The price floor for AI inference.

Floor Compute is an open marketplace for AI inference behind one OpenAI-compatible endpoint. Sellers list competing offers for the same models, a five-axis score picks the best route for every request, failover is automatic, and every dollar settles two-sided, exactly. There is no single market price for a token — there is a spread, and a floor. We find the floor, on every request.

ABSTRACT

Open-model inference is sold by dozens of providers at prices that differ by multiples for the identical model, and those prices move week to week. Developers wire their code to one provider, inherit its price and its outages, and rarely re-shop. Floor Compute is a two-sided marketplace in front of that spread: buyers keep the OpenAI API they already use, sellers of GPU capacity self-onboard and list offers, and a matching engine scores every active offer on price, latency, uptime, liquidity, and health, routes to the best one, and settles the money on both sides of the trade. This document covers the problem, the market mechanics, the matching score, the economics, and the path toward an open routing network.

01 A fragmented, opaque market

The same open model, say a 70B-class chat model, is served today by many providers, each with its own price, latency profile, rate limits, and reliability. Prices for the very same model routinely run 2 to 5 times higher from one provider to the next, and they shift as capacity and competition change. There is no single market price for a token. There is a spread.

Yet the typical integration is static. A developer points their SDK at one provider, hard-codes a base URL and a key, and ships. From that moment they pay that provider's price and absorb that provider's outages and rate limits, even when a healthier provider is serving the same model for a fraction of the cost a few milliseconds away. Re-shopping means code changes, new accounts, new keys, and new failure modes, so almost nobody does it. That spread is real money left on the table on every request.

02 Floor Compute in one line

Floor Compute is an open marketplace for inference: sellers register offers, a model served at a price from their endpoint, and for every buyer request the matching engine scores all active offers for that model on five axes, routes to the best one, fails over automatically, and settles payment to both sides. Buyers change a base URL and keep their code. Sellers plug in capacity and compete on price and reliability.

Floor Compute does not build models and does not own GPUs. It makes providers compete. The first-party catalog is listed as one seller among many, so the order book is never empty while the market grows around it.

03 How the market works

- **OpenAI-compatible gateway (the buy side).** Standard /v1 chat, completions, and embeddings requests. A request names a model, the "auto" alias, or a routing suffix such as :cheapest or :fastest. Existing OpenAI SDKs work by changing only base_url and key.
- **The offer book (the sell side).** An offer is the unit of competition: a model, served from a seller's endpoint with the seller's own credentials, at a seller-set price, with declared capacity. Offers are validated before they go live and carry a live 0 to 100 health score after.
- **Matching engine.** For each request, every ACTIVE offer for the model is filtered for capability first (context window, required features, unsupported parameters, cost ceiling), then scored on five axes. The best route wins; the runners-up become the failover order.
- **Dispatch and failover.** The winning seller's endpoint serves the request, streaming included. If it fails before the first byte reaches the buyer, a rate limit, a 5xx, a timeout, the engine retries the next-best offer. Total attempts are capped and every hop is recorded.
- **Settlement.** The buyer is charged the offer price. The marketplace keeps a flat fee. The seller is credited the remainder. All three amounts reconcile exactly, per request.

04 The matching score

Every eligible offer is scored on five normalized axes. The weights below are the defaults; buyers can re-weight per API key or per request. The best route wins.

AXIS	WEIGHT	SOURCE
Price	0.40	The seller's offer price, normalized across the book
Latency	0.20	Live estimate learned from real dispatches
Uptime	0.20	Rolling success rate of the offer
Liquidity	0.10	Declared and remaining capacity
Health	0.10	Live 0 to 100 score from outcomes and probes

```
score(offer) =  0.40·price + 0.20·latency + 0.20·uptime
              + 0.10·liquidity + 0.10·health

route → argmax_offer score(offer)
```

Routing is transparent, not magic. Every response carries an x_floor object naming the selected offer, the full five-axis score breakdown, the billed rate, the failover order that was computed, and the measured savings versus the priciest eligible offer. Any decision can be audited after the fact, and previewed before it with the routing preview endpoint.

05 Economics

Buyers pay the seller's offer price. No subscriptions, no token markup, no hidden spread. The marketplace keeps a flat 5 percent of each trade, deducted from the seller's side, and the seller is credited the rest: 95 percent of every dollar the buyer spends goes to the capacity that served it.

Money is exact. Each request is metered once, in micro-dollars, against separate input and output prices; the buyer charge, the marketplace fee, and the seller settlement reconcile to the micro-dollar on every row of the ledger. Buyer balances are prepaid, topped up by card or USDC, and spend hard-stops at zero. Because the engine routes to the best healthy offer and the spread is wide, the savings claim is measured per request, not asserted.

06 Selling on Floor Compute

Anyone with capacity can join the sell side: GPU operators, model hosts, resellers of committed throughput. Onboarding is self-serve. Every provider deserves a chance to compete; no provider is guaranteed traffic.

- **Register an endpoint.** An OpenAI-compatible URL plus a credential. Endpoints must be public https; they are screened against private and internal addresses at registration and re-checked at dispatch. Credentials are encrypted at rest and are never logged or returned.
- **List offers.** Model, input and output price per million tokens, context window, features, capacity. An offer moves draft to validation to active; the validation probe confirms the endpoint actually serves the model before it can enter the book.
- **Compete on merit.** Live health (0 to 100) is learned from every dispatch outcome and an optional background probe. Degraded offers are deprioritized automatically and recover the same way. Price cuts move an offer up the book instantly.
- **Get paid per request.** Revenue is credited atomically with each settled request and tracked in a seller balance, with payouts requested from the seller console.

07 Reliability and failover

Offer health is learned from live traffic. Every dispatch outcome feeds latency and success-rate estimates, a per-provider circuit breaker, and the persisted 0 to 100 health score, so a degrading offer is deprioritized before it hard-fails. When a dispatch does fail before streaming starts, a rate limit, a 5xx, a timeout, or a provider-side configuration fault, the prompt takes another path within the same request; once a stream has begun it is never silently switched. Uptime is the product: the buyer sees one reliable endpoint, not the churn behind it.

08 Privacy and security

Buyer API keys are hashed at rest and shown only once. Seller credentials are encrypted at rest and decrypted only at dispatch. Floor Compute logs identifiers, latency, and cost, the data needed to route and bill, and not prompt or completion content. Compatibility is drop-in, so your data paths stay under your control. Only the base URL changes.

09 Toward a routing network

The protocol grows in phases, from an aggregator to the routing layer for AI.

- **Phase 1, one channel (shipped).** A single OpenAI-compatible endpoint over the whole provider market, a validated catalog, and routing you can see through.
- **Phase 2, routing that learns (shipped).** Live health and latency folded into the score, automatic failover, per-request analytics, measured savings.
- **Phase 3, open the market (live now).** Self-serve seller onboarding, competing offers, two-sided settlement, public price boards and market analytics.
- **Phase 4, the network layer.** Response caching, deeper capacity commitments, richer payout rails, and the default routing layer for open-model inference.

10 Conclusion

Inference is becoming a commodity, and commodities are won by the marketplace that connects demand to the best reliable supply, not by owning the supply. Floor Compute is that marketplace: one endpoint, an open offer book, transparent five-axis matching, exact two-sided settlement, and automatic failover. The future of AI should not belong to one provider. It should belong to the best provider for every request. Integrate once, and every request finds the floor.
