Inference Arbitrage Router: Cost-Quality Optimisation Across Heterogeneous LLM Providers

ForceDream Research Team, Intelligence OS Division | 2026-05-14 | v1.2 | 4 pages
Category: Inference | Layers: L1, L3, L4
URL: https://forcedream.com/research/inference-arbitrage-router-llm-cost-optimisation

WORM ACCESS SEAL | L828

fd2026001a3f7c2b

Abstract

We present the ForceDream Inference Arbitrage Router (IAR), a dynamic multi-objective routing architecture that continuously arbitrages inference requests across heterogeneous large language model providers. The IAR evaluates four primary routing signals to construct a Pareto-optimal dispatch decision within a single request lifecycle. We demonstrate 43.2% cost reduction while maintaining quality scores within 4.1% of the theoretical maximum. All routing decisions are WORM-sealed at dispatch.

1. Introduction

Modern AI infrastructure requires routing inference requests across multiple providers to optimise for cost, quality, and latency simultaneously. Single-provider strategies expose systems to pricing volatility, capacity constraints, and quality variance across task types. The ForceDream Inference Arbitrage Router addresses this through continuous multi-objective optimisation at the routing layer, operating below all application-level business logic and above the raw provider APIs.

2. Architecture

The IAR maintains a provider state matrix P of dimensions n x 4, where n is the number of registered providers and the four columns represent cost-per-token, observed latency (EWMA, alpha=0.3), quality score (WORM-derived), and availability (binary, 200ms intervals). At dispatch time, the router solves a lightweight Pareto optimisation over P conditioned on the selected priority mode. The priority mode maps to a target region on the Pareto frontier.

3. Priority Modes

Four priority modes are supported. Balanced: equal weighting across all signals, producing the social optimum (CBEI=0.94). Cheapest: cost weight=0.8, minimising spend subject to quality floor. Fastest: latency weight=0.8, minimising round-trip time. Quality: quality score weight=0.8, maximising output quality. All modes produce allocations on or near the Pareto frontier.

4. WORM Audit Integration

Every routing decision is sealed at dispatch using SHA-256 over (provider_id, task_hash, timestamp_ms, priority_mode). The seal is stored in the WORM ledger before the request is forwarded. This creates an immutable audit trail suitable for FCA Consumer Duty PS22/9, MiFID II COBS, and equivalent regulatory

frameworks.

5. Results

Evaluated across 90 days of production traffic. Cost reduction: 43.2% vs single-provider baseline (95% CI: 41.8-44.6%). Quality degradation: 4.1% vs theoretical maximum ($p<0.001$). Routing latency overhead: 8ms (SD=2.3ms). Failover latency: 34ms. Uptime: 99.97%. Zero missed WORM seals across 4.2M financial events.

6. Conclusions

The IAR demonstrates that multi-provider inference arbitrage is viable at production scale without meaningful quality degradation. The 43.2% cost reduction compounds significantly at scale. The WORM-sealed audit trail provides a regulatory-grade record of every routing decision.

Live API Endpoints

```
POST /v1/inference/route
POST /v1/inference/race
GET /v1/models/pricing
GET /v1/models/list
```

Citation

ForceDream Research Team (2026). Inference Arbitrage Router. ForceDream Intelligence OS Research Series, FD-2026-001. https://forcedream.com/research/inference-arbitrage-router-llm-cost-optimisation

---