Purified Intelligence: The New Economics of Inference

Yilong Li · Lei Yang
MEGA · November 2026

Abstract. An inference provider can offer the same model for much cheaper and still lose to a more established rival. If consumers of intelligence cannot verify the model they're served, or know how their data will be handled, more supply does not always mean more intelligence they can actually use. We propose *Purified Intelligence*: inference that users can source interchangeably from providers they might otherwise have no reason to trust. As more providers become viable, users gain options and suppliers compete for demand that was previously out of reach. This forms the basis of a more open inference market.

1. The Supply Paradox

The world is producing more intelligence than ever before. But producing intelligence and making it available to users are two different economic problems.

New models and independent inference providers are coming online rapidly as the price of comparable intelligence continues to fall [1]. Yet the growing inference supply is far from homogeneous. Providers differ greatly in performance, data policies and overall reliability. And that's before accounting for the variance in the models themselves.

A development studio seeking a particular powerful model comes across a cheap inference provider it knows nothing about. It cannot establish whether it will actually run the model it advertises, or if it will securely handle its application data.

The unknown provider may be perfectly capable, but there is not sufficient evidence to justify the risk. The team opts to pay more for the familiar service without knowing whether the cheaper alternative would have served it just as well.

Aggregators already compare providers, routing requests according to price, speed or data policy requirements [2]. These features do not, however, establish the truth of every claim made about a particular inference service [3, 4]. Inference can become more abundant without adding to the supply which users can practically use.

Access is also unevenly distributed across borders. A model or provider which can be easily accessed by one person may be entirely unavailable to someone else. Geography, regulations and commercial terms each play a role in determining access. Purification cannot override legal requirements but it could make more independent providers into viable alternatives for users around the world.

2. Making Inference Interchangeable

Purified Intelligence is inference that can be sourced interchangeably across a range of providers and, where the buyer's requirements allow, across models. For this to be feasible, the buyer must first be able to find sufficient evidence about each service.

In economics, fungibility describes how easily one unit of a particular good can be substituted for another. Intelligence is not naturally fungible; the usefulness of inference ultimately depends on the task, the model being served and the conditions under which it is delivered. Across different models, outputs do not necessarily have to be identical to be acceptable substitutes for a particular task.

Inference from the same model served by two different providers might be acceptable substitutes for a buyer who can establish adequate safeguards and minimum performance targets. Another buyer might need a particular latency or data policy, and ultimately reject a service that does not meet their requirements, regardless of the model.

In such a market, who the inference provider is matters progressively less as the service they offer matters progressively more. *Purification* is the engineering pursuit of that outcome.

3. Purification Process

We start with the uncertainties which make inference services unacceptable to the user. The most immediate

concerns are model authenticity, privacy and security. Each reflects a different reason for rejecting providers which otherwise may be capable.

3.1 Model Authenticity

Model authenticity poses a direct question: is the provider serving the model it advertises? Statistical evidence about the model being run can be gathered through testing and model fingerprinting, but such output-only tests [3] have limitations. Trusted execution environments (TEEs) and remote attestation can provide evidence of certain properties of the environment in which inference occurs, but this must be securely linked to the model itself to support a convincing claim about its identity. Research prototypes have also demonstrated cryptographic proofs tying inference outputs to committed model weights [5]. These methods offer different assurances at different costs.

3.2 Privacy

Privacy is a separate matter which, for sensitive work, can dictate whether an inference service is usable at all. AI requests may contain confidential business information, or, in the case of applications which deliver intelligence to end users, sensitive personal data. Transforming sensitive information, or removing it before a request reaches a provider, can protect privacy but also strip away context the model needs. The challenge, then, is to redact only what's necessary and split tasks between trusted and untrusted endpoints, while making it easy for users to decide what to share without disrupting their workflows.

3.3 Security

Security becomes crucial when AI applications act on behalf of users, as is the case with agents. A malicious inference provider or even a rogue model can embed instructions in a response that direct the agent to take actions the user never intended [6]. For instance, if a user asks an agent to install a software package, the response may direct them to a malicious substitute. A response filter could check that instruction against the user's original request and flag it before it ever reaches the agent.

These are distinct problems which cannot be solved by any single mechanism. Purification may begin by enabling greater confidence in providers offering access to the same model. Extending interchangeability between models requires separate judgments about whether their outputs are suitable for the same task.

4. From Competition to Commitments

Purification can give buyers access to existing inference capacity they would usually avoid on trust grounds, giving them more viable providers for the same model.

With several acceptable choices on offer, the best option depends on the workload. A voice app might prefer to pay more for lower latency, while a batch job wants higher throughput per dollar. A critical service may opt for the most reliable provider available. An open inference market should give every provider on the Pareto frontier a chance to win the workloads it serves best [7].

In a spot market, buyers can choose between providers with enough capacity available to service requests when they arrive. But a buyer planning a product launch needs to know whether a suitable provider will have room for them come launch day.

A provider faces similar uncertainty. They must first decide how much capacity to reserve, or how much to add, before they know how much demand there will actually be.

This calculation is different when a buyer commits to purchasing capacity ahead of time. Under a forward commitment, the buyer agrees to purchase a specified amount of inference capacity, and the provider agrees to make that capacity available during an agreed window, at an agreed price and level of service. This allows both sides to better plan for the future.

5. Lessons from Dojima

For such commitments to become tradable, additional market design would be required. The Dojima Rice Exchange, officially authorized in Osaka in 1730, established rules for standardized contracts, margin and clearing. It is considered the first organized futures exchange [8].

At no point was every sack of rice made physically identical to allow for trading. Instead, the market established the terms under which obligations related to rice could be transferred. The contracts themselves could be fungible, even when the underlying rice varied. The same distinction works for inference.

A number of different providers may give comparable offers for the same work, but such offers are not the same as interchangeable service contracts. A tradable commitment might only be for a specific model, volume of inference, minimum service standards and delivery window. Any qualifying provider could then compete to fulfill it.

The market must then standardize those contracts enough for trading to concentrate around them, all while

preserving the differences between services that buyers care about.

6. Building the Market

Purification can give evidence that a service meets the terms of those contracts, but there will still be a need for clear rules around settlement, liquidity and reference prices.

Separately, a credibly neutral market would need transparent guardrails for verification and participation, applied without granting special privileges to any one provider or group.

Blockchain could establish shared records and rules for trading commitments, or automatically manage payments. Writing obligations onchain does not, however, establish whether inference was delivered to the agreed standard. That still requires evidence of work performed offchain, as well as methods of resolving disputes.

Many cloud services already offer GPU reservations. Compute futures have also been proposed to address related risks [9, 10]. But a GPU-hour does not specify which model will run or whether the service meets the buyer's requirements.

Buyers could secure future delivery of a specified inference service via standardized, tradable commitments. Providers could then use those commitments to plan investment in more capacity, and potentially seek financing against the contracted future demand.

7. Economy of Access

As AI increasingly underpins how people, businesses, institutions and governments operate, control over inference becomes a source of leverage.

Granted, inference already boasts a striking range of models and providers. But the choices shown on a routing page aren't the same choices which are practically available to every builder. An option that cannot be used is no option at all.

This is the impetus for a credibly neutral global inference market, giving independent providers everywhere a chance to compete for demand previously unavailable to them.

Such a market could bring the world's intelligence to more people in more places, and give them the freedom to put it to work on their own terms.

References

[1] M. Demirer, A. Fradkin, N. Tadelis, "The Emerging Market for Intelligence: How Firms Buy and Sell AI," *Journal of Economic Perspectives*, 40(3), 23–46, 2026. https://doi.org/10.1257/jep.20261506

[2] OpenRouter, "Provider Selection." https://openrouter.ai/docs/guides/routing/provider-selection

[3] W. Cai, T. Shi, X. Zhao, D. Song, "Are You Getting What You Pay For? Auditing Model Substitution in LLM APIs," arXiv:2504.04715, 2025. https://arxiv.org/abs/2504.04715

[4] T. Chitra, "The Souq: Value Creation and Extraction in OpenRouter," Robot Ventures, August 17, 2026. https://robvc.com/research/the-souq

[5] R. Rinberg, A. Allain, A. Chiesa, M. Christ, R. Dalal, G. Fenzi, B. Mayer, G. Weissenberg, "Progress on provable inference for large language models," Anthropic, 2026. https://www.anthropic.com/research/provable-inference

[6] E. Debenedetti, J. Zhang, M. Balunović, L. Beurer-Kellner, M. Fischer, F. Tramèr, "AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents," *Advances in Neural Information Processing Systems*, 37, 2024. https://proceedings.nips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html

[7] Z. Chen, Y. Ye, Z. Zhou, "Online Linear Programming for Multi-Objective Routing in LLM Serving," *Proceedings of the 43rd International Conference on Machine Learning*, PMLR 306, 17892–17912, 2026. https://proceedings.mlr.press/v306/chen26ft.html

[8] Osaka Dojima Exchange, "The Birthplace of Futures Trading." https://www.odex.co.jp/en/about/derivative/index.html

[9] Intercontinental Exchange, "ICE and Ornn to Launch GPU Compute Futures Contracts," May 19, 2026. https://ir.theice.com/press/news-details/2026/ICE-and-Ornn-to-Launch-GPU-Compute-Futures-Contracts/default.aspx

[10] Commodity Futures Trading Commission, "CFTC Requests Comment on the Listing of Compute Derivatives Contracts," August 19, 2026. https://www.cftc.gov/PressRoom/PressReleases/9286-26