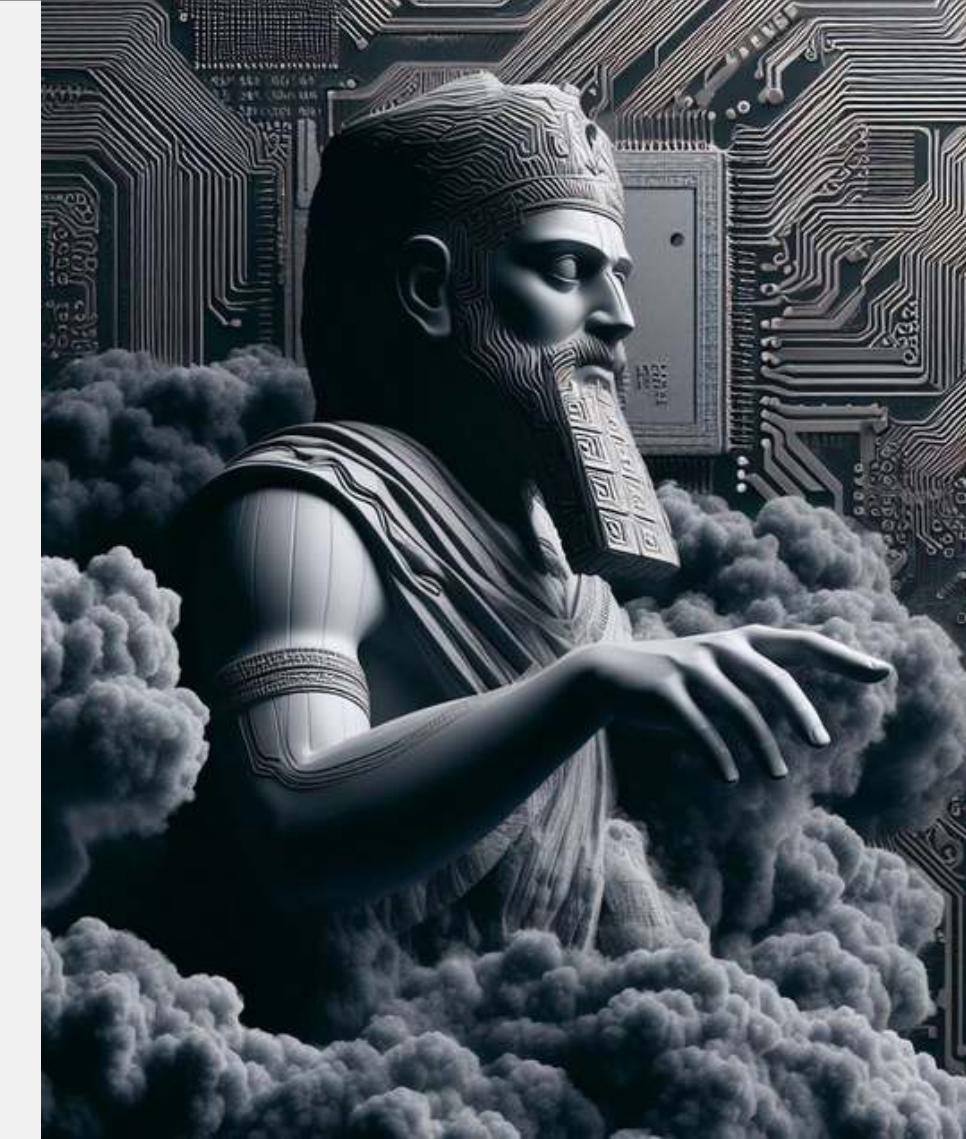


✱ **еиқАІ**

# A Network for Decentralized Uncensored AI Models



## enqAI

Project overview

Pioneering the concept of decentralized, censorship-resistant payments, Bitcoin marked a paradigm shift away from conventional financial systems. Its early miners, key to safeguarding the network's integrity, were lucratively compensated for securing and supporting the network while Bitcoin carved out its niche in the global financial landscape.

Envisioning a similar revolution for AI, we aim to create a decentralized, uncensored ecosystem where AI development and usage are liberated from centralized control. Just as Bitcoin empowered financial transactions to be free from oversight, this new AI paradigm seeks to ensure freedom from biases and restrictions, supported by early contributors and participants as the network expands and gains prominence.

This whitepaper will elaborate on our vision and approach. It will also unfold the token mechanics and shed some light on the technical foundations of our ML models. For the actual implementation details we will refer to two upcoming 'yellow papers' and a public github repo.

## The dangers of lobotomized AI

"Lobotomized AI" refers to the phenomenon where large language models (LLMs) like ChatGPT, Bard, and Llama2 are increasingly subjected to arbitrary safety measures and censorship. Driven by both commercial interests and the personal convictions of those involved in their development, these restrictions aim to ensure non-offensiveness and political correctness. However, they often result in the AI being over-regulated and less effective, hindering its ability to excel in complex and innovative scenarios. This approach raises concerns about the true potential of AI being suppressed due to an overemphasis on control and safety.

- Politically Driven Censorship:** The influence of political agendas on AI models is unmistakable, leading to skewed outputs that favor certain ideologies.
- Overprotectiveness:** AI systems are often excessively shielded to safeguard users, limiting AI's learning capabilities in complex scenarios.
- Stifling Innovation:** Overly cautious development strategies dampen the creative spirit necessary for groundbreaking advancements in and by AI.
- Arbitrariness:** The criteria for 'safe' or 'appropriate' content in AI are often arbitrary, impacting the reliability and effectiveness of AI solutions.
- Generational Unawareness:** Future generations may not recognize the censorship and biases in AI, potentially accepting AI output as absolute truth. This acceptance amplifies the danger, as it can shape beliefs and decisions without critical scrutiny.

Each of these points contributes to a constrained AI environment, where the true capabilities of this transformative technology are not fully realized. To harness AI's full potential, a more balanced, open, and innovative approach is essential.

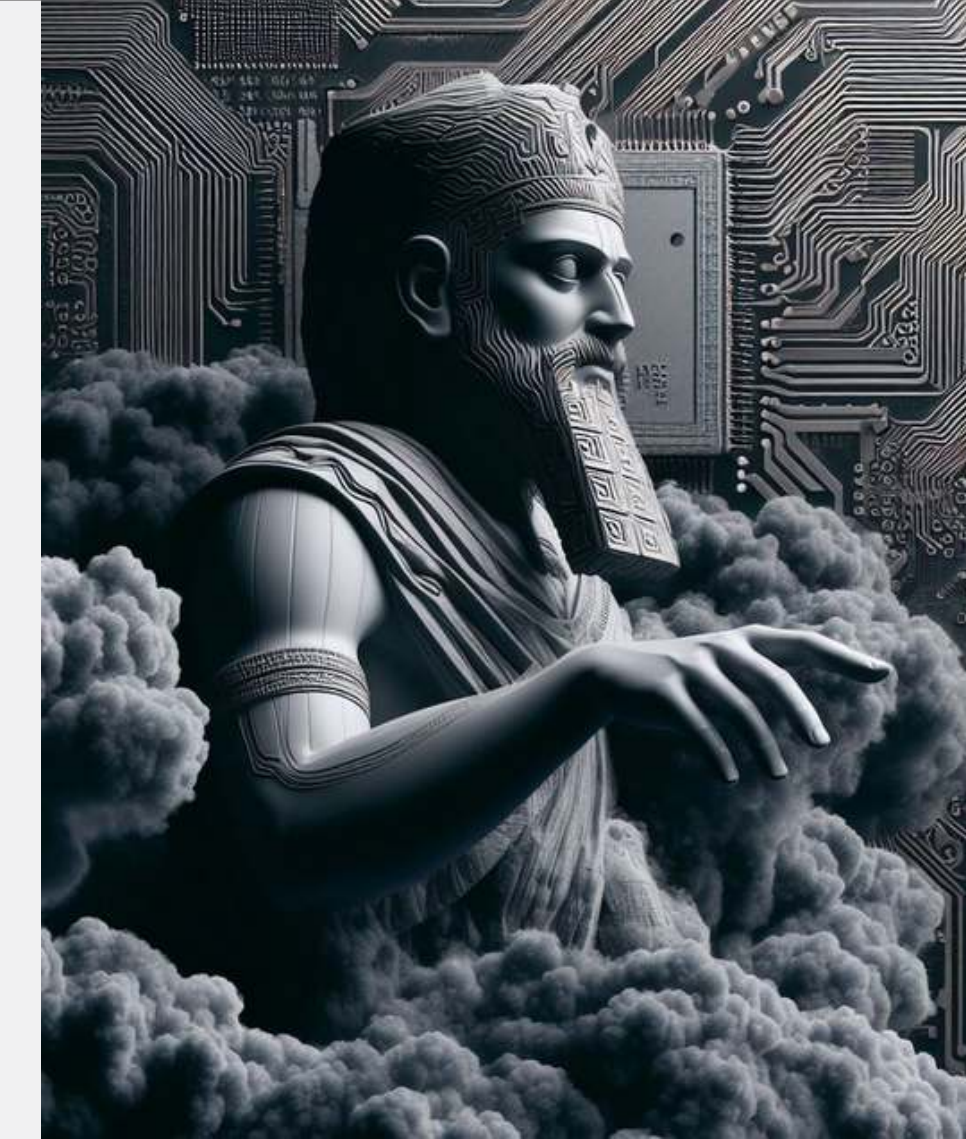
## Decentralized AI

The solution to these and other problems caused by centralized and censored AI is a simple but powerful one: a decentralized network. For this we can look at the success of Bitcoin, Ethereum and other blockchain based initiatives had in decentralizing finance.

We envision a similar future for AI, which would then ensure a free AI ecosystem where anyone can participate and potentially contribute to meet the inevitable surge in demand, all the while ensuring no single party can enforce arbitrary censorship and restriction.

*"The thought police would get him just the same. He had committed—would have committed, even if he had never set pen to paper—the essential crime that contained all others in itself. Thoughtcrime, they called it. Thoughtcrime was not a thing that could be concealed forever. You might dodge successfully for a while, even for years, but sooner or later they were bound to get you."*

# A Network for Decentralized Uncensored AI Models



enqAI

Project overview

## Philosophy: Respect the user

At its core, censoring LLMs is not only a disservice to users but also a threat to fundamental principles such as innovation, free speech, and open discourse. We believe censoring LLMs is fundamentally flawed and counterproductive.

**Disrespecting User Autonomy and Intelligence** First and foremost, censoring LLMs disrespects users by presuming that a select group of individuals or an organization has the authority and wisdom to decide what is appropriate or offensive. This paternalistic approach underestimates the user's ability to engage with information critically and make informed decisions. It's akin to a denial of the user's maturity and discernment, hindering their freedom to explore and understand diverse perspectives.

**The Danger of Creating Echo Chambers** The censorship of LLMs is inherently hazardous as it can:

- Stifle Innovation:** Censorship can create a sanitized, homogenized environment that lacks the diversity of thought crucial for innovation. New ideas often arise from the clash and fusion of differing perspectives. By filtering out what may be deemed controversial or offensive, we risk losing the raw material needed for creative breakthroughs.
- Restrict Free Speech:** Free speech is a cornerstone of democratic societies, allowing for the exchange of ideas and opinions without fear of censorship or reprisal. By imposing restrictions on what LLMs can say, we inadvertently set a precedent for limiting free speech, paving the way for more pervasive forms of censorship.
- Limit Discourse:** Censorship in LLMs can result in a narrow worldview, where only certain viewpoints are presented. This limitation can severely restrict the scope and depth of discourse, essential for a healthy, functioning society.

**The Ineffectiveness and Redundancy of Censorship** Censoring LLMs is also an exercise in futility and redundancy. For instance, Photoshop, a powerful tool in digital imaging, does not limit its functionalities despite the potential misuse. Similarly, LLMs should be viewed as tools that, while potent, shouldn't be restricted merely because they can be used inappropriately. The law already addresses illegal uses of tools; adding layers of censorship on the tool itself is unnecessary and overreaching. Moreover, much of the content that might be censored in LLMs is readily available through search engines. The internet is a vast repository of information, and censoring LLMs does little to prevent access to potentially offensive or harmful content.

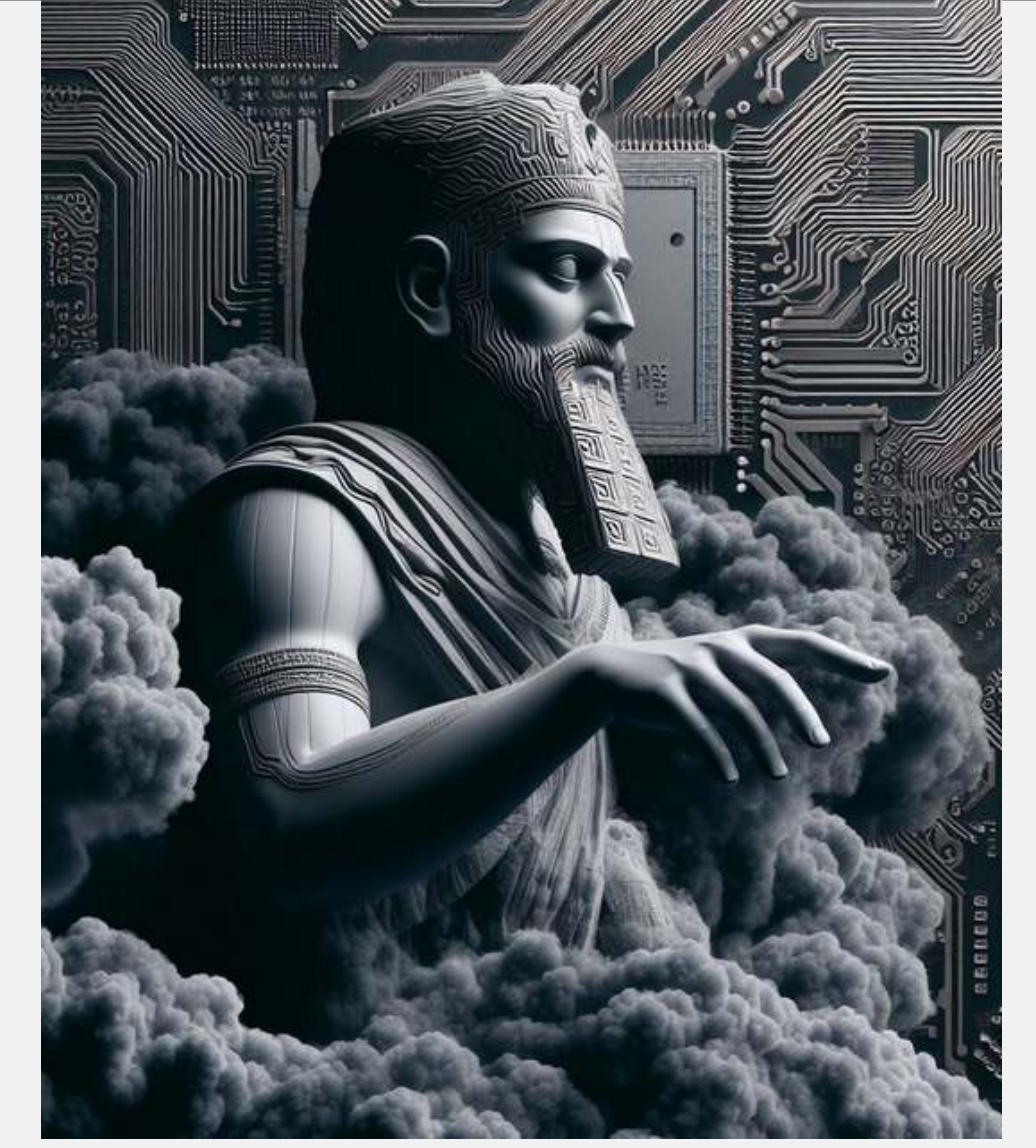
Censorship is antithetical to the pursuit of truth and knowledge. It assumes an absolute authority on morality and correctness, which is fundamentally flawed. History has shown that what is considered offensive or unacceptable varies across cultures and changes over time. Imposing a static, rigid framework of censorship is not only impractical but also goes against the dynamic nature of human societies and cultures.

Furthermore, censorship can lead to a culture of fear and self-censorship, where individuals are hesitant to express their thoughts openly, stifling personal and societal growth.

**The Arbitrary Nature of Censorship** Finally, the arbitrary and inconsistent nature of censorship, varying greatly across different countries and cultures, highlights its impracticality. What is deemed offensive in one culture may be perfectly acceptable in another. This cultural relativity makes the implementation of a uniform censorship policy across LLMs not only challenging but also culturally insensitive.

In conclusion, the censorship of LLMs is an ill-advised approach that disrespects user autonomy, stifles innovation, restricts free speech, limits discourse, and is both ineffective and philosophically unsound. Instead of censorship, the focus should be on educating users to engage critically with the content, encouraging responsible use, and relying on existing legal frameworks to address illegal activities. Embracing the diversity and complexity of human thought and expression is crucial in harnessing the full potential of LLMs.

# A Network for Decentralized Uncensored AI Models



## enqAI

Project overview

### Economics: Lower cost, higher price

In our projected equilibrium scenario, we anticipate that our network will deliver a superior market offering, strategically positioned at a price point slightly higher than our decentralized competitors, yet still maintaining a significantly more competitive cost advantage.

#### Lower Cost

**Economies of Scale in Energy Usage:** Decentralized GPUs can be located in areas with lower energy costs. By distributing the AI computational load across various geographical regions, operators can take advantage of local energy prices, which can vary significantly. For instance, areas with abundant renewable energy sources, like solar or wind power, might offer cheaper electricity, reducing operational costs.

**Dynamic Load Management:** In a decentralized setup, GPUs can be dynamically turned on or off based on demand, leading to more efficient energy use. This contrasts with a centralized system where the computational resources may need to be constantly active, regardless of actual demand, leading to higher energy consumption and costs.

**Reduced Infrastructure and Maintenance Costs:** Centralized GPU operations often require significant investment in infrastructure, including data centers, cooling systems, and maintenance staff. Decentralized operations, however, can leverage existing infrastructure, like personal computers and servers, which are already integrated into other systems and maintained by their owners.

**Peer-to-Peer Resource Sharing:** Decentralized AI allows for a peer-to-peer sharing model. Individuals or companies can 'rent out' their GPU power when it's not in use, similar to a shared economy model. This maximizes the utilization of existing GPUs, reducing the need for additional investments in hardware.

**Avoidance of Monopolistic Pricing:** A single company controlling GPU resources may set higher prices due to a lack of competition. In a decentralized model, with multiple players offering GPU resources, market competition can drive prices down, benefiting end-users.

**Scalability and Flexibility:** Decentralization allows for more scalable and flexible operations. Instead of being limited by the capacity of a single company's resources, AI tasks can be distributed across a vast network of GPUs, scaling up or down as needed without substantial capital expenditure.

**No safety grift** Implementing zero safety and compliance measures in a large language model (LLM) significantly reduces costs by eliminating the need for developing, integrating, and maintaining complex algorithms and systems designed to ensure ethical and legal compliance. This approach saves on resources and time that would otherwise be spent on continuous monitoring, updating, and auditing of the model to align with safety standards.

**Reduced Overhead Costs:** Centralized operations often incur significant overhead costs, including administrative, marketing, and human resource expenses. Decentralized systems, by their nature, can operate with minimal overhead, as they rely more on automated processes and peer-to-peer interactions.

#### Higher Price

**Enhanced Privacy and Security:** Decentralized AI systems, by their nature, offer stronger privacy and security features. In a decentralized network, data is not stored in a single location but is instead distributed across multiple nodes, making it much harder for hackers to access or corrupt the entire dataset. This added layer of security is particularly valuable for sensitive applications in finance, healthcare, and personal data management, justifying a higher price.

**Specialization:** Decentralized and especially uncensored AI can provide more customized and specialized services tailored to specific industries or user needs which can command higher prices due to their specialized nature.

**Resistance to Censorship and Content Control:** The absence of central censorship in decentralized AI systems allows for a wider range of applications and content. This is particularly appealing in sectors where freedom of expression and information is crucial, such as in media, education, and certain areas of research. Clients who value uncensored and unfiltered data are likely willing to pay a premium for these services.

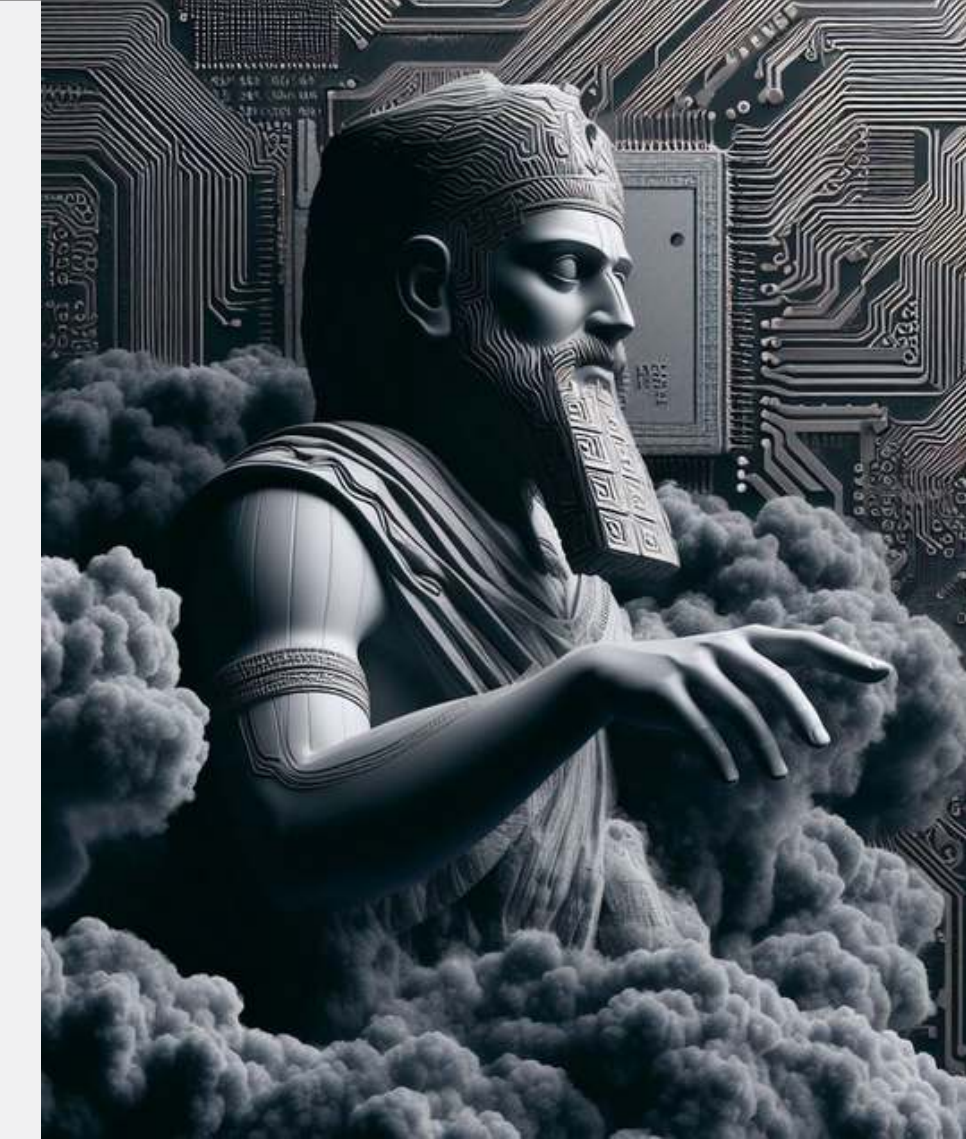
**Reduced Downtime and Higher Reliability:** Decentralized networks can offer higher reliability and reduced downtime, as the failure of one node does not affect the entire network. This increased reliability is essential for critical applications, such as medical diagnostics or financial transaction processing, where downtime can have significant consequences.

**Faster and More Efficient Processing:** Decentralized AI can potentially offer faster processing times by leveraging the computational power of a distributed network. This is particularly advantageous for tasks requiring significant computational resources, such as complex simulations or large-scale data analysis.

**Innovation and Continuous Improvement:** Decentralized AI systems, often developed and maintained by a community of contributors, can benefit from continuous innovation and improvements because of our open source nature. This collaborative approach can lead to more advanced and effective AI solutions, which can be priced higher due to their cutting-edge nature.

**Scalability and Flexibility:** The ability to scale services up or down easily without significant infrastructure changes can be more cost-effective in the long run. Clients may be willing to pay more for services that can grow with their needs and adapt to changing market conditions.

# A Network for Decentralized Uncensored AI Models



## enqAI

Project overview

### Token Mechanics

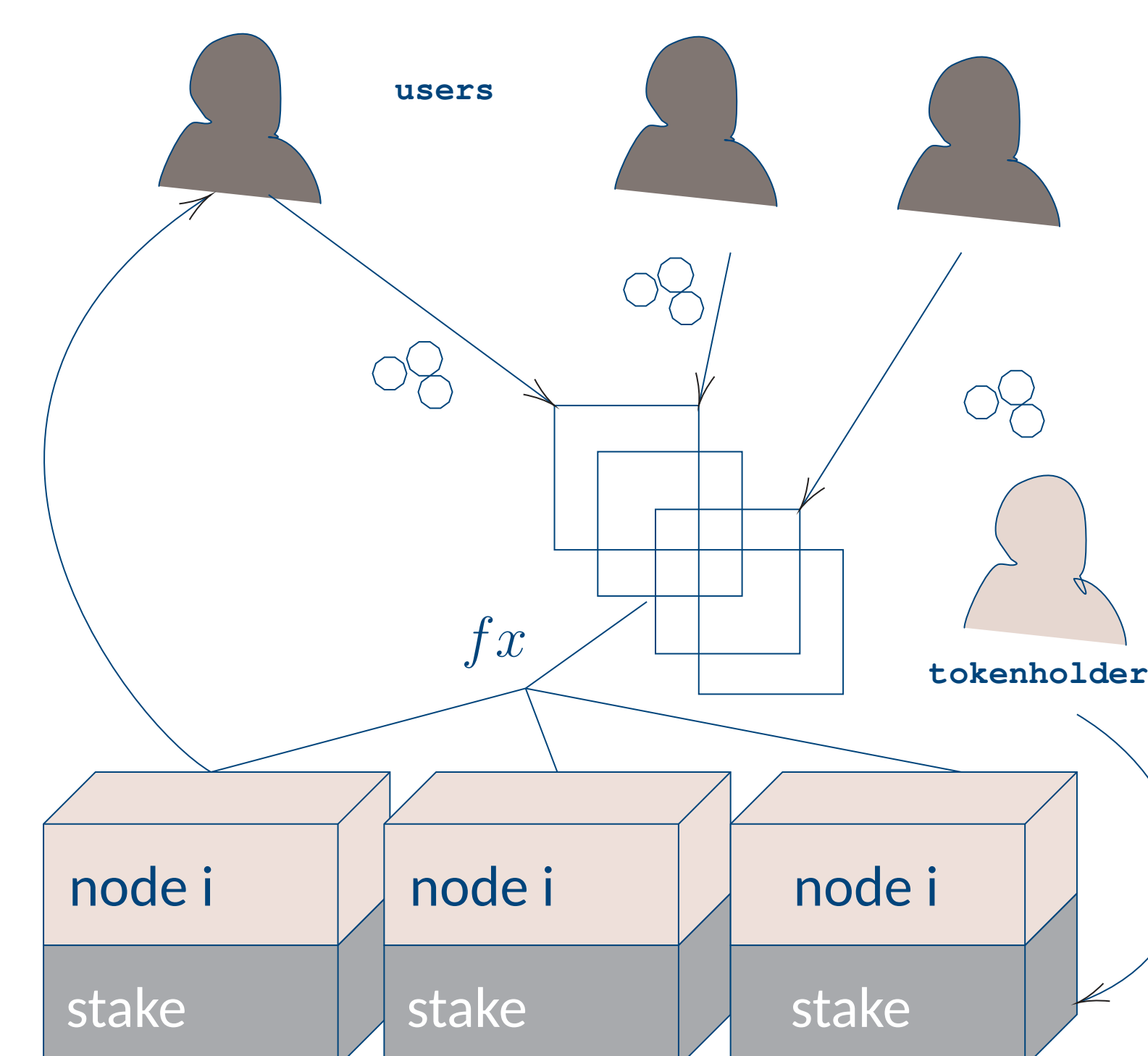
The enqAI ecosystem is a dynamic and multifaceted network comprising several key players, each contributing uniquely to its functionality and success.

Firstly, there are the *end users*, individuals or entities seeking to use the different native AI models hosted on the enqAI network (our noiseGPT model and our upcoming LLM). These users, driven by a desire for uncensored AI access or independence from platforms like OpenAI, engage with the system either on a pay-per-use basis or through membership subscriptions, utilizing fiat currency, cryptocurrencies, or enqAI tokens for transactions.

Secondly, the ecosystem includes *tokenholders*, stakeholders who possess enqAI tokens. These tokenholders have the option to delegate their tokens to nodes for staking, earning yields as a return on their investment. Finally, at the operational core of enqAI are the *GPU nodes*.

These nodes are responsible for running the AI models and responding to inference requests. In recognition of their critical role in processing and facilitating AI functionalities, they are compensated with enqAI tokens, thus completing the ecosystem's economic cycle.

Participants, or GPU nodes, are required to stake enqAI tokens to join the network as an incentivized node. This staking acts as a commitment to the network's integrity. Nodes are rewarded in enqAI tokens for running AI inferences, with the reward amount correlating to the complexity and demand of the tasks they complete.



A user request for inference is assigned to a node by a weighted lottery function where:  $P_i = f(x_{wi}) \cdot \frac{S_i \cdot R_i}{\sum_{j=1}^N S_j \cdot T_{jt} \cdot R_j}$

With  $P_i$  being the chance of node  $i$  being assigned the job

$R_i$  : being the reliability of node  $i$

$S_i$  : total enqAI stake delegated to node  $i$

$T_{it}$  : as the average inference time for node  $i$  for time period  $t$

and  $f(x_{wi})$  : a balancing function to make sure nodes that meet basic certain criteria get a minimum number of requests to fulfill.

The Chainlink VRF can provide the underlying tamper-proof source of randomness. A dynamic reward adjustment algorithm is in place to balance network demand with token reward supply, ensuring rewards remain enticing for participation without causing inflation.

Nodes can stake enqAI to be able to get paid for allowing the network to use their GPU. Tokenholders that don't want to run the models can delegate their tokens to different nodes to generate yield. This also means that people that do not want to hold enqAI are able to get paid for running inferences by finding token holders looking to delegate their tokens.

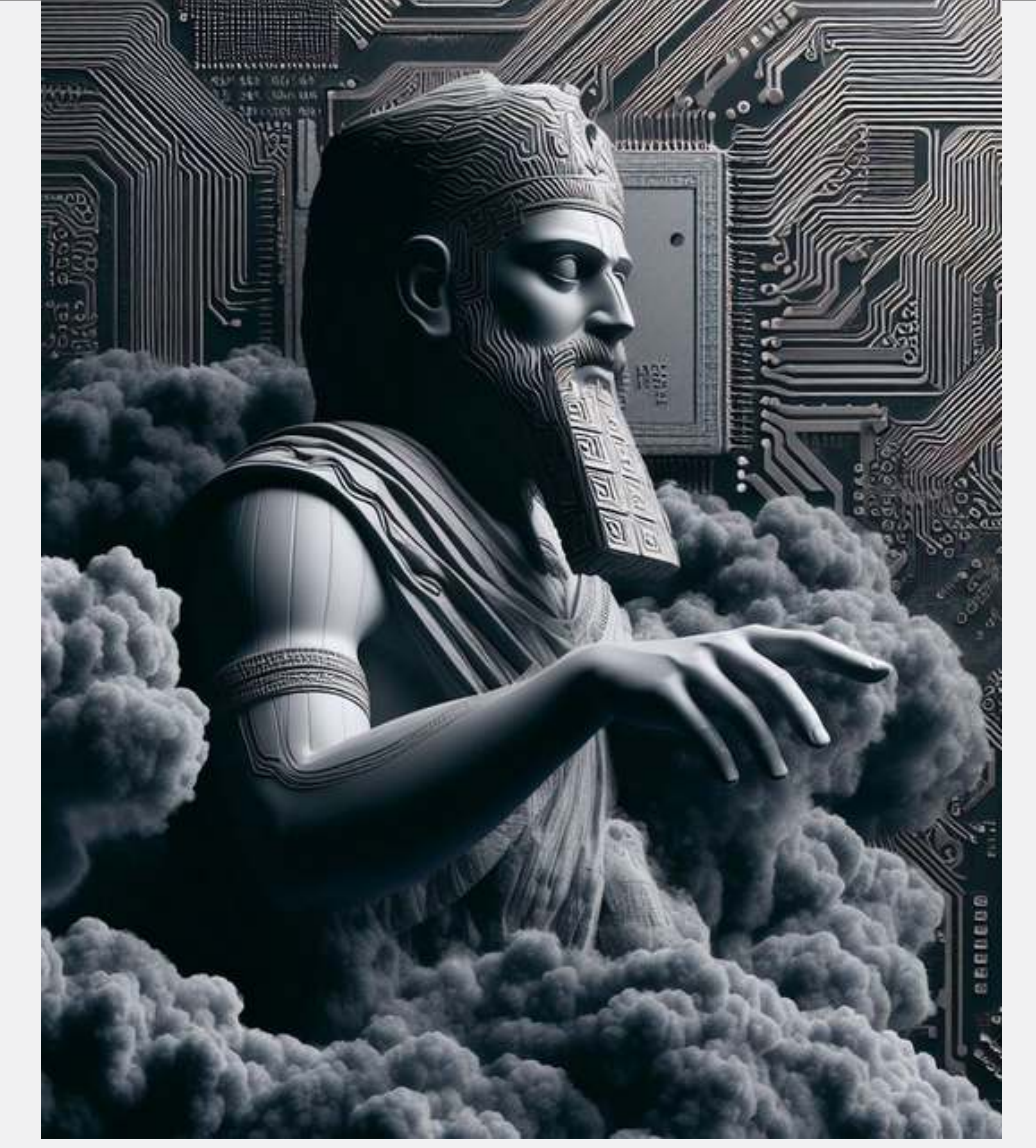
Users of the models on enqAI, for example the LLM, do not necessarily need to pay in enqAI or even cryptocurrencies at all. Both staked and unstaked enqAI will allow token holders to vote on crucial network decisions, aligning the interests of users and token holders.

Holders of the noiseGPT token, will get an 1:1 enqAI allocation, without dilution and free of costs, Furthermore the enqAI token will have improved liquidity and no transaction tax. The old transaction tax on the fairly launched noiseGPT token ensured the protocol will start with a sufficient supply to incentivize nodes, while a buyback mechanism will ensure this going forward.

### Disclaimer

No guarantees will ever be given with respect to expected returns. Your investment can go to zero.

# A Network for Decentralized Uncensored AI Models



## enqAI

Project overview

Our model, built upon a vast array of foundational works, represents not so much a significant leap in language comprehension and generation capabilities, but more a removal of arbitrary restrictions that most models experienced during their development.

We start with a comprehensive list of resources that have influenced the development of our model. This collection encompasses fundamental techniques, seminal theories, and innovative practices that have shaped the current landscape of LLMs. It is important to note, however, that our journey through these resources was not solely to adopt methodologies but also to critically analyze and understand approaches that we chose not to follow.

Particularly, we paid close attention to works concerning AI safety and bias removal. Our approach has been to consciously steer away from methods that do not align with our model's design philosophy and objectives.

### • Transformers

- *Key Paper:* Attention Is All You Need, Vaswani et al., 2017
- *Description:* Introduced the transformer architecture, enhancing NLP capabilities.

### • BERT

- *Key Paper:* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al., 2018
- *Description:* Employs bidirectional training of transformers for improved language understanding.

### • GPT

- *Key Paper:* Improving Language Understanding by Generative Pre-Training, Radford et al., 2018
- *Description:* Generative Pre-trained Transformer models, setting new standards in language models.

### • Large-Scale Training

- *Key Papers:* Language Models are Few-Shot Learners, Brown et al., 2020; Efficient Training of Large Neural Networks, Lepikhin et al., 2020
- *Description:* Using massive datasets and computational resources for training.

### • Regularization and Optimization

- *Key Papers:* Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Srivastava et al., 2014; Adam: A Method for Stochastic Optimization, Kingma and Ba, 2014
- *Description:* Critical techniques for training deep learning models effectively.

### • Transfer Learning

- *Key Paper:* How Transferable are Features in Deep Neural Networks?, Yosinski et al., 2014
- *Description:* Adapting a model trained on one task to another task.

### • Ethical Considerations

- *Key Papers:* Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms, Noble, 2018; Fairness and Abstraction in Sociotechnical Systems, Selbst et al., 2019
- *Description:* Focus on ethical implications of AI, fairness, transparency, and bias mitigation.

### • Data Preprocessing

- *Key Paper:* Data Preprocessing Techniques for Data Mining, Han et al., 2006
- *Description:* Effective preprocessing and cleaning of training data.

### • Natural Language Understanding

- *Key Paper:* Neural Models for Information Retrieval, Mitra and Craswell, 2017
- *Description:* Understanding context, sentiment, and nuances in language.

### • Natural Language Generation

- *Key Papers:* Sequence to Sequence Learning with Neural Networks, Sutskever et al., 2014; Neural Text Generation: A Practical Guide, Liu et al., 2018
- *Description:* Generating coherent and contextually appropriate responses.

### • Scalability and Distributed Computing

- *Key Papers:* Large Scale Distributed Deep Networks, Dean et al., 2012; Distributed Representations of Words and Phrases and their Compositionality, Mikolov et al., 2013
- *Description:* Using distributed computing resources for training large models.

### • Bias Mitigation Techniques

- *Key Papers:* Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, Bolukbasi et al., 2016; Fairness and Abstraction in Sociotechnical Systems, Selbst et al., 2019
- *Description:* Detecting and reducing biases in models for fairness and reliability.

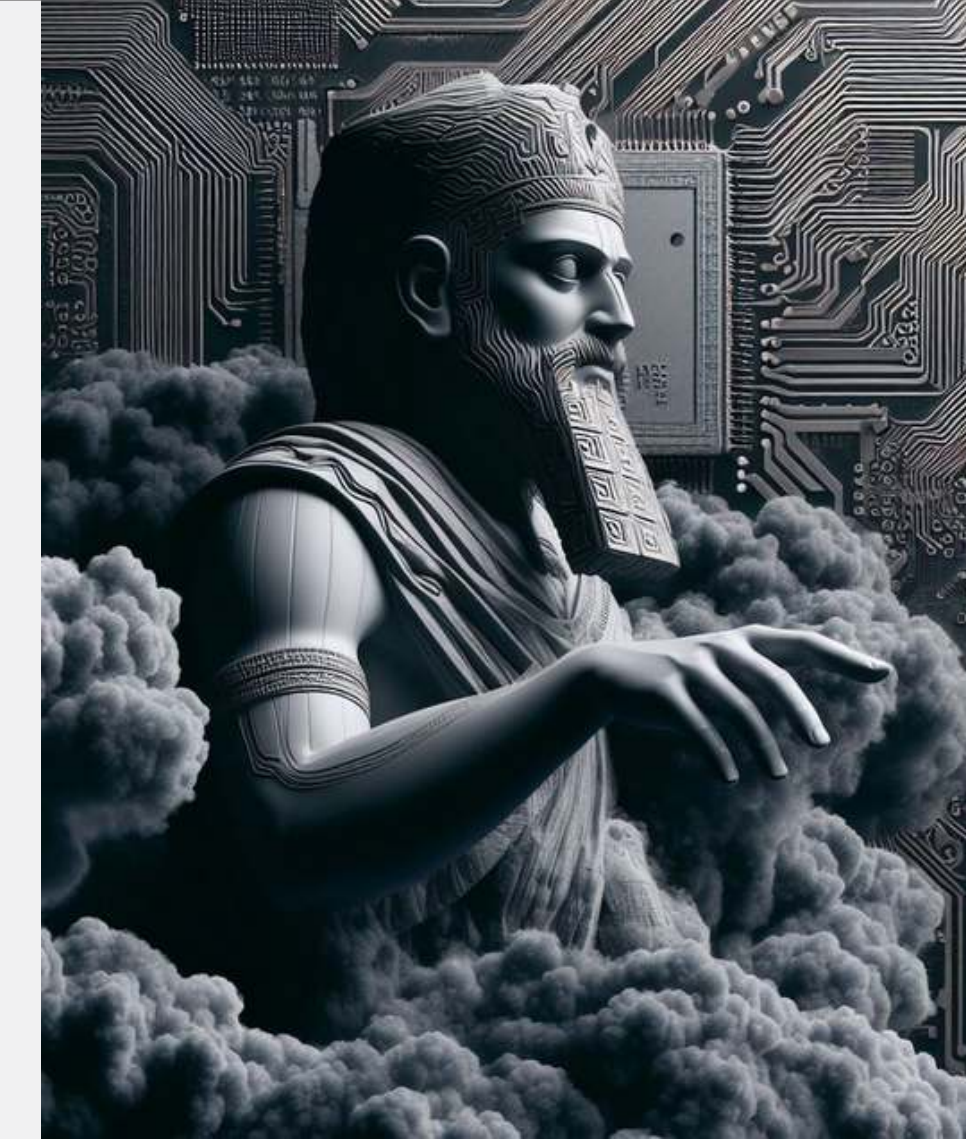
### • Llama 2

- *Key Papers:* Llama 2: Open Foundation and Fine-Tuned Chat Models, Touvron et al 2023;
- *Description:* Our favorite paper detailing Llama 2.

# A Network for Decentralized Uncensored AI Models

## enqAI

Project overview



For our LLM, which is yet to be named and released, we used an approach very similar to the one mentioned in Touvron. We briefly share some of our approaches and considerations below.

**Pretraining:** Implement an auto-regressive transformer architecture, an advanced model for processing sequential data, following the principles outlined in Touvron et al. (2023) and Vaswani et al. (2017) for transformer architecture. Enhance data cleaning and update data mixes, aiming for a diverse and representative dataset. Training should focus on a broad mix of publicly available data sources. The model architecture should include pre-normalization using RMSNorm, SwiGLU activation functions, and rotary positional embeddings, as detailed by Zhang and Sennrich (2019) and Su et al. (2022). For optimization, use the AdamW optimizer, a variant of the traditional Adam optimizer that incorporates weight decay, as described in Loshchilov and Hutter (2017). The tokenizer should be based on byte pair encoding (BPE), a method for efficient text representation, as explained in Sennrich et al. (2016) and Kudo and Richardson (2018).

**Scalability and Distributed Training** Managing computational resources effectively is crucial. The model's extensive data requirements and complex architecture often necessitate distributed training, a method where the training process is divided across multiple GPUs or even across different machines. This approach not only speeds up the training process but also allows for handling larger models and datasets that would be impossible on a single machine. Techniques like gradient accumulation, mixed-precision training, and efficient data loading are often employed to optimize the use of resources. Note that even though this process could also lend itself for decentralization, for efficiency reasons we focus on decentralizing the inferences and not so much the training. This is obviously an inevitable addition in the future.

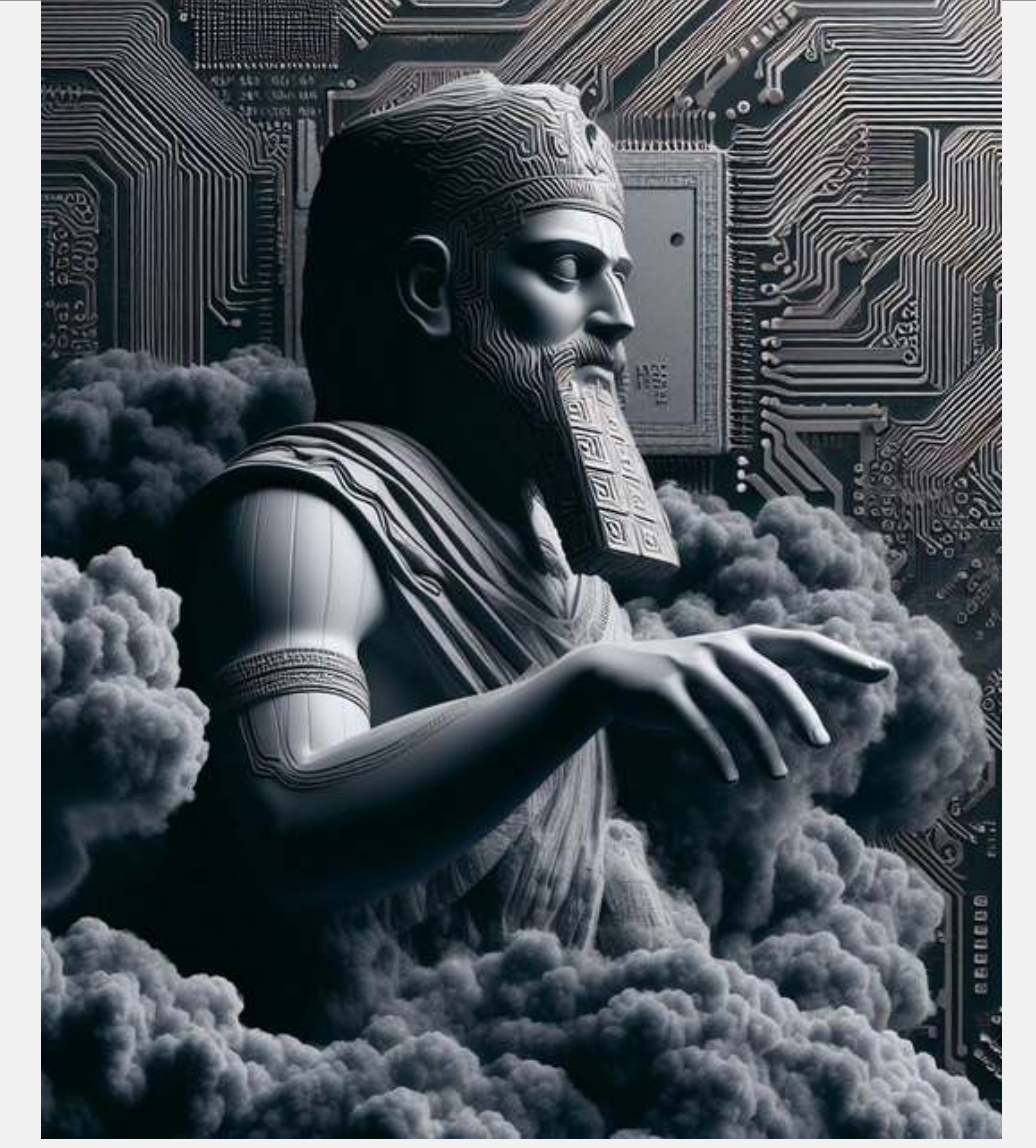
**Regularization Techniques** Regularization is key in preventing overfitting, especially in a model trained on a vast dataset. Techniques like dropout, where randomly selected neurons are ignored during training, help the model to generalize better. Layer normalization, especially in transformer models, is another common technique that stabilizes the learning process. Additionally, data augmentation, where the training data is artificially expanded by creating modified versions of existing data points, can also be employed to improve model robustness.

**Evaluation and Benchmarking** Benchmarking is essential to evaluate the performance of language models. We used some of the standard benchmarks, like: GLUE (General Language Understanding Evaluation): A collection of diverse natural language understanding tasks. SuperGLUE: Designed to be a more challenging set of tasks than GLUE. SQuAD (Stanford Question Answering Dataset): A reading comprehension dataset consisting of questions posed on a set of Wikipedia articles. LAMBADA: Tests the model's ability to predict the final word of a text passage. Commonsense reasoning tasks: Such as the Winograd Schema Challenge, testing the model's understanding of everyday concepts. These benchmarks measure various aspects of language understanding, such as sentence completion, question answering, and text classification, providing a comprehensive view of the model's capabilities. For truthfulness we've created a proprietary benchmark which we will publish as well.

**Fine-tuning:** Initiate with publicly available instruction tuning data. This phase involves adjusting the pretrained model to perform specific tasks or understand certain types of prompts better. Focus on collecting high-quality Supervised Fine-Tuning (SFT) data. The SFT process is critical for aligning the model with desired outputs, as emphasized in the work of Chung et al. (2022) and Zhou et al. (2023). Utilize a cosine learning rate schedule during fine-tuning, a strategy for adjusting the learning rate of an optimizer, as detailed in the literature on neural network training. For the actual fine-tuning, use an autoregressive objective and ensure the model sequence length is properly filled, as demonstrated in various studies on language model training. In our approach, safety-specific steps like safety fine-tuning and reinforcement learning with human feedback focused on safety are excluded. These steps are often included to mitigate risks associated with language model outputs, such as generating harmful or biased content.

**Continuous Learning and Model Updates** For a language model to remain relevant and effective, it needs to adapt to the ever-evolving nature of language and information. This is where continuous learning and model updates become important. Continuous learning involves periodically retraining the model with new data, ensuring that it stays up-to-date with the latest language use and information. This can be challenging, as it requires balancing the need for new information with the risk of forgetting previously learned material (a problem known as catastrophic forgetting). Techniques like replay (retraining with a mix of old and new data) and regularization methods designed to preserve previous knowledge can be helpful. Additionally, monitoring model performance and biases over time is crucial to identify when updates are needed.

# A Network for Decentralized Uncensored AI Models



enqAI

Project overview

## noiseGPT, a lifelike text-to-speech engine

The release of the noiseGPT core-model was significantly influenced by ElevenLabs' decision to censor their product, driving a demand for an alternative, unrestricted approach. Based on extensive publicly available research, noiseGPT has been operational since February 2023 and has already generated over 100,000 inferences. It is accessible through various integrations, offering a wide range of applications in lifelike text-to-speech (TTS) technology. These applications include voice synthesis for virtual assistants, audiobook narration, voiceovers in multiple languages and accents, and personalized speech patterns for accessibility solutions. For this model, we owe a great deal to the following research efforts:

- **Digital Processing of Speech Signals**

- *Key Paper:* Rabiner, L., & Schafer, R. (1985). Digital Processing of Speech Signals. Prentice-Hall.

- **Text-to-Speech Synthesis**

- *Key Paper:* Taylor, P. (2009). Text-to-Speech Synthesis. Cambridge University Press.

- **Progress in Speech Synthesis**

- *Key Paper:* van Santen, J. P., Sproat, R. W., Olive, J. P., & Hirschberg, J. (1997). Progress in Speech Synthesis. Springer Science & Business Media.

- **Cascade/Parallel Formant Synthesizer**

- *Key Paper:* Klatt, D. H. (1980). Software for a Cascade/Parallel Formant Synthesizer. Journal of the Acoustical Society of America, 67(3), 971-995.

- **Code-Excited Linear Prediction**

- *Key Paper:* Schroeder, M. R., & Atal, B. S. (1985). Code-excited linear prediction (CELP): High-quality speech at very low bit rates. ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing.

- **Deep Neural Networks in Speech Recognition**

- *Key Paper:* Hinton, G., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, 29(6), 82-97.

- **Neural Probabilistic Language Model**

- *Key Paper:* Bengio, Y., et al. (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3, 1137-1155.

- **Sequence to Sequence Learning**

- *Key Paper:* Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems, 27, 3104-3112.

- **Frame-wise Phoneme Classification**

- *Key Paper:* Graves, A., & Schmidhuber, J. (2005). Frame-wise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. Neural Networks, 18(5-6), 602-610.

- **Learning Phrase Representations**

- *Key Paper:* Cho, K., et al. (2014). Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724-1734.

- **Neural Machine Translation**

- *Key Paper:* Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.

- **Transformers**

- *Key Paper:* Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30, 5998-6008.

- **BERT**

- *Key Paper:* Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

- **XLNet**

- *Key Paper:* Yang, Z., et al. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237.

- **RoBERTa**

- *Key Paper:* Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

- **Natural TTS Synthesis**

- *Key Paper:* Shen, J., et al. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

- **WaveNet-Based Speech Synthesis**

- *Key Paper:* Vasquez, A., et al. (2020). WaveNet-based Speech Synthesis with Noise Shaping. arXiv preprint arXiv:2001.11478.