuning of models.

 doi:10.20944/preprints202307.2142.v1

- Null Hypothesis (H0): There is no significant difference in the interest shown by early adopters of Llama 2 between the medical domain and other domains.
- Alternative Hypothesis (H1): Early adopters of Llama 2 prioritize the medical domain significantly more than other domains, indicating a greater interest in utilizing LLMs for medical applications.

Just like the research conducted for the GPT-3 model [3] after its launch in early 2023, early adopters of Llama 2 also show a keen interest in the medical domain. This emphasis on the medical field highlights the significant role of LLMs in medicine and its impact on human life. Consequently, the null hypothesis can be confidently rejected.

7. Responsible AI and Ethical Considerations

The adoption of AI models like Llama 2 in real-world applications necessitates a robust commitment to responsible AI practices and ethical considerations. This section explores the responsible use of Llama 2 by early adopters and sheds light on the ethical challenges that arise with its widespread utilization.

As Meta mentioned in its paper, the open release of language models (LLMs), including Llama 2, holds the potential to bring substantial benefits to society [1]. However, it is essential to acknowledge that LLMs are novel technologies that inherently carry certain risks associated with their usage [53,54,55]. As of now, testing of Llama 2 has primarily been conducted in English, and it is practically impossible to cover all possible scenarios that might arise during real-world deployments.

Given the potential risks, Meta emphasizes the importance of responsible and safe usage of Llama 2 and its variant Llama 2-Chat. Before integrating these models into any applications, developers are strongly advised to conduct thorough safety testing and tuning that aligns with the specific use-cases and contexts of their projects. Such an approach will help mitigate potential ethical and societal implications that might arise from unchecked deployments [1].

To assist developers in responsibly deploying Llama 2 and Llama 2-Chat, Meta provides a dedicated responsible use guide [29]. This guide outlines best practices, guidelines, and considerations for ensuring the safe and ethical integration of the models. Additionally, code examples are made available to facilitate the implementation of these practices effectively.

In Section 5.3 of Meta's paper, further details of their responsible release strategy are presented. This section delves into the methodologies and frameworks employed to address ethical concerns, data privacy, fairness, transparency, and other critical aspects that underpin the responsible deployment of Llama 2 and Llama 2-Chat [1].

Meta's proactive approach to responsible AI release underscores its commitment to promoting ethical AI practices within the AI community. By providing developers with the necessary tools, guidelines, and insights, Meta seeks to foster an ecosystem where the benefits of LLMs like Llama 2 can be harnessed responsibly while minimizing potential risks and ensuring positive societal impact [28,29]. Collaboration between the AI research community and developers will be crucial in continuously refining and optimizing the responsible use strategies for Llama 2, driving the AI field toward a more ethical and sustainable future.

In the realm of Artificial Intelligence (AI), with particular emphasis on Language Model Models (LLMs) exemplified by Llama 2, a multitude of ethical considerations and salient facets come to the fore, which bear profound significance in fostering the development and implementation of responsible AI practices. Some of the most pivotal aspects are outlined herewith:

1. Bias Mitigation and Fairness: Early adopters' experiences with Llama 2 have highlighted the importance of addressing biases in AI outputs. As a pre-trained model trained on diverse data sources, Llama 2 may inadvertently inherit biases present in the training data. Researchers and developers must implement robust techniques to identify and mitigate biases to ensure fairness and equitable outcomes across diverse user populations [56,57,58].
2. Transparency and Interpretability: The complexity of deep learning models like Llama 2 can present challenges in understanding their decision-making processes. To promote transparency

 doi:10.20944/preprints202307.2142.v1

and interpretability, early adopters have emphasized the need for methods that provide insights into the model's internal workings. Future research should focus on developing techniques to make AI models more interpretable, enabling users to comprehend the rationale behind model's predictions [59].

3. Privacy and Data Protection: Llama 2's success heavily relies on the vast amount of data used during pretraining. Early adopters recognize the significance of safeguarding user data and respecting privacy concerns. Employing privacy-preserving methods, such as federated learning or differential privacy, can uphold the confidentiality of user data while ensuring the model's effectiveness [60,61].
4. Ethical Use-Cases and Societal Impact: As AI technologies like Llama 2 become increasingly integrated into various domains, early adopters have stressed the importance of identifying and promoting ethically sound use cases. Research should extend to analyze the societal impact of Llama 2's deployment, considering potential consequences on individuals, communities, and societal values. Striking a balance between innovation and responsible AI practices is crucial to harness the full potential of LLMs while mitigating unintended negative effects [62,63].
5. Continuous Monitoring and Auditing: To maintain ethical AI practices, early adopters advocate for continuous monitoring and auditing of Llama 2's performance. Regular assessments can help identify potential biases or deviations in the model's behavior, enabling timely adjustments to ensure compliance with ethical standards [64,65].
6. End-User Empowerment and Informed Consent: As AI models like Llama 2 become integral to user experiences, early adopters have emphasized the significance of end-user empowerment and informed consent. Users should be well-informed about the AI's involvement in their interactions and have the right to control and modify the extent of AI-driven recommendations or decisions [66].

The utilization of Llama 2 by early adopters brings forth important considerations regarding responsible AI and ethics. By addressing bias, promoting transparency, ensuring privacy, and evaluating societal implications, the AI community can uphold ethical standards and foster trust in the deployment of Llama 2 and other AI models. Moving forward, continuous research and collaboration are vital in fostering responsible AI practices that contribute positively to society while unlocking the full potential of Meta's innovative open-source pre-trained model.

8. Future Directions and Research

The frontier of Language Model Models (LLMs) is brimming with possibilities, with notable contributions from the open-source pretrained LLM "Llama 2." The ongoing evolution of LLMs fuels the exploration of crucial research avenues for researchers and practitioners alike. A paramount direction involves the pursuit of larger and more sophisticated models, exemplified by Llama 2's remarkable up to 65 billion parameters, with the aim of elevating context comprehension, refining reasoning capabilities, and producing precise, contextually relevant responses [32]. These advancements hold the potential to revolutionize natural language processing, opening doors to novel applications and enhanced performance across various domains.

An imperative focus within LLM research revolves around tackling biases in the models. The possibility of biases in the training data leading to biased outputs raises concerns about fairness and equity in AI applications. Hence, researchers are diligently delving into techniques to alleviate biases within the Llama 2 model, striving to create systems that are more resilient and impartial in their language generation [1]. Moreover, considerable efforts are directed towards enhancing the interpretability of LLMs, allowing users to comprehend the decision-making processes and offering explanations for the models' outputs. This facet plays a pivotal role in cultivating trust in AI systems, promoting accountability, and reinforcing transparency in their functioning.

Moreover, there is a keen interest in exploring the multimodal capabilities of LLMs, including the dynamic Llama 2 model [1]. By seamlessly integrating visual and textual information, new frontiers emerge, facilitating a more comprehensive and contextually-aware approach to content understanding and generation. Multimodal LLMs, exemplified by Llama 2, possess the potential to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 August 2023 doi:10.20944/preprints202307.2142.v1

revolutionize diverse applications, ranging from image captioning and video analysis to virtual assistants. As Llama 2 continues its advancement, the fusion of language comprehension with visual data is anticipated to yield increasingly sophisticated and human-like AI systems. The open-source nature of Llama 2 paves the way for researchers and developers to actively contribute to and derive benefits from these exciting advancements in the realm of language modeling.

To date, the assessment of Llama 2 [1] performance has been primarily reliant on benchmarks provided exclusively by its developers, which encompass datasets such as TriviaQA, SQuAD, and GSM8K, among others. These benchmarks are widely acknowledged for their reliability. Nonetheless, there is a growing consensus regarding the necessity of conducting independent evaluations employing supplementary benchmark tools and datasets. Notably, the LAMBADA dataset presents a language modeling and word prediction benchmark that aligns well with the proficiencies of LLMs [67]. This dataset allows for evaluating LLMs' predictive capabilities when tasked with filling in missing words within sentences. Similarly, the RACE (RACE: Large-scale ReAding Comprehension Dataset from Examinations) dataset serves as an evaluation resource to assess LLMs' competence in comprehending and responding to questions based on given passages [67]. Furthermore, the SuperGLUE benchmark, as an extension of GLUE, encompasses more challenging tasks that are particularly apt for evaluating the capabilities of advanced LLMs [68]. It serves as a means to assess how effectively LLMs navigate complex language understanding tasks, including ReCoRD and CommonsenseQA. Consequently, incorporating these additional benchmark tools contributes to a more comprehensive evaluation of LLMs' performance, particularly with respect to their language modeling, comprehension, and reasoning abilities.

Finally, although the license agreement may impose certain obstacles for those seeking substantial profits from Llama, initiatives like LLaMA2.ai [43] are deemed necessary. The success of ChatGPT lies in the user's ease of running such models without the need to install software or deploy models themselves [3]. A compelling proposition entails Meta's potential development of a comparable tool to ChatGPT, encompassing both an API and ChatBot UI. Additionally, this tool could be endowed with the capability of being fine-tuned by user-generated inputs on current topic tasks. Consequently, the acquired knowledge from these tasks could be incorporated into the pretrained model as updates, thereby augmenting its adaptability and usefulness.

The development of such models necessitates collaboration between the parent company and early adopters. Early adopters, through domain-specific projects and fine-tuning, play a crucial role in unlocking market demands. Subsequently, the parent company and the open-source community take charge of meeting these demands and providing the necessary offerings.

9. Discussion

The investigation into early adopters' utilization of Meta's new open-source pre-trained model, Llama 2, has provided valuable insights into the practical implications and challenges associated with its application in AI projects. This section discusses the key findings and implications of this study, as well as identifies potential avenues for future research in the realm of AI model adoption and development.

- The investigation into application diversity and effectiveness of Llama 2 indicates a notable level of interest among early adopters across a broad spectrum of AI projects. These adopters have demonstrated successful deployment of the model on multiple platforms and technologies, particularly when fine-tuned for domain-specific tasks. This observation underscores the model's versatility and effectiveness in addressing various AI tasks, making it a potential solution of interest for researchers and developers seeking a unified model suitable for multiple applications.
- Early Adopters' Feedback and Challenges: Despite the limited availability of feedback, early adopters reported encountering minimal challenges in both deployment and fine-tuning processes of Llama 2. This outcome reflects favorably on Meta for synchronously launching the model with model recipes [40], which seemingly contributed to a smooth user experience and implementation for the adopters.

 doi:10.20944/preprints202307.2142.v1

- Cross-Model Comparisons: In order to obtain a comprehensive assessment of Llama 2's standing within the AI landscape, it is suggested that future studies conduct comparative analyses with other prominent pretrained models. By undertaking cross-model comparisons, researchers can glean valuable insights into Llama 2's distinctive contributions, advantages, and areas in which it outperforms existing alternatives. Such analyses would aid in elucidating the specific strengths and capabilities of Llama 2, contributing to a more holistic understanding of its potential in the field of artificial intelligence.
- Extended Use Cases and Domains: While the present study provides insights into early adopters' deployment of Llama 2 in specific applications, future research endeavors could extend the investigation to encompass its implementation across additional domains and diverse use cases. Exploring Llama 2's potential in emerging fields, such as healthcare, finance, and environmental sciences, would not only exemplify its versatility but also widen its potential impact across various industries and research domains.

Finally, the examination of early adopters' utilization of Meta's Llama 2 pretrained model has provided valuable insights into its multifaceted capabilities, effectiveness, and encountered challenges. The knowledge gained from this study lays a solid foundation for future efforts aimed at refining the model, addressing ethical implications, enhancing scalability, and exploring novel applications. By leveraging the collective feedback and experiences of early adopters, the AI community can engage in a continuous process of evolution, propelling advancements and breakthroughs in the realm of artificial intelligence.

10. Conclusions

In conclusion, Llama 2 represents a significant milestone in the field of natural language processing. Its open-source nature, combined with its capabilities for research and commercial applications, holds the potential to empower diverse users to explore and implement innovative solutions responsibly. This paper has not only explored the foundational elements of the Llama 2 model but also investigated how these early adopters harness its capabilities in their AI projects. Ten days after the launch of Llama 2, the fine-tuning of the model for medical-specific domains and chatbots demonstrates a prevailing trend among researchers towards a pursuit of a more robust and contextually appropriate AI framework, aimed at fostering a higher level of quality and ethical standards for future AI applications. Moreover, we have discussed the implications of Llama 2's adoption on the broader open-source AI landscape, addressing challenges and opportunities for developers and researchers seeking to develop cutting-edge AI solutions. This study marks an early exploration of the Llama 2 pre-trained model, laying a promising foundation for forthcoming research investigations. By embracing ethical considerations, we can harness the power of Llama 2 to drive positive impacts across various domains.

Author Contributions: Conceptualization, K.I.R. and N.D.T.; methodology, K.I.R., N.D.T. and D.K.N.; validation, K.I.R. and N.D.T.; formal analysis, K.I.R., N.D.T. and D.K.N.; investigation, K.I.R., N.D.T. and D.K.N.; resources, K.I.R.; data curation, K.I.R.; writing—original draft preparation, K.I.R. and N.D.T.; writing—review and editing, K.I.R., N.D.T. and D.K.N.; visualization, K.I.R.; supervision, D.K.N.; and N.D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas,

M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; Scialom, T. Llama 2: Open Foundation and fine-tuned chat models. arXiv 2023, arXiv:2307.09288. [Google Scholar]

2. Meta and Microsoft introduce the next generation of Llama. Available online: https://ai.meta.com/blog/llama-2/ (accessed on Jul 28, 2023).
3. Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. Future Internet 2023, 15, 192. [CrossRef]
4. Dillion, D.; Tandon, N.; Gu, Y.; Gray, K. Can ai language models replace human participants? Trends in Cognitive Sciences 2023, 27, 597–600. [CrossRef]
5. Rahali, A.; Akhloufi, M.A. End-to-End Transformer-Based Models in Textual-Based NLP. AI 2023, 4, 54-110. [CrossRef]
6. Piris, Y.; Gay, A.-C. Customer satisfaction and natural language processing. Journal of Business Research 2021, 124, 264–271. [CrossRef]
7. Dash, G.; Sharma, C.; Sharma, S. Sustainable Marketing and the Role of Social Media: An Experimental Study Using Natural Language Processing (NLP). Sustainability 2023, 15, 5443. [CrossRef]
8. Arowosegbe, A.; Oyelade, T. Application of Natural Language Processing (NLP) in Detecting and Preventing Suicide Ideation: A Systematic Review. Int. J. Environ. Res. Public Health 2023, 20, 1514. [CrossRef]
9. Tyagi, N.; Bhushan, B. Demystifying the role of natural language processing (NLP) in Smart City Applications: Background, motivation, recent advances, and future research directions. Wireless Personal Communications 2023, 130, 857–908. [CrossRef]
10. Tyagi, N.; Bhushan, B. Demystifying the role of natural language processing (NLP) in Smart City Applications: Background, motivation, recent advances, and future research directions. Wireless Personal Communications 2023, 130, 857–908. [CrossRef]
11. Pruneski, J. A.; Pareek, A.; Nwachukwu, B. U.; Martin, R. K.; Kelly, B. T.; Karlsson, J.; Pearle, A. D.; Kiapour, A. M.; Williams, R. J. Natural language processing: Using artificial intelligence to understand human language in Orthopedics. Knee Surgery, Sports Traumatology, Arthroscopy 2022, 31, 1203–1211. [CrossRef]
12. Mukhamadiyev, A.; Mukhiddinov, M.; Khujayarov, I.; Ochilov, M.; Cho, J. Development of Language Models for Continuous Uzbek Speech Recognition System. Sensors 2023, 23, 1145. [CrossRef]
13. Ahmed, A.; Leroy, G.; Lu, H. Y.; Kauchak, D.; Stone, J.; Harber, P.; Rains, S. A.; Mishra, P.; Chitroda, B. Audio Delivery of Health Information: An NLP study of information difficulty and bias in listeners. Procedia Computer Science 2023, 219, 1509–1517. [CrossRef]
14. Wang, J.; Xu, G.; Yan, F.; Wang, J.; Wang, Z. Defect transformer: An efficient hybrid transformer architecture for surface defect detection. Measurement 2023, 211, 112614. [CrossRef]
15. Drosouli, I.; Voulodimos, A.; Mastorocostas, P.; Miaoulis, G.; Ghazanfarpour, D. TMD-BERT: A Transformer-Based Model for Transportation Mode Detection. Electronics 2023, 12, 581. [CrossRef]
16. Philippi, D.; Rothaus, K.; Castelli, M. A Vision Transformer architecture for the automated segmentation of retinal lesions in spectral domain optical coherence tomography images. Scientific Reports 2023, 13. [CrossRef]
17. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.-S.; Khan, F.S. Transformers in Remote Sensing: A Survey. Remote Sens. 2023, 15, 1860. [CrossRef]
18. Panopoulos, I.; Nikolaidis, S.; Venieris, S. I.; Venieris, I. S. Exploring the performance and efficiency of Transformer models for NLP on mobile devices. arXiv 2023, arXiv:2306.11426. [Google Scholar]
19. Ohri, K.; Kumar, M. Supervised fine-tuned approach for automated detection of diabetic retinopathy. Multimedia Tools and Applications 2023. [CrossRef]
20. Li, H.; Zhu, C.; Zhang, Y.; Sun, Y.; Shui, Z.; Kuang, W.; Zheng, S.; Yang, L. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. arXiv 2023, arXiv:2303.08446. [Google Scholar]
21. Lodagala, V. S.; Ghosh, S.; Umesh, S. Pada: Pruning assisted domain adaptation for self-supervised speech representations. 2022 IEEE Spoken Language Technology Workshop (SLT) 2023. 10.1109/SLT54892.2023.10022820

22. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; Han, W.; Huang, M.; Jin, Q.; Lan, Y.; Liu, Y.; Liu, Z.; Lu, Z.; Qiu, X.; Song, R.; Tang, J.; Wen, J.-R.; Yuan, J.; Zhao, W. X.; Zhu, J. Pre-trained models: Past, present and future. AI Open 2021, 2, 225–250. [CrossRef]
23. Prottasha, N.J.; Sami, A.A.; Kowsher, M.; Murad, S.A.; Bairagi, A.K.; Masud, M.; Baz, M. Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning. Sensors 2022, 22, 4157. [CrossRef]
24. Xu, Z.; Huang, S.; Zhang, Y.; Tao, D. Webly-supervised fine-grained visual categorization via deep domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2018, 40, 1100–1113. [CrossRef]
25. Tang, C. I.; Qendro, L.; Spathis, D.; Kawsar, F.; Mascolo, C.; Mathur, A. Practical self-supervised continual learning with continual fine-tuning. arXiv 2023, arXiv:2303.17235. [Google Scholar]
26. Skelton, J. Llama 2. A model overview and demo tutorial with Paperspace Gradient. Available online: https://blog.paperspace.com/llama-2/ (accessed on Jul 28, 2023)
27. Hugging Face llama-2-7b. Available online: https://huggingface.co/meta-llama/Llama-2-7b (accessed on Jul 28, 2023).
28. Llama 2 - Resource Overview - META AI. Available online: https://ai.meta.com/resources/models-and-libraries/llama/ (accessed on Jul 28, 2023).
29. Llama 2 - Responsible Use Guide. Available online: https://ai.meta.com/llama/responsible-use-guide/ (accessed on Jul 28, 2023).
30. Llama 2 License Agreement. Available online: https://github.com/facebookresearch/llama/blob/main/LICENSE (accessed on Jul 28, 2023).
31. Inference code for Llama Models - GitHub. Available online: https://github.com/facebookresearch/llama/tree/main (accessed on Jul 28, 2023).
32. Hugging Face Llama 2 Models. Available online: https://huggingface.co/models?other=llama-2 (accessed on Jul 28, 2023).
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. arXiv 2023, arXiv:1706.03762. [Google Scholar]
34. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with Subword units. arXiv 2016, arXiv:1508.07909. [Google Scholar]
35. Shazeer, N. Glu variants improve transformer. arXiv 2020, arXiv:2002.05202. [Google Scholar]
36. Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; Wang, H. Preference ranking optimization for human alignment. arXiv 2023, arXiv:2306.17492. [Google Scholar]
37. Taecharungroj, V. "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data Cogn. Comput. 2023, 7, 35. [CrossRef]
38. Sotnikov, V.; Chaikova, A. Language Models for Multimessenger Astronomy. Galaxies 2023, 11, 63. [CrossRef]
39. Maroto-Gómez, M.; Castro-González, Á.; Castillo, J. C.; Malfaz, M.; Salichs, M. Á. An adaptive decision-making system supported on user preference predictions for Human–Robot Interactive Communication. User Modeling and User-Adapted Interaction 2022, 33, 359–403. [CrossRef]
40. Facebookresearch/llama-recipes: Examples and recipes for Llama 2 model. Available online: https://github.com/facebookresearch/llama-recipes (accessed on Jul 28, 2023).
41. Karpathy/LLAMA2.C: Inference llama 2 in one file of pure C. Available online: https://github.com/karpathy/llama2.c (accessed on Jul 28, 2023).
42. Flagalpha/LLAMA2-Chinese. Available online: https://github.com/FlagAlpha/Llama2-Chinese (accessed on Jul 28, 2023).
43. A16Z-infra/LLAMA2-chatbot. Available online: https://github.com/a16z-infra/llama2-chatbot (accessed on Jul 28, 2023).
44. Liltom-Eth Liltom-eth/LLAMA2-webui. Available online: https://github.com/liltom-eth/llama2-webui (accessed on Jul 28, 2023).
45. Kennethleungty/llama-2-open-source-llm-cpu-inference. Available online: https://github.com/kennethleungty/Llama-2-Open-Source-LLM-CPU-Inference (accessed on Jul 28, 2023).
46. Soulteary Soulteary/docker-LLAMA2-chat. Available online: https://github.com/soulteary/docker-llama2-chat (accessed on Jul 28, 2023).
47. Dataprofessor/Llama2. Available online: https://github.com/dataprofessor/llama2 (accessed on Jul 28, 2023).

48. AYAKA14732/llama-2-jax. Available online: https://github.com/ayaka14732/llama-2-jax (accessed on Jul 28, 2023).
49. Alpha-VLLM/LLAMA2-accessory. Available online: https://github.com/Alpha-VLLM/LLaMA2-Accessory (accessed on Jul 28, 2023).
50. AIANYTIME/LLAMA2-Medical-chatbot. Available online: https://github.com/AIAnytime/Llama2-Medical-Chatbot (accessed on Jul 28, 2023).
51. Anakin87/LLAMA2-Haystack. Available online: https://github.com/anakin87/llama2-haystack (accessed on Jul 28, 2023).
52. Anakin87/LLAMA2-Haystack. https://github.com/anakin87/llama2-haystack (accessed on Jul 28, 2023).
53. Bender, E. M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 2021.
54. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S.; Hawkins, W.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W.; Legassick, S.; Irving, G.; Gabriel, I. Ethical and social risks of harm from language models. arXiv 2021, arXiv:2112.04359. [Google Scholar]
55. Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Daumé III, H.; Dodge, J.; Evans, E.; Hooker, S.; Jernite, Y.; Luccioni, A. S.; Lusoli, A.; Mitchell, M.; Newman, J.; Png, M.-T.; Strait, A.; Vassilev, A. Evaluating the social impact of Generative AI systems in Systems and Society. arXiv 2023, arXiv:2306.05949. [Google Scholar]
56. Li, Y.; Zhang, Y. Fairness of chatgpt. arXiv 2023, arXiv:2305.18569. [Google Scholar]
57. Abramski, K.; Citraro, S.; Lombardi, L.; Rossetti, G.; Stella, M. Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students. Big Data Cogn. Comput. 2023, 7, 124. [CrossRef]
58. Rozado, D. The Political Biases of ChatGPT. Soc. Sci. 2023, 12, 148. [CrossRef]
59. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 2019, 8, 832. [CrossRef]
60. Mazurek, G.; Małagocka, K. Perception of privacy and data protection in the context of the development of Artificial Intelligence. Journal of Management Analytics 2019, 6, 344–364. [CrossRef]
61. Goldsteen, A.; Saadi, O.; Shmelkin, R.; Shachor, S.; Razinkov, N. Ai Privacy Toolkit. SoftwareX 2023, 22, 101352. [CrossRef]
62. Hagerty, A.; Rubinov, I. Global AI Ethics: A review of the social impacts and ethical implications of Artificial Intelligence. arXiv 2019, arXiv:1907.07892. [Google Scholar]
63. Khakurel, J.; Penzenstadler, B.; Porras, J.; Knutas, A.; Zhang, W. The Rise of Artificial Intelligence under the Lens of Sustainability. Technologies 2018, 6, 100. [CrossRef]
64. Minkkinen, M.; Laine, J.; Mäntymäki, M. Continuous auditing of Artificial Intelligence: A conceptualization and assessment of tools and Frameworks. Digital Society 2022, 1. [CrossRef]
65. Mökander, J.; Floridi, L. Ethics-based auditing to develop trustworthy AI. Minds and Machines 2021, 31, 323–327. [CrossRef]
66. Usmani, U. A.; Happonen, A.; Watada, J. Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) 2023. [CrossRef]
67. Zeng, C.; Li, S.; Li, Q.; Hu, J.; Hu, J. A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets. Appl. Sci. 2020, 10, 7640. [CrossRef]
68. Eleftheriadis, P.; Perikos, I.; Hatzilygeroudis, I. Evaluating Deep Learning Techniques for Natural Language Inference. Appl. Sci. 2023, 13, 2577. [CrossRef]