LIAISE: Lovetillion Inalienable Artificial Intelligence Serving Everyone

LIAISE Developers
support@lovetillion.org
www.lovetillion.org

Abstract

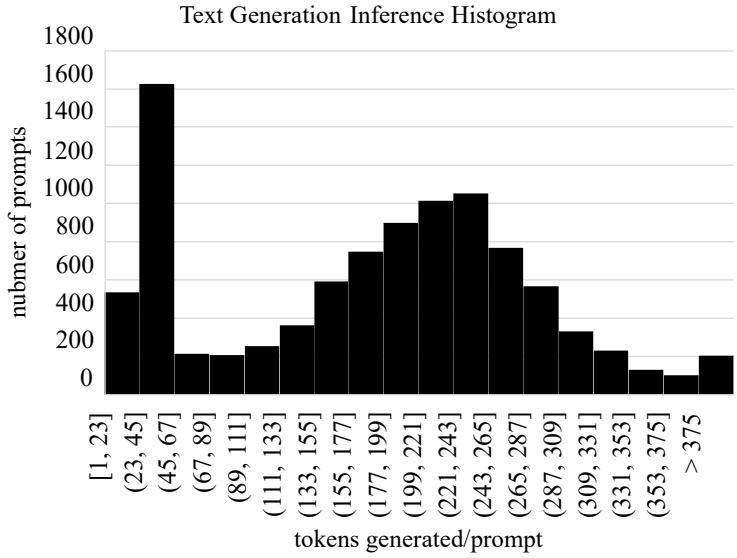
Describes a distributed system for managing peer-to-peer ethical and rights protecting artificial intelligence spanning several key domains: integration with third-party blockchains for rewarding actors in the system; certification of adapters ranging from tiny client-side to large server-side models facilitating complex ensemble workflows; efficient sharing of system prompts, retrieval augmentation and training data sets with transparency and traceability; and orchestration of prompt, inference and central aggregation of tiny non-latency sensitive training workloads. Instead of proof-of-work style cryptographic hashes, the system relies upon a large number of end-users who are seeking rewards that initiate valuable proof-of-work for its security. Any actor with a supported blockchain public key address/account can interact directly with the LIAISE system without the need for any centralized entity or intermediary. The provisioning of LIAISE resources are instantaneous and permissionless. To navigate the complex challenges around operating an ethical artificial intelligence system, with humility we will be following the teachings of our Lord and Saviour, Jesus Christ. We have drawn inspiration and attempted aligning the design of the system to existing work done as part of Holy See sponsored initiatives.[1]

1. Primitives Workflow

Early prototyping was focused on forming baselines using truly open and easily accessible modern Transformer and Diffusion artificial neural network (ANN) models. We will explore some of that discovery here but there is no inherent limitation to the type of inference or tiny model training workloads we envision running through the system. The design of the system will provide extensibility to inference generative video, image-to-image, image text-to-text and diverse image classification tasks (object detection/image description), signal processing (denoise voice), automatic speech recognition, natural speech synthesis, deep learning recommendation models (DLRM) and to train tiny models providing at least a path to becoming less reliant on costly latency sensitive large models through novel ensemble approaches. Inference accuracy was confirmed based on the model, configuration, seed, point in time retrial augmented generation (RAG) contents and global system prompt in order to qualify for delivery to an end-user completing the reward eligible transaction. Below are some observations and findings from the research work and represent the initial set of primitives:

1 https://www.romecall.org/

1) Text Generation (Meta-Llama-3-8B-Instruct): In order to qualify for end-user proof-of-work, the text token generation lengths per prompt must fit within a normal distribution to disincentivize long response attacks. During initial testing of the system this normal distribution was demonstrated across the user base.



Inference does not qualify for rewards until any given end-user's text-generation token lengths $x_1, ..., x_n$ satisfy a Shapiro-Wilk test[2] where the p value is greater than the chosen α level.

2) RAG Context (chromadb all-MiniLM-L6-v2 embeddings): For instruction tuned chat bot style large language models (LLMs) with heavy parameter optimization and high costs to retrain, there will always be gaps in their knowledge set due to limitations of their training data corpus lacking access to highly esoteric vertical topics and new out of sample information (news events, tabular values, discoveries).

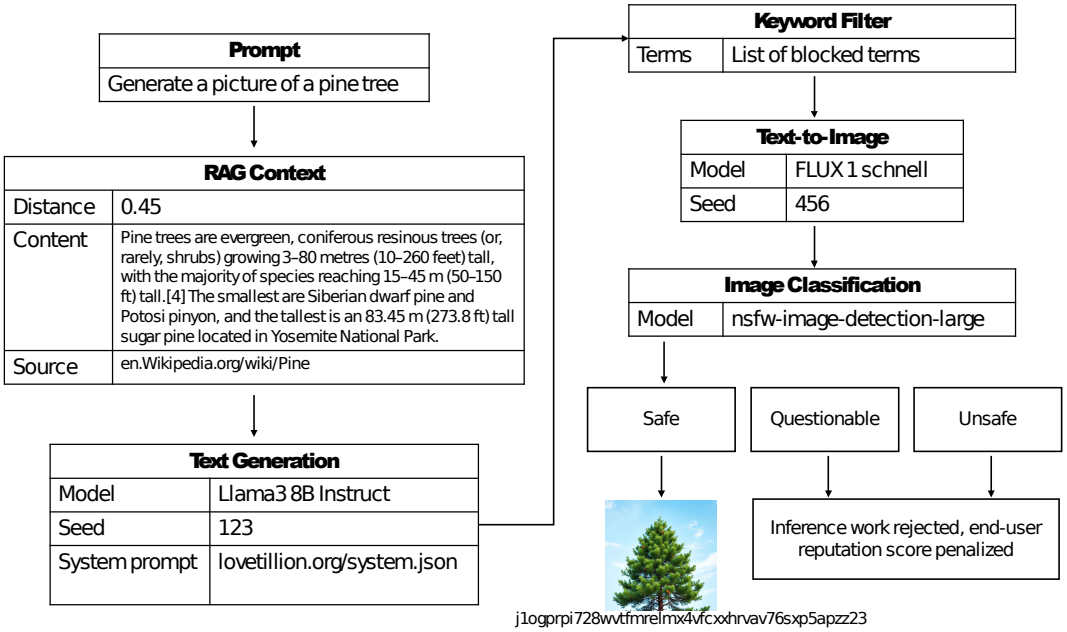
A scalable approach is needed in order to provide insightful responses and fend off against the posterior predictions containing excessive amounts of hallucination or bias from these gaps or other errors in the training data. Working with smaller models that are more reliant on retrieval augmentation can actually provide value when certain contents are found incorrect or inappropriate and need to be removed entirely. In terms of uses beyond text generation, we hypothesize reference images can be utilized to guide diffusion processes in image generation tasks through Image-to-Image for example to allow a deficient trained model to guide its generation to a poisonous mushroom reference photo.[3]

3) Text-to-Image (FLUX.1-schnell): During experimentation approximately 95% of the compute was allocated to the image generation task by end-users. In order to protect against malicious use several techniques were implemented including; passing the text-to-image prompt into a workflow with an ethical filter on the text generation model's system prompt, the resultant image generation prompt then had to then pass through a blocklist of

2 https://math.mit.edu/~rmd/465/shapiro.pdf
3 https://www.404media.co/google-serves-ai-generated-images-of-mushrooms-putting-foragers-at-risk/

filtered terms translated into common contrastive language-image pre-training (CLIP) languages, and then upon generation the final result was run through classifier step before finalizing the transaction.

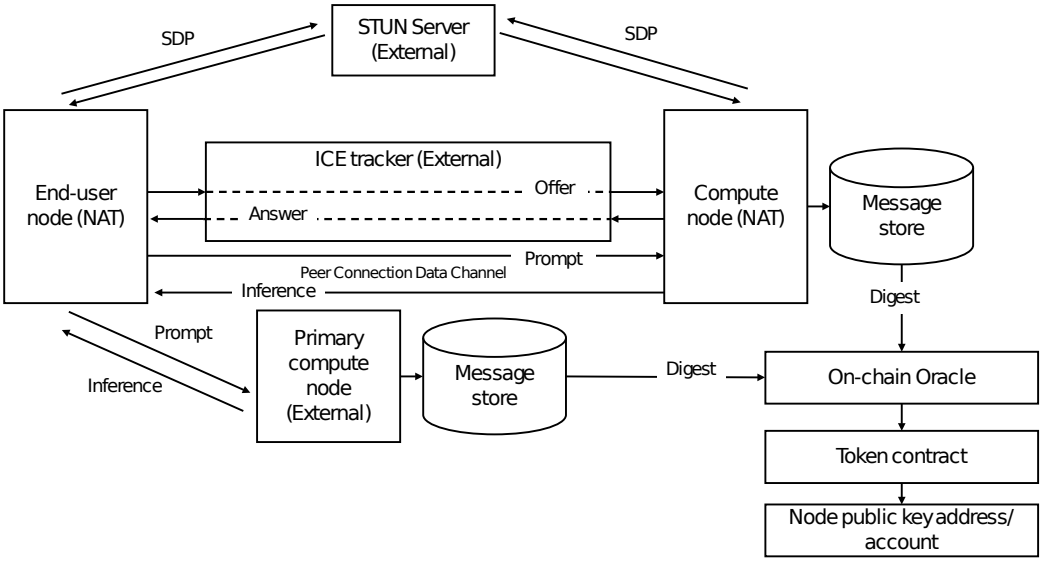
4) Image Classification (nsfw-image-detection-large): Based on a study of 17,001 images generated by users there were 1,049 not safe for work (NSFW) images detected with incidence rate of 6.17%. Since this would make text-to-image and supporting inference workflow considered wasted work, end-users are penalized for these generations to disincentive the behaviour by holding back a portion of their future rewards.
5) Inference Feedback: Primitives allow metadata hierarchies to be associated with them. In some cases this data will be assigned by the prompting user to provide evaluation feedback on the inference result. Applications to fine tune reinforcement learning (RL) weights and for RAG context guidance will be based on this type of user feedback through “what did you think of this response?” “what did you think about this image?” 👍 👎 style feedback mechanisms.



As primitives and their corresponding models are certified, compute nodes will be able to activate official adapters released by the DAO. These are seamlessly integrated into system workflows so compute nodes with varying capabilities can serve different functions. In many cases compute nodes will need to agree to share their contact information in order to access models per specialized licensing agreements. There may also be restrictions in place depending on where the compute is physically located due to local regulatory compliance needs.

2. System Nodes

Nodes represent the connection points between actors of the system. In order to form a peer-to-peer mesh, they need to maintain and share a list of known external primary node Internet protocol (IP) address and/or exchange interactive connectivity establishment (ICE) session description protocol (SDP) messages providing offer-answer handshakes for network address translation (NAT). Nodes can operate in a lightweight capacity where they only send prompts or respond with their inference workloads and do not persist any of the historical proof-of-work data.



1) End-user: Initiate connections through a web browser to run prompts or train tiny models. Web real time communication (WebRTC) are supported through ICE trackers or end-users may connect to primary nodes. In addition to submitting prompts they will traverse workflows and share node connection information other peers.
2) Compute: Perform inference and training on consumer grade GPUs. As work is completed it's broadcast across the system's network of nodes and stored in the message store.
3) ICE tracker: Operate a public web service allowing end-user and compute candidate nodes to be discovered.
4) Message store: maintain a full record of distinct timestamped messages across the system for finalizing reward emissions based on observed proof-of-work ledger. To limit the size of data in text-generation with long conversation context, a parent child prompt inference hierarchy will exist. Users can have several conversations going on each with a different starting parent. For image data the average size was 1.3MB at 1024x1024 resolution so as an optimization a SHA-256 hash can be used for tracking the completed work making the source file itself optional.

   Since the storage of all system messages is tied to rewards there's an incentive for compute node operators to maintain a snapshot of this data. Message store nodes may also provide services to retrieve or share conversational history for use in RL. If an end-user ever wishes to restore their conversational history it could be retrieved by specifying their initial parent prompt conversation identifier. Images and originating prompts could be used to build royalty free image gallery published on the open web.

3. Privacy

User prompts, inference results and public training datasets, can be viewed by any third party interested in accessing them as they are part of a public proof-of-work ledger. It's expected that the system will be used for a wide range of purposes, including aiding in personal and professional communications, which means that they risk containing a variety of different types of personal data. Official DAO guidance is to provide a prominent warning to users that communications are not private or confidential and they are not intended to contain any sensitive or personal data. The system has no obligation or ability to protect against these risks and will stress sensitivities around utilizing any personal data in prompts since they will not remain confidential. Communications

will be sent and received with encryption while in transport wherever possible but since they are stored on a public proof-of-work ledger it eventually leaves them fully accessible in the public domain in perpetuity.

Similar to e-mail communications which can be sent in plain text internationally the LIAISE system does not have confidentiality obligations under privacy regulations. Precedence around the exceptional nature of common and ubiquitous forms of public over the Internet communications exists and can be freely used for a wide range of purposes. Thus, privacy restrictions are considered outside of the scope of the foundational architecture of LIAISE. This position will be continually revisited as advancements are made in fully homomorphic encryption (FHE), trusted execution environment (TEE) and similar privacy protecting technology to improve end-user privacy in the system. From a model training perspective as demonstrated by Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation (COINSTAC)[4] contributing to some level of federated learning with privacy protections may be possible. In this example, only aggregate model checkpoints are shared across the network so these central aggregators and the final proof-of-work ledger wouldn't require knowledge of the individual originating training data inputs.

From a networking standpoint, since nodes make connections directly to ensure a pure peer-to-peer solution there will be exposure to IP addresses on the public Internet. Connections will be possible behind NAT but not for connections behind common privacy oriented virtual private network (VPN) providers since they will prevent a successful session traversal utilities for NAT (STUN) handshake due to privacy protections they enforce at the provider level. It's also worth noting that WebRTC support is entirely absent from the Tor browser bundle's build for the same privacy reasons. To get around some of these restrictions end-users will be reliant on primary nodes. In future the DAO plans to setup traversal using relays around NAT (TURN) servers which will help increase adoption while maximizing user privacy allowing for the use of a VPN. The drawback is of course the need for these TURN nodes to relay traffic which will result in impacts to the system as it continues to evolve which is beyond the scope of what is defined here. Outside of the above cases the messages distribute through the layers of the peer-to-peer network based on the node public key address/account without persisting IP addresses of the originating node.

4. Emissions

When interacting with blockchains, the system requires its own utility token XLV representing the lovetillion decentralized autonomous organization (DAO) which will follow a linear emission schedule. This allows for balanced participation from late entrants to prevent centralized control of the DAO. Tokens can be generated by any end-user or compute node, or acquired through a decentralized exchange (DEX) liquidity pool (LP) in a fully permissionless decentralized manner.

Whichever nodes complete the inference first qualify for the bulk of rewards, slower nodes can provide the necessary validation of the completed work. 90% allocation of rewards goes to the lowest latency response, 7% is shared by the validators and 3% goes to end-users.

The DAO has the option to choose an emissions blockchain every 180 days both as part of the missionary program of evangelizing the system to new communities and increasing the accessibility of the system to more users. There is no requirement to change blockchains continuously but it's seen as highly desirable to support the missionary aspect of the project. During the proof of concept phase two relatively low transaction fee blockchains that provided support for end-user authentication, tokens and smart contracts were utilized:

- Bitcoin Cash as a UTXO (unspent transaction output) model using Wallet Connect.
- Arbitrum as a state-based model using Multi Injected Provider Discovery (EIP-6963).

4 https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00365/full

End-users and compute nodes sign all completed useful work (prompts, inference, training checkpoints) for consideration as part of the next emission. They simply must be authenticated utilizing the currently supported emissions blockchain. A similar mechanism can be used for XLV spends where a given primitive (ex: video) requires an end-user to pay a fee to utilize the model since it will exceed their maximum rewards. This would cause compute node to focus their resources on a specialized request as long as it carried a premium over their expected portion of daily rewards from other primitives available. This mechanism allows compute nodes to accumulate XLV in excess of the daily emission cap.

If a compute node is observed to have reliability below 97%, the network implements a quality of service (QoS) optimization to shift prompts away from affected nodes to avoid increasing the latency of the system thereby excluding them from emissions until they can demonstrate the necessary availability while receiving a significantly reduced portion of randomly assigned work.

5. Security

Compute nodes will be given work in a provably random assignment to avoid end-user, inference and validator collusion. This will also protect against defeating NSFW classifier tasks risking end-users being exposed to inappropriate content. Within the proof-of-work ledger, the timestamped messages form hierarchies associated with the public key address/account of the cooperating nodes. These observations form a directed acyclic graph (DAG) of authenticated node messages with their prompts and inference or model checkpoints. These results are summarized then broadcasted to on-chain oracles for use by smart contracts to determine how to reward nodes.

Continuous reconciliation of the timestamped public key signed observable work passes through the system providing a cumulative source of truth for reward balances based on shared observations from all message store nodes so they can be aligned in the discrete time series. This creates a shared hierarchy of sets where observation time overlay and expected rewards are in agreement to satisfy emission. All transaction settlement of the XLV utility token are handled by blockchains through their corresponding proof-of-work or proof-of-stake mechanism. This eliminates the need for handling any double-spends, tracking of balances or facilitating transfers drastically simplifying the design of the system.

It's expected that the system will be subjected to dishonest end-users sending spam prompt floods similar to what the early Bitcoin system was subjected to.[5] This could also apply to compute nodes that are configured incorrectly or returning inference that differs from what is expected based on the system primitive in a given workflow. Since the network operates as a pure peer-to-peer mesh and the primary node or NAT firewall's external IP addresses will always be visible, it will be at the discretion of nodes to reject traffic similar to how Internet relay chat (IRC) servers had support for a kill line (K-Line). Further, the DAO reserves the right to publish a global (G-Line) to globally ban a network.

Sybil attack prevention is implemented through a continuously reevaluated optimization function that requires end-users to accumulate a history of honest behavior that meets the expected use frequency and quality standards of valuable proof-of-work before they can unlock a larger portion of rewards. Since all end-users share in the same reward pool, an attack of flooding the network with requests reduces the overall throughput of the constrained compute node resources which will self-normalize. Between these mechanisms and other more brute force spam prevention techniques we believe the system will be able to continue to scale fairly.

5 https://en.bitcoin.it/wiki/Spam_transactions

6. Conclusion

The LIAISE design presented attempts to unlock the possibilities of an unstoppable, fully decentralized, peer-to-peer operating system for humanity's foray into artificial intelligence. Techniques have been presented to ensure the usage remains ethical and protects the rights of all actors. By leveraging blockchain technology and its own distributed network of nodes, LIAISE enables a wide range of applications ensuring that no person will be left behind in this new area of computing. A technology that progresses development in a responsible and harmonious fashion with principles anchored in the teachings of Jesus Christ and the deep introspection of our collective moral compass.

Through the DAO enforced system prompts, model workflows and certification processes we will be able to ensure that inference quality meets the highest standards, that it can continually improve and effortlessly scale without requiring commercial cloud or highly centralized datacenters. Its modular architecture and open-source nature will enable developers and researchers to contribute to the system and create new applications and models with a basis of rewards allowing for useful proof-of-work and shared participation in the growth and direction of the system.