INFERIUM - THE PIONEER ML-DRIVEN INTELLIGENT STORE AND AGGREGATOR FOR AI INFERENCE.

Inference is the process of running live data through a trained AI model to make a prediction or solve a task.

PROBLEM

There are some key problems related to the current lack of an AI inference store and aggregator:

1. Complexity in AI model selection
2. Lack of communication between end-users and developers
3. Lack of awareness of new models
4. Lack of privacy and compliance protection during AI inference tasks

DEMAND

AI technology demand can be divided into two main categories:

1. Tailored demand: the lists of specific groups actively seeking AI solutions include:
- AI companies
- AI developers
- Private enterprises
- End users

2. Sector demand: the demand is not limited to generic AI models, this category underscores the importance of niche models tailored for various industries:
- Trading
- E-commerce
- Healthcare
- Medicine
- Customer Service
- Marketing
- Education

SOLUTION

These problems and demands highlight the critical need for a unified platform that could streamline model selection and improve usability in AI applications. And Inferium is the pioneer ML-driven intelligent store and aggregator for AI inference.

OVERALL TECH ARCHITECTURE

The overall tech architecture in Inferium presents a detailed view of a system designed for AI model creation, deployment, and user interaction. Inferium integrates key components like AI companies, AI developers, Machine Learning Engine (ML Engine), Users, Inferium, GPU Computing Layer, and Data Layers to streamline AI model development, validation, and deployment. As a central hub, it links AI stakeholders—companies, developers, clients, and users—to meet varied AI demands. Its data layers and advanced computational capabilities ensure real-time performance and scalability, providing an all-in-one solution for today's AI challenges.

The flow of tech architecture in Inferium:

- From AI companies and AI developers to Inferium: AI models are developed and submitted to Inferium for management and deployment.
- From Inferium to users: Users access Inferium to find and select the most appropriate AI model based on their requirements.
- Data gathering: The data layers collect datasets necessary for the AI models to perform inference, which is then utilized by Inferium.
- Processing with ML engine: The ML engine evaluates the performance of various AI models, learning which models best meet user needs and improving accessibility.
- Computational power: The GPU computing layer provides the necessary resources to execute models, ensuring fast and efficient computations.

CUTTING-EDGE FEATURES

The innovative functionalities of Inferium focus on emphasizing 3 core principles: Verify, Evaluate, and Aggregate.

1. Automated Inference Verification System: Streamlined deployment process in minutes for developers through our automated testing framework to validate for any code errors or security issues.
2. Inference Performance Measurement: Daily automated testing of listed AI models with Inferium's unique metrics to ensure reliability and robustness.
3. ML-Driven AI Model Aggregator: Compare multiple AI models simultaneously to easily identify the most effective model through direct performance assessments, user needs, and interaction patterns.

4. Blockchain Pre-Integration: Pre-integrate with smart contracts on multiple blockchain platforms allows direct plug-in for AI models, enhancing interoperability and reducing setup time for developers
5. Inference Mastery: Introduce a competition-based gamification of inference models on specific problem sets, with the best models being rewarded each season.
6. End User and Developer Incentives: Promotes community engagement and model quality through rewarding mechanisms. Users who provide testing and feedback are incentivized, while developers are rewarded for high-performing models, creating a dynamic and innovative ecosystem.
7. Enhanced Privacy Protection: Integrate homomorphic encryption technology (FHE) that allows users to run AI models on sensitive data while maintaining privacy and security.
These innovations make Inferium a powerful platform for both developers and end-users in the AI landscape.

INFERENCE PERFORMANCE MEASUREMENT AND AGGREGATOR

There is a comprehensive framework for measuring AI model performance through three main components: adaptive metrics, human evaluation, and cross-model comparison.
- Adaptive metrics: Use industry-standard metrics tailored to each type of AI model. For example, use precision, recall, and F1-score for classification models, and mean squared error (MSE) or mean absolute error (MAE) for regression models.
- Human evaluation: Integrate user feedback directly into performance metrics, allowing models to be rated not just on technical performance but also on user satisfaction and practical utility.
- Cross-model comparison: Use multiple top benchmarks to compare performance across different models, highlighting strengths and weaknesses.

PROOF OF INFERENCE

The Proof of Inference categorizes features associated with three different validation methods: TEES (Trusted Execution Environments), ZKPROOF (Zero-Knowledge Proofs), and Optimistic approaches. The image lists several features that are evaluated across the three methods:

Workflow for Proof of Inference:
- Submit Inference Task: The user submits input data to the Inferium platform.

- Run AI Model: ROFL (Runtime Off-chain Logic) executes the AI model on the input data and generates results.
- Generate Cryptographic Proof: ROFL creates a cryptographic proof of the computation, using a TEE or zero-knowledge proof.
- Secure Transmission: ROFL securely sends the proof and results to RONL via RPC.
- Verification: RONL verifies the proof against known criteria and attestations.
- Blockchain Recording: If valid, RONL records the inference results and proof on the blockchain for transparency and trust.

ADAPTIVE METRICS

The adaptive metrics feature a comparative table detailing the performance of different AI models across six evaluation metrics, alongside their average scores. Each model's performance is quantified as a decimal, representing its effectiveness in various assessments like MMLU, HellaSwag, HumanEval, BBHard, GSM-8K, and MATH.

HUMAN EVALUATION

The human evaluation emphasizes the importance of qualitative assessments in AI model performance. Through a systematic scoring approach and user feedback, it aims to balance technical metrics with human judgment, ultimately enhancing the evaluation process for AI models. This ensures that the models not only perform well according to metrics but are also favored by end-users for their practical effectiveness.

1. Model Score:
- Total Score Formula: The total model score is calculated using the formula:

Total Score = $\alpha \times MS + \beta \times US$

α (alpha) and β (beta) are weight factors used to balance the contributions of the metric score (MS) and the user score (US).

$\alpha = 0.4$: Weight assigned to the Metric Score (40%).

$\beta = 0.6$: Weight assigned to the User Score (60%).

- Scoring Breakdown:
  + MS (Metric Score): This score is based on various metrics applied to the model's performance. It reflects the average score across those metrics.
  + US (User Score): This score measures user interaction with the model. Each request or download contributes to calculating this score.

- Task Scoring: The image outlines various tasks and their corresponding points, showcasing how developers can earn points:
  + Download/Clone: 1 point
  + Deploy on Space: 1 point
  + Good Open Feedback: 1 point
  + Average Rating > 3: 3 points
  + Voting 1st/2nd/3rd Best on Comparison: 3–1 points based on rank
  + Crowdsourced Judging: 100 points for 1st place in Model PVP and 50 points for being in the Top 5 in Model PVP

All metrics are normalized to a scale of 0 to 100 for consistency.

2. Emphasizing Human Judgment:

- Interface Overview: Features an interface for Deploying and Comparing Models. The interface allows users to submit inputs and compare the outputs of different models.
- Process Steps:
  + Submit Input: Users can upload or browse to submit an input image or data for comparison.
  + Receive Output: The output from multiple models is displayed for comparison. Each model is anonymized with labels (e.g., Model A, Model B, Model C) to ensure unbiased feedback.
- Comparison & Ranking: Users can examine outputs side by side and rank the models based on their preferences. A prompt asks users, “Which model do you like best?”, where they can select their top three preferences (Top 1, Top 2, Top 3).

INFERENCE MASTERY: LEADERBOARD

The leaderboard is dynamically updated in real-time, providing immediate access to the most current scores and rankings of AI models within the marketplace per category. This cultivates a competitive atmosphere, incentivizing developers to iteratively enhance their models in pursuit of higher rankings.

The leaderboard table lists various AI models, and here's what each column represents:

- Ranking: The position of the model in the leaderboard based on its score.
- Name: The name of the AI model.
- Score: The performance score of the model, with higher scores indicating better performance.

- Model Type: The classification of the model, such as CNNs (Convolutional Neural Networks), RNNs (Recurrent Neural Networks), etc.
- Use Cases: The specific applications or tasks that the model can perform, like prediction, object detection, or content generation.
- Developer (Dev): The name of the developer or team who created the model.
- Users: The number of users utilizing the model, indicating its popularity and adoption within the platform.

INFERENCE MASTERY: MODEL PVP

The Model PVP presents an engaging and incentivizing environment for AI developers and judges alike. This promotes high-quality AI model development and validation by combining crowdsourced evaluation with structured judging criteria and recognition.

- Public model PVP: Open to all participants, this competition allows anyone to publicly submit and compete with their AI models for rewards and recognition.
- Enterprise model PVP: A tailored competition exclusively for enterprise-level models, offering companies a platform to showcase and benchmark their proprietary AI solutions in different niche sectors against their competitors. All results will be kept private and confidential for each enterprise.

Crowdsourced judging evaluation: Judges will consider relevant criteria based on each model. There is an on-chain proof-of-judging mechanism in place to ensure quality and integrity during the judging process. This includes backend checks for judges to maintain standards. Judges will receive recognition and rewards for their contributions, promoting engagement and encouraging participation in future evaluations.

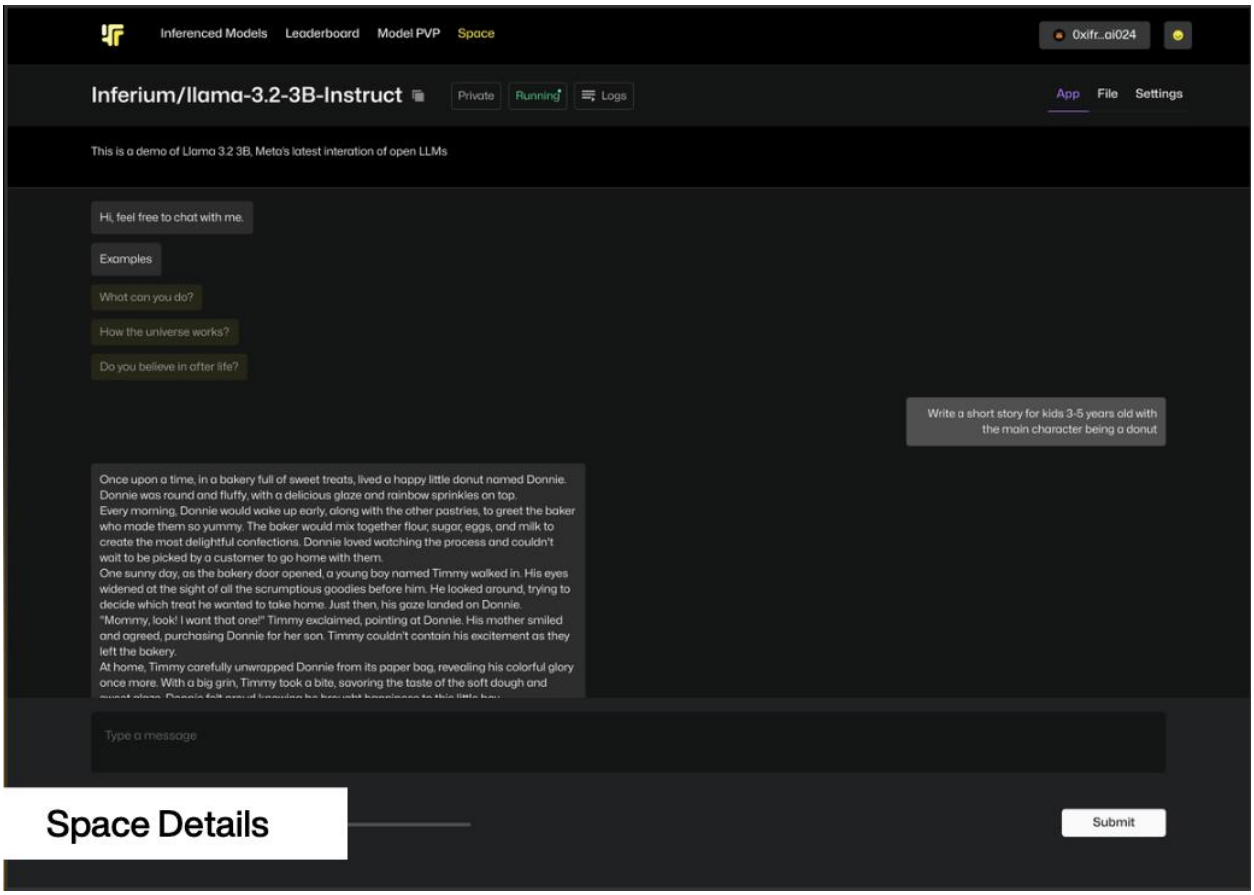
Judge eligibility criteria:

- Participated in previous tournaments and awarded a prize
- Having models that have more than 50K usages and avg. rate > 4
- Strong education background and experience (LinkedIn, Patent cross-check)
- Pass code challenges

GENERAL PROCESS

The process outlines a structured approach for bringing AI models from development to deployment, ensuring robust validation and engagement from both developers and users. The use of blockchain enhances transparency and trust in the system, making it easier to manage and verify the performance of AI models.

USE CASE



BUSINESS MODEL

Inferium's business model showcases how various elements interact to create value. It outlines a symbiotic relationship between advertisers, users, AI companies, and the Inferium platform, creating a dynamic ecosystem where value is generated through participation, services, and rewards.

TOKEN UTILITY

The $IFR token is central to the Inferium ecosystem, providing various utilities, including developer incentives, transaction fees, and premium services. This provides an overview of how Inferium encourages participation and creates incentives for both users and developers within its ecosystem.

- End user and developer incentives: Reward users for testing and feedback to boost model quality and community engagement. Simultaneously, incentivize

developers with rewards for high-performing models to promote a dynamic and innovative environment.
- Staking to become validators: Users must stake IFR tokens to become validators. In return, validators receive rewards for their contributions.
- Transaction fees for user space: Each user has a designated space limit for comparing and using models. Exceeding this limit will incur a transaction fee.
- Premium service fee: Premium services include tailored model customization for user deployment, pay-per-inference for developers, and 24/7 walkthrough and advisory.

USER SPACE PACKAGE

The user space packages for Inferium detail various tiers of service, their features, and pricing. Here's a simple package:

Free Tier:
- Hardware: Basic CPU (2 CPUs, 16 GB RAM)
- GPU: No GPU support
- Storage: 50 GB (free)
- Features:
  + Access to basic space for deploying models
  + Options for public or private spaces
  + Community support
- Price: Free

Developer Tier:
- Hardware: Upgraded CPU (8 CPUs, 32 GB RAM)
- GPU: No GPU support
- Storage: 100 GB
- Features:
  + Includes all Free Tier features
  + Enhanced performance for model deployment
  + Development mode with SSH/VS Code support
  + Priority support
- Price: Starts from $5 per month

Pro Tier:
- Hardware: Custom hardware
- GPU: Yes
- Storage: 500 GB

- Features:
  + All Developer Tier features
  + Access to higher compute resources
  + Early access to new features
  + Access to private datasets
  + Dedicated inference API
- Price: Starts from $9 per month

Enterprise Tier:
- Hardware: Custom hardware
- GPU: Yes
- Storage: Custom
- Features:
  + All Pro Tier features
  + Single Sign-On (SSO) and SAML support
  + Team account support
  + Support for audit logs and managed billing
  + Deployment on private infrastructure
  + Enhanced security and compliance features
  + Enterprise model PvP (player vs. player)
  + Enterprise model evaluation
- Price: Starts from $25 per user per month

PAY-PER INFERENCE

The Pay-Per-Inference pricing model for using AI inference services in Inferium details costs associated with processing different types of data. This model requires both a dataset and GPU resources for the inference process.

Types of Data and Pricing: The pricing structure for various data types is as follows:
- Text Data: $0.003 per 1,000 tokens
- Audio: $0.003 per minute
- Video: $0.10 per minute
- Image: $0.010 per image

External GPU Pricing: The pricing for using external GPUs is based on contracts with specific providers:

- Aethir: Pricing based on Inferium's exclusive contract.
- IO.net: Pricing based on Inferium's exclusive contract.
- Others: Pricing based on Inferium's exclusive contract.

MARKET STRATEGY

The strategy emphasizes building a Strong Market Foundation with a focus on Unique Selling Propositions (USPs), establishing partnerships, and fostering community engagement.

Key Metrics:
- 31K+ users
- 50K+ verified emails
- 4,500+ inference requests daily
- 1K+ models in review
- 100+ ambassadors

Strong market foundation: emphasizes the establishment of partnerships, a strong focus on unique selling propositions (USPs), and targeted outreach strategies for both developers and end users. It fosters engagement and collaboration within the AI community while ensuring the platform remains innovative and appealing to its users.
- Establish Partnerships with AI projects, VCs, Data Layers, GPU Layers
- Huge focus on USPs
- Product roadmap aligned with token emission for continuous innovation
- Developer outreach through targeted advertising, hackathons, and workshops
- End user onboarding through social campaigns

Pre-Launch activities:
- Figma followed by MVP: Develop a prototype (Minimum Viable Product) using Figma, which is a design tool for creating user interfaces.
- Social points and waitlist campaign: Include initiatives to incentivize early sign-ups and reward users for engaging with the brand on social media.
- Airdrop campaign: Distribution of tokens or incentives to early adopters to generate buzz and encourage participation.
- Aggressive developer model interest campaign: Target outreach to developers to stimulate interest in building and deploying models on the platform.

Post-Launch activities:
- Alpha version release: Launch the initial version of the platform for users to start interacting with the core features.
- Integration with data market and computation cloud: Collaborate with and integrate services from data markets and computing platforms to enhance functionality.
- Staking to be validators: Allow users to stake tokens and participate as validators in the platform's ecosystem, helping maintain network integrity.
- Building measuring framework combining ARC-c, ARC-e, BoolQ, HellaSwag, WinoGrande, etc.

AI COMMUNITY PLAN

Inferium aims to build an inclusive AI community by empowering students, developers, and enterprises. The goal is to encourage collaboration, learning, and innovation through various initiatives that provide opportunities to explore, engage, and run inference on AI models.

AI University (AIU) Program is a central part of the community plan, focusing on engaging students and providing them with practical experience in AI.
- Target: The AIU Program specifically targets students from top universities, focusing on fields like Economics, IT, and Engineering. This focus ensures that participants have relevant academic backgrounds that can contribute to the AI field.
- Purpose: The primary aim is to create a space where students can run inferences on AI models in their fields.
- Execution: To achieve its goals, the program will implement several key activities: monthly workshops, mini hackathons, and an ambassador program to grow engagement through student leaders.

SUPPORTERS

Inferium is grateful for the backing of an exceptional group of supporters who contribute their expertise and resources to empower our mission. Our supporters include:
- X Ventures
- Insignius Capital
- Hodl Ventures

- Connectico
- X21 Digital
- Tempest Ventures
- Messier M87
- Roger Lim: General Partner at NGC Ventures
- Igor Grim: Founder of Rivalz
- Lester Lim: Founder at X21 Digital
- 0x_ZHUANG: CMO at Scallop
- Deepak: Founder at Altcoin Alerts
- @defi_mochi
- @CryptoDaku_
- @charles48011843
- @Tufanoglu_
- @bachkhoabnb
- @danhtan68
- @MartinHo9999
- and more

PARTNERS

Inferium collaborates with a diverse range of innovative partners to drive innovation in AI and Web3 technologies, including:

- Aethir: Building Scalable Decentralised Cloud Infrastructure (DCI) for gaming and AI – Website | X
- io.net: The AI Compute-as-a-Currency, powered by $IO™ Internet of GPUs™ – Website | X
- Rivalz: The AI world abstraction layer, powered by a dual-chain infrastructure – Website | X
- Oasis: The privacy layer for Web3 and AI – Website | X
- Push: The communication layer of web3, powering decentralized notifications & chats for wallets – Website | X
- AgentLayer: A decentralized autonomous AI agent blockchain protocol – Website | X
- Mind Network: An FHE restaking layer for PoS and AI networks – Website | X
- Segmind: A cloud orchestration platform for Generative AI, offering flexible infrastructure options, and PixelFlow, a rapid prototyping and deployment app – Website | X

- GAIMIN: Making AI affordable, powering AI with the world's largest distributed computational network of gamers – Website | X
- Solidus AI Tech: The compute marketplace, Al marketplace and AITECH PAD, powering the future of AI with our HPC data center – Website | X
- Cluster Protocol: The decentralized AI infrastructure – Website | X
- DIN: The first modular AI-native data pre-processing layer – Website | X
- Glacier: The first data-centric blockchain to supercharge AI & DePIN at scale – Website | X
- NeuroMesh: The world's largest AI model trained by everyone, for everyone – Website | X
- Carv: Building the largest modular data layer for gaming & AI – Website | X
- Aggregata: Accelerate Safe SuperIntelligence with decentralized AI data – Website | X
- Humans: Building the AIverse – Website | X
- Private AI: Building $PGPT, AI data governance with privacy – Website | X
- LAIKA AI: The first Web3 AI SuperApp, powered by on-chain data – Website | X
- Datalayer: Blockchain protocol for personal AI & DePIN wearables – Website | X
- Tenzro: Enabling everyone, everywhere to contribute and participate in development of AI – Website | X
- and more

ROADMAP

Inferium's roadmap simplifies goals into structured phases, reflecting a clear vision for expanding AI inference services, technological advancements, and community engagement.

Q2 – Q4 2024:

- Inference Model Listing
- Serverless API
- Social Points System
- Review and Rate Model
- Human Evaluation Beta
- MVP Launch
- Magic Search Beta
- User Space Beta

- Scoring System Beta
- Public IDO (Initial DEX Offering)
- Token Launch
- B2B Clients Scaling and Onboarding

Q1 – Q2 2025:

- Tournaments
- Judge Validation
- Model Validation
- External GPU Integration
- Model Leaderboard
- Subscription Tier and Compute Unit Setup
- Payment Integration
- Dedicated API
- Beta Launch
- Testnet Proof of Inference
- Testnet Proof of Human Evaluation
- CLI (Command-Line Interface) Library

Q3 – Q4 2025:

- Mainnet Proof of Human Evaluation (include Judging)
- Mainnet Proof of Inference
- Official Launch
- Dataset Module
- Private Data
- Private Infrastructure
- Audit Logs
- SSO (Single Sign-On) Integration
- FHE (Fully Homomorphic Encryption) Integration
- 3rd Party Deploy Cloud Integration

2026 and Forward:

- Platform Refinement
- Advanced Compute Options
- Full Inferium Hub
- Developer Professional Mode
- Quantum Computation
- Accuracy Improvement for Human and Tech Evaluation

- Inferium Library

This roadmap simplifies Inferium's goals into structured phases, reflecting a clear vision for expanding AI inference services, technological advancements, and community engagement.

CONTACT

For more information, feel free to connect with Inferium via:

Website: https://www.inferium.io/

Twitter: @InferiumAI (https://x.com/inferiumAI)

Telegram community: @InferiumAI (https://t.me/inferiumAI)