Comparison of LLM Pricing Models: Paper Token vs Provision Support

Introduction to LLM Pricing Models

LLM providers offer different pricing models to balance cost, performance, and scalability. Two commonly discussed models are: - Paper Token Model: Pay-as-you-go based on token usage. - Provision Support Model: Subscription or reserved compute with predictable cost and speed.

Comparison Overview

Key aspects of comparison include: - Cost - Credits vs Tokens - Processing Speed - SLA & Reliability

How Tokens Work (Example)

Tokens are small units of text (≈4 characters or ¾ word). Example: User prompt: 'Write a poem about AI' → ~6 tokens. Billing = (Input tokens + Output tokens) × Price per token.

How Credits Work (Example)

Credits are an abstraction used by some platforms (e.g., Windsurf, Hugging Face). 1 Credit = a fixed number of tokens or compute usage. For example, 100 credits might equal 1M tokens or fixed GPU hours.

Processing Speed Impact

- Paper Token Model: Shared compute, variable latency, may throttle. - Provision Support Model: Dedicated compute, predictable speed, SLA-backed performance.

Summary Table

See the comparison table below.

Conclusion & Recommendations

- Paper Token Model: Best for casual use, experimentation, and cost-sensitive workloads. - Provision Support Model: Best for production, enterprise apps, requiring low latency and reliability.

Feature	Paper Token Model	Provision Support Model
Cost	Pay-as-you-go, per token	Fixed or subscription-based
Credits/Tokens	Tokens billed directly	Credits mapped to compute or tokens
Processing Speed	Variable, depends on load	Predictable, SLA-backed
Consistency	May vary under high demand	High consistency

SLA Support	Generally none	Enterprise-grade SLAs
Best For	Casual/dev workloads	Production/enterprise workloads