AI gateways: Can telcos turn AI cost control into a business?

AI gateways are becoming an important control layer for managing the cost, routing and governance of AI usage. AT&T is using this capability internally to optimise its own AI estate, while China Mobile and China Telecom are taking similar capabilities to external customers.

Xiaoya Sun, Research Analyst

September 2026

AI is moving from experimentation into everyday enterprise use, and the cost and complexity of managing that usage is growing with it. As organisations adopt more models across more applications and workflows, they need better ways to control spend, route requests efficiently and maintain visibility over how AI is being used. One response is simply to restrict usage through budgets, quotas or hard limits. That keeps spending under control, but it can also limit how extensively organisations use AI, creating demand for ways to make each unit of consumption more efficient instead.

That is one of the problems the AI gateway has emerged to address. Rather than allowing each application, employee or agent to connect independently to different AI models, an AI gateway sits between them and the models they use, providing a central point from which an organisation can manage AI traffic. The functionality varies by provider, but commonly includes routing requests between models, monitoring token consumption and cost, caching repeated requests, applying access and security policies, and providing observability across the AI estate. In cost optimisation, the central idea is simple: not every task needs the newest, largest or most expensive model, so an intelligent gateway can direct a request to the model that best balances cost, performance, quality and latency.

A sizeable market has already formed around this capability, with hyperscalers, specialist AI infrastructure companies and networking and security vendors all offering versions of an AI gateway. Telcos are now entering the same space. Telcos are now entering the same space, using broadly similar technology but applying it in different ways.

AT&T: using an AI gateway to control its own AI spend

One example of how telcos are applying this capability is AT&T, which has built its AI gateway primarily for internal use. In July 2026, the operator described its approach under the label tokenomics, treating tokens primarily as a cost to be managed rather than a product to be sold (for now). With average consumption of around 45 billion tokens a day, AT&T has a strong incentive to manage AI spend more efficiently, and its proprietary gateway sits at the centre of that effort by routing each prompt to the most cost-effective model capable of completing the task.

The logic is straightforward: defaulting to the newest and largest frontier models is unnecessarily expensive when many tasks can be handled by smaller or cheaper models without sacrificing the required performance. AT&T's in-house gateway therefore uses cache-aware routing to weigh speed and cost against expected output quality, selecting the most appropriate model for each task and even switching models during a multi-turn session where that makes sense. The operator says this approach has reduced AI costs by as much as 90%.

Cost optimisation is only one part of the broader stack. AT&T also combines the gateway with specialised telco models, including OTel 2.0, an open-source model post-trained on more than 400 billion telecom-specific tokens, alongside governance and security capabilities. Together, these form part of what AT&T describes as a sovereign AI stack. For now, however, the role of the gateway itself is primarily internal: helping the operator use AI more efficiently while retaining control over model selection, data and costs.

Chinese operators: taking the gateway to customers

Chinese operators have built much the same core capability, but are taking it in a more explicitly commercial direction. Since early 2026, the term tokenisation has become prominent in China's telecoms industry, including at MWC Shanghai, where token-based approaches to AI monetisation featured in discussions around new operator business models. In practice, the model is relatively straightforward: operators pool AI compute and models, route customer requests between them, meter the resulting usage and sell access to that capacity. The token becomes the billing unit, while the gateway and orchestration layer underneath determine how efficiently that service can be delivered.

China Mobile has set out this approach through its Mixture of Models and Agents (MoMA) platform, which combines compute, orchestration and services into an end-to-end layer through which customers can access different AI models. At its base is a unified API gateway connected to more than 300 models, including China Mobile's own Jiutian model alongside third-party options such as DeepSeek, Doubao, Qwen and Kimi. An intelligent routing engine then analyses each request and applies one of three strategies — cost first, performance first or balanced — to match it to the most appropriate model. The service layer turns those capabilities into applications and services for sectors including government, finance, industry, healthcare and education, and China Mobile claims its inference engine, routing and optimisation techniques have reduced unit token costs by around 30% (see figure below).

China Mobile's MoMA platform displayed at MWC Shanghai 2026

Source: STL Partners

China Telecom has approached the same monetisation problem from an infrastructure angle. Its TokenHub is presented as a proprietary distributed inference network, combining the operator's own Xingchen AI foundation models with a selection of third-party alternatives. A self-built AI inference gateway sits in front of it, directing inference requests and allocating tokens to individual tasks as they arrive. TokenHub is sold as a monthly subscription: the entry-level tier provides 25 million tokens for RMB 29 (USD 4) per month, rising to 680 million tokens for RMB 699 (USD 104) per month at the top of the range.

Where can telcos add value?

As AI gateways become a more established part of the AI stack, the opportunity for telcos is likely to lie less in the gateway itself than in what they can combine it with. Routing, metering and policy control are already offered by a growing number of providers, so a standalone gateway is becoming increasingly difficult to differentiate. Operators may have a stronger proposition where the gateway sits within a broader bundle of connectivity, compute, security, sovereignty and managed services.

A sensible starting point is internal use. AT&T shows how an operator can use a gateway to reduce inference costs and gain greater control over its own AI estate, while also building practical experience of running the technology at scale. That capability could then form part of a broader customer proposition, as China Mobile is beginning to test.

The question for telcos is therefore not whether they can build an AI gateway, but whether they can combine it with other assets in a way that creates a sufficiently differentiated offer. In an increasingly crowded market, the strongest opportunity may be to treat the gateway as one component of a broader AI proposition rather than as a standalone product.

Xiaoya Sun is a Research Analyst at STL Partners, specialising in AI monetisation and satellites.

Get in touch with the author to learn more

xiaoya.sun@stlpartners.com

Or visit STL Partners' AI Hub

https://stlpartners.com/ai/