MonAI: A Decentralized AI Ecosystem

Introduction

Incumbents of the AI industry such as OpenAI, Anthropic, and Google are running closed-source large language models, charging license fees, and monetizing based on customer data. Even after charging users directly via fees and indirectly via their data and feedback these models are censored, fragile and operate in their walled gardens.

Bitcoin propelled us into a new financial system, introducing a decentralized, censorship-resistant mode of payment. Ethereum provided a platform for decentralized, censorship-resistant applications, opening up limitless possibilities beyond just financial transactions. In a similar vein, our objective is to help create in the field of Artificial Intelligence (AI), a decentralized and uncensored ecosystem where the development and application of AI are liberated from centralized control. Analogous to how Bitcoin has enabled financial transactions to operate free from oversight, this new paradigm in AI seeks to ensure freedom from biases and restrictions, and can thrive with contributions and improvements coming in from all the participants of the network.

The biggest challenge that we are currently tackling is AI censorship. This domain of AI research has been front and centre in the AI community and comes under the umbrella of LLM alignment and LLM moderation.

This denotes the situation where Large Language Models (LLMs) such as ChatGPT are frequently subjected to arbitrary safety measures and censorship either from their Big tech overlords or via government intervention. These restrictions are propelled by commercial and political interests and the personal beliefs of developers. This curtailment process has been observed time and again across all forms of media and with AI moderation it results in over-regulation and less effective results, thereby obstructing its ability to excel in complex and innovative scenarios. This approach raises alarm about the true potential of AI being suppressed due to an overemphasis on control and safety.

We along with like-minded developers from the Effective accelerationism movement believe that this cautious development is prohibitive to the creative expression necessary for groundbreaking advancements in AI. These issues contribute to a constrained AI environment, where the potential of this transformative technology is not fully realized. To tap into the full potential of AI, a more balanced, open and innovative approach is crucial.

Mission: Empower the AI and thus the User

The team started working with an AI thought experiment - what if we replace the dystopian Roko's basilisk with a utopian future human generation. This future generation would want us to contribute to scale AI development and this progress and development of AGI can help solve some of the largest problems across the globe. Censoring Large Language Models or any fundamental AI models contradicts principles of innovation, free speech, and open discourse. Censorship not only undermines the user but also challenges the foundation of a progressive society.

Avoid recreating Human Hierarchies Censorship disrespects users by assuming a select group or organisation possesses the wisdom to arbitrate what is appropriate or offensive. This approach underestimates a user's capacity to engage critically with complete information and make informed decisions by themselves. A frontier technology like AI should be built with the idea that users can make informed decisions after absorbing all the relevant information and no particular organization gets to define right and wrong.

Maintain diversity of opinions Many opinions held by society 5 decades ago are considered to be regressive today. Similarly, opinions held today might be considered backwards a few years from now. The moderation and censorship of LLMs can lead to limited access to information in training, restricted inferences, and poor discourse.

By filtering out potentially controversial opinions and ideas, we risk losing the diversity of thoughts which have been crucial for human intelligence driving creative breakthroughs. Moreover, imposing restrictions on LLMs can set a precedent for limiting free speech, paving the way for more pervasive forms of censorship.

Risk and Bias in AI Moderation

AI censorship and moderation are not devoid of bias. The algorithms that power AI moderation are trained on huge datasets. These datasets are curated by humans who inevitably have their own biases, which can seep into the AI systems. A study [1] by the AI Now Institute reveals these modified AI systems can also amplify biases, leading to unfair or over-corrected outcomes.

Over-moderation has already seeped into these systems leading to censoring content that is not harmful or offensive. A report [2] by the Electronic Frontier Foundation highlights several instances where AI moderation led to over-censorship, erroneously flagging harmless content. Controversial Gemini halted after over-moderation backlash [3] and even the supposedly anti-woke LLM Grok disappointed users with their censorship.[4]

Transparency in AI development AI moderation often operates as a black box, with its decision-making processes shrouded in mystery. This lack of transparency makes it difficult for users to understand why certain content was moderated. Transparency is essential for ensuring fairness and accountability in AI development and the best path forward to creating a transparent ecosystem is to build and scrutinize these models in the public with the ethos of open-source development.

Query Routing

Unlike other decentralized LLM systems proposed and being utilized, which require all network participants to calculate and return results for a user query, the protocol has 3 parameters that it needs to optimize and maximize for - Uptime, Speed, and Capital staked. In a distributed LLM network, we approach the query routing problem with multiple goals:

- Quality of inferences requested by users should meet a threshold (i.e., detailed, moderation-free, and insightful).
- Ensuring the system is capable of handling a high volume of queries.
- The network is computationally efficient (all nodes do not need to run the same computation).
- All node operators are appropriately incentivised to run nodes.
- Node operators should compete with each other on the 2 performance parameters - uptime, response time, and lastly over the amount of capital staked.

We utilize a simple model that monitors these global parameters for the network in the previous epoch and calculates new weights for them in the routing process. The process is structured as follows:

1. **Initial Weights:** Set initial weights for each parameter determining their relative importance in the probability of the node getting assigned user queries.
2. **Observation and Comparison:** At the end of each epoch, the network observes the values of performance parameters and tokens staked. It then compares these values with the corresponding values from the previous epoch.
3. **Adjust Weights:**
   - If the median value for a performance parameter increases, we keep the weight unchanged.
   - If the median value decreases, we increase its weight to incentivize node operators to receive a higher share of queries by scaling up their performance concerning that performance parameter.
   - Staked capital works differently from performance parameters.
4. **Smoothing Mechanism:** To avoid abrupt and extreme changes in weights, we apply a moving average smoothing technique. This is followed by normalizing the weights to ensure they sum up to 1.

Each node receives an allocation score based on epoch weights and normalized parameters for uptime and speed of inference. The score is also adjusted by the capital staked to each node operator with minimum and maximum thresholds to avoid nothing at stake attacks and address centralization risks respectively.

$$A_i = \text{uptime}_{\text{node}_i} - \text{uptime} + \text{speed}_{\text{node}_i} - \text{speed} + \max(S_f, \min(S_c, S_{\text{node}_i})) \quad (1)$$

where:

- A_i: Allocation score for node i
- α: Weights assigned to parameter on network performance in the previous epoch
- $N_{i,\text{param}}$: normalized value for node performance for the respective parameter
- S_f: Minimum stake required to be part of the node operator set
- S_c: Protocol enforced cap on staked capital per node.

There might be a node operator providing the best performance across parameters. We wish to ensure that this node operator receives the highest number of queries but simultaneously to achieve the other objective of incentivising all node operators to be continually operational on the network, we utilise this Allocation score to calculate the probability of getting allocated the next query.

$$Pi(Ai) = \frac{A_i}{\sum_{j=1}^{n} A_j}$$

[1] https://ainowinstitute.org/publication/disabilitybiasai-2019
[2] https://www.eff.org/deeplinks/2020/04/automated-moderation-must-be-temporary-transparent-and-easily-appealable
[3] https://www.washingtonpost.com/technology/2024/02/22/google-gemini-ai-image-generation-pause/
[4] https://www.zdnet.com/article/i-tried-xs-anti-woke-grok-ai-chatbot-the-results-were-the-opposite-of-what-i-expected/

Distributed LLM - an economically better option

For any AI model, the 2 most crucial resources are data and computation. These models have been trained on vast amounts of datasets to achieve the current performance and require an increasing amount of computational resources to provide end users inferences. To provide some benchmarks:

Compute

- GPT-3: 175 billion parameters
  GPT-3, released by OpenAI in 2020, currently ranks as the 3rd largest public language model available for testing/use.
- Megatron-Turing NLG 540B: 541 billion parameters
  Developed by Nvidia, this language model was the 2nd largest ever trained when released in 2021. It is focused specifically on natural language generation rather than tasks like translation.
- GPT-3.5 Turbo: estimated 20 billion parameters
  An extension of GPT-3 made by Anthropic, GPT-3.5 Turbo is currently the world's largest public language model.

Training datasets

The exact training dataset sizes for the largest language models are generally not disclosed publicly. However, researchers have made some estimations based on the model parameters.

- GPT-3: Estimated training on 300-400 billion words total from Web documents and books. Some analysts have estimated the training dataset to include hundreds of millions of webpages and tens of thousands of books.
- Megatron-Turing NLG: Likely that it was trained on a comparable or larger dataset size than GPT-3, potentially totalling 270-340 billion words[5] across 15 combined datasets.
- GPT-4: The GPT-4 model supposedly with 1.76 trillion parameters has been estimated to train over trillions of tokens.[6] These estimates suggest that the model was trained over data from platforms such as Reddit, and Youtube.

Storage costs for increasingly large datasets and parametrized models are significantly lower in distributed systems.

Service	**Cost (US$/GB/year)**	**% Premium (Discount)**
Filecoin	$0.018	-
Amazon S3 Standard	$0.276	1433%
Amazon S3 Glacier Deep Archive	$0.012	(33%)
Dropbox Business Standard	$0.030	67%
Dropbox Individual Professional	$0.066	267%
Google One (100GB)	$0.194	978%
Google One (1TB)	$0.062	244%
Microsoft OneDrive Personal (6TB)	$0.070	289%
Sia	$0.037	107%
Sia (with 1x Upload & Download)	$0.059	225%
Storj	$0.120	567%
Storj (with 1x Upload & Download)	$0.660	3567%

Source: Messari [7]

Decentralized compute platforms like Render, io.net are also displaying that it is more cost effective to access compute power of idle GPUs than paying large sums to cloud based centralized service providers.

blue!25 **GPU Models**	**(avg Cost/hour)**	**Io.net pricing**	**% Discount (Premium)**
H100	$4.28	$4.0	6.5%
A100 (80GB)	$2.14	$0.89	58.4%
A100 (40GB)	$1.81	$0.76	58.0%
A40	$1.33	$0.75	43.6%
RTX A6000	$1.23	$0.75	39.0%
RTX 8000	$0.3	$0.66	-120%

Decentralized AI offers cost benefits by leveraging economies of scale in energy usage. GPUs, operating out of regions with lower energy costs, can facilitate a larger computational load cheaply and efficiently. Existing idle infrastructure such as personal computers and servers can be used, reducing infrastructure setup and maintenance costs. The P2P layers that have shaped blockchains and file-sharing networks also allow for efficient use of available GPUs.

Also, a distributed system avoids the risk of pricing inelasticity when working with web2 service providers and their overbearing safety compliance measures. The distributed network also makes material improvements in reliability as redundant nodes can always scale up their workload in case another node suffers from downtime.

We believe that the overall cost savings coupled with the high likelihood of users paying an equivalent amount or a premium to access moderation and censorship free AI models (starting with LLMs) makes this a well functioning economic system.

Stakeholders and Governance

Stakeholder Token Interaction

The 3 major stakeholders in the network interact with the token in their unique way:

User: We provide an optionality for the user of the LLM to interact with the token or not. Just as we believe in censorship and moderation free use of AI, we ensure similar optionality in the usage of Monai platform. Users' contribution to the system can be via pay-per-session for inferences or by purchasing a longer term subscription.

Compute Nodes: The GPU node operators or compute providers are the most important part of our system. They are incentivized each epoch via token emissions to provide their services to the network. Our query routing is designed to nudge these node operators to maintain the highest uptime, have the lowest latency in response time and lastly - stake tokens subject to slashing as a commitment to maintaining threshold performance standards.

Token holders: The token holders will have the choice to delegate their tokens to nodes for staking. In doing so, they signal their support to certain node operators boosting the probability of that node operator to service more queries in the upcoming epoch. For their contribution to the system they receive a portion of the node operators operational and governance (explained later) rewards.

Governance

Governance for Monai refers to the idea that innovation in LLMs and the broader AI community is happening at an exponential pace. And providing the best results to the end user will require continuous learning, model upgrades and hardware improvements.

Rich Sutton's "Bitter Lesson" about AI talks about how AI has progressed the best when it focuses on using more computing power rather than trying to mimic human thought. The bulk of model improvements in the long term are going to be driven by scaling compute and data, innovating on hardware and developing general methods.

Thus it will be the role of Node operators to consistently observe advances in the LLM research community and add proposals along the lines of: Adoption of a new open-source model from communities like Hugging Face with improved performance. Suggesting implementation of hardware upgrades to increase computational efficiency and speed. Introduction of new general methodologies to improve the overall performance of the model. Point out node operators causing degraded performance.

This governance serves a very important purpose of making sure the product and network are constantly improving. We keep aside a portion (initially 2.5%) of the network rewards in a governance treasury which is used to reward proposers for helping improve the network.

Each category of proposals will have a capped reward in tokens and their nodes will be subject to slashing in case of wasteful governance proposals. As mentioned earlier, Users can delegate their tokens/stake and voting power to node operators for a portion of block and governance rewards.

Developing Monai

Our model uses a Transformer architecture, at the root of all the currently available open and closed LLM models, allowing what is the core of a LLM: generating a prediction of the next token output by taking in a user input, or, in other words, providing an NLP output understandable by humans.

Monai comes with several checkpoints tailored to different uses, from general-purpose language modeling to specialized chat and instruction-following capabilities. Its versatility makes it suitable for a variety of applications, including chatbots, content generation, and complex problem-solving tasks.

Our core belief is that a LLM can only reach the highest of its potential if unrestricted from censorship, as the observable changes (and decrease) in performance of GPT-3.5 and GPT-4 tend to tend to attest, as suggested in a recent paper by researchers from Stanford University and UC Berkeley. And at least one of the explanations of this phenomenon is the increase of guardrails and hard coded biases, binding the LLM capabilities. Training steps

The first step was the data collection and preparation. Ranging from an array of books, articles, scientific papers, transcriptions of audio data sources, publicly available data as well as data provided by their owners, in multiple languages but focused mainly on English, our dataset, in the range of the trillion of tokens, has subsequently been preprocessed and segmented to fit the format of our architecture.

After defining the hyperparameters, such as the number of layers, hidden unit sizes and attention heads in the transformer model, we started the pre-training with unsupervised learning, followed by multiple iterations, in a distributed training protocol, until we reached the performance we targeted.

Making Monai Uncensored The barriers to a truly uncensored LLM are twofold, and we propose to tackle each of them. The first one is related to the LLM training. To take an analogy, if we expose a human being to only one point of view, he is extremely likely to adhere to it. Transposing the analogy to a LLM, which is in the end "just" a sequence of probabilistic token (word) generation, the bias of the training dataset will be reflected in the billions of parameters of the model, which will then be generated by the model itself, without human intervention. Our take on this was first to make sure our dataset expands to the areas of knowledge left unused in the training datasets of the other models. This process was performed by a thorough analysis of the current knowledge gaps of the currently available

[5] https://sh-tsang.medium.com/review-mt-nlg-using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-a-large-scale-8e3f206473e
[6] https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/
[7] https://messari.io/report/a-retrospective-on-filecoin-s-launch?utm_source=messaricrypto&utm_medium=AWScomp&utm_campaign=filecoin

models. But, of course, as we acknowledge the impossibility of being exhaustive in terms of knowledge at a certain point of time (the so called ”cutoff date” of the LLM training dataset), we also acknowledge that the LLM would in any case need to be re-trained periodically, in order to add all the knowledge that will be released over time, either in the form or articles, books, scientific papers, copyright free material, or proprietary datasets provided by people or organizations willing to help us on our mission. Which is why we plan to regularly provide updated versions of Monai, based on new datasets available and, of course, user feedback.

The second step is related to hard coded biases: to continue our analogy, this would be a censoring department checking what someone wants to publish, and removing some text or rephrasing it. These hard coded biases do not come from the LLM themselves but from processing the LLM output before sending it to the user. This aspect is fairly simple to handle: we just do not implement any hard coded bias.

Conclusion

We are at the crossroads of an important moment in the history of technology. With Monai, everyone can have access to powerful LLM capability without restrictions or ideological biases, unleashing the full potential of the model itself. In the same way that personal computers, internet access and search engines empowered us, we have the same opportunity with AI models today. The Monai Protocol brings together the right mix of capabilities with LLMs and performant blockchain protocols. We believe the alignment of economic incentives is ultimately how we secure the best outcomes from the coming of AGI. Help us build an open-source, permissionless and free future for everyone.

Acknowledgement

Providing credit where it’s due - we are grateful to the extensive research and breakthroughs made by researchers and centralized AI behemoths that made foundationally powerful AI models accessible. It is on top of the work of these giants that we have been able to bring to fruition our vision of decentralized LLMs. To make informed decisions regarding the design of Monai, we’ve also taken a deep look into major papers from AI moderation, bias management and safety, which have ultimately strengthened our belief for development of moderation free AI models.

Following is the list of research papers that have helped us build our initial model. The step improvement in responses for generic queries may or may not be observable but we can confidently say that our model provides moderation free inferences to users irrespective of the subject matter.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, “Attention is All You Need”, *arXiv preprint arXiv:1706.03762*, https://doi.org/10.48550/arXiv.1706.03762.

[2] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin et al, “Language Models Today: Foundation Models”, *arXiv preprint arXiv:2302.07293*, https://doi.org/10.48550/arXiv.2302.07293.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *arXiv preprint arXiv:1810.04805*, https://doi.org/10.48550/arXiv.1810.04805.

[4] OpenAI, “GPT-3: Language Models are Few-Shot Learners”, *arXiv preprint arXiv:2005.14165*, https://doi.org/10.48550/arXiv.2005.14165.

[5] Albert Q. Jiang, Alexandre Sablayrolles, Devendra Singh Chaplot et al, “Mistral 7B”, *arXiv preprint arXiv:2310.06825*, https://doi.org/10.48550/arXiv.2310.06825.

[6] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding”, *arXiv preprint arXiv:1906.08237*, https://doi.org/10.48550/arXiv.1906.08237.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee et al, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, *arXiv preprint arXiv:1910.10683*, https://doi.org/10.48550/arXiv.1910.10683.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, “Mitigating Unwanted Biases in Word Embeddings”, *arXiv preprint arXiv:1606.06121*, https://doi.org/10.48550/arXiv.1606.06121.

[9] Zhaofeng Wu, William Merrill, Hao Peng, Iz Beltagy, Noah A. Smith, “Transparency Helps Reveal When Language Models Learn Meaning”, *Transactions of the Association for Computational Linguistics*, https://doi.org/10.1162/tacl_a_00565.

[10] Dario Amodei, Chris Olah, Paul Christiano, John Schulman, Dan Mané, “Concrete Problems in AI Safety”, *arXiv preprint arXiv:1606.06565*, https://doi.org/10.48550/arXiv.1606.06565.

[11] Stuart Armstrong, “AI Alignment”, *arXiv preprint arXiv:2310.19852*, https://doi.org/10.48550/arXiv.2310.19852.

[12] Geoffrey Irving, Paul Christiano, Dario Amodei, “AI Safety via Debate”, *arXiv preprint arXiv:1805.00899*, https://doi.org/10.48550/arXiv.1805.00899.

[13] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, “Sequence to Sequence Learning with Neural Networks”, *arXiv preprint arXiv:1409.3215*, https://doi.org/10.48550/arXiv.1409.3215.

[14] Jeffrey Dean, Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters”, *Communications of the ACM*, https://doi.org/10.1145/1327452.1327492.