Preprints.org

Article

Not peer-reviewed version

Expert Routing and Self-Preferencing on AI Platforms

Xufeng Zhang *, Han Li , Shenghui Bao

Posted Date: 13 March 2026

doi: 10.20944/preprints202603.1069.v1

Keywords: AI platforms; self-preferencing; routing; expert advice; platform neutrality

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Expert Routing and Self-Preferencing on AI Platforms

Xufeng Zhang [1,*], Han Li [2] and Shenghui Bao [3]

[1] Resp AI Research Lab, CIIOE, Xiamen, 361000, China
[2] Lanzhou University, Lanzhou, 73000, China
[3] National University, Manila, 1008, Philippines
* Correspondence: xufeng@nau.edu

Abstract

This paper develops an industrial-organization theory of AI routing, modeling how a dual-role platform allocates user queries between an in-house model and an outside expert. We formulate this as a delegated allocation problem featuring endogenous quality investment and data feedback. The expert sets access prices and initial quality, while the platform routes queries by difficulty. Early traffic routed to the expert enhances its future quality through learning-by-serving. In equilibrium, routing follows a cutoff rule. The platform's self-preferencing acts as a tax on outside expertise, raising routing thresholds, reducing outside demand, and compressing both current quality investment and future learning gains. Decentralized routing introduces three inefficiency wedges compared to a dynamic first best: an access-markup wedge, a bias wedge, and a data-feedback wedge. The third is unique to AI routing because traffic allocation directly dictates learning opportunities. Consequently, neutrality and access-pricing remedies are complementary but insufficient together, as platforms fail to internalize the future value of outside learning. This model provides a tractable framework for analyzing AI gateways and router governance.

Keywords: AI platforms; self-preferencing; routing; expert advice; platform neutrality

JEL Classification: L13; D82; D83; L86; O33

---

1. Introduction

Large language models and adjacent AI systems have created a new layer of intermediation. In many commercially relevant deployments, users do not choose directly among transparent menus of models. Instead, a platform, enterprise gateway, assistant, or API aggregator receives the query, evaluates the available systems, and decides which model answers. Engineering practice has made this problem explicit under the label of routing. A router may send simple requests to a cheap in-house model, difficult requests to a stronger frontier model, and niche requests to a specialized system. Recent computer-science contributions have shown that such routing can materially improve the cost–quality frontier of AI deployment. Chen et al. [1] study budget-aware cascades, Ong et al. [2] learn routers from preference data, and Panda et al. [3] cast the deployment problem as contextual-bandit routing under budget constraints. Those papers are important for system design, but they typically treat the router as an algorithmic object rather than as a strategic economic actor.

From an industrial-organization perspective, that abstraction is too strong. The router is often a vertically integrated intermediary with market power, a business model, and its own model on the menu. Once the routing layer is recognized as an intermediary, familiar questions from platform economics and biased intermediation immediately reappear. If the gateway owns one candidate model, will it self-preference that model? If outside experts must pay for access or accept unfavorable API terms, how are allocation, price, and investment distorted? If traffic to outside experts also generates data, learning-by-serving, or capability improvement, who internalizes the long-run value of referral?

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

In AI markets, routing is therefore not only a prediction problem. It is also a market-design problem in which the intermediary allocates both current demand and future learning opportunities.

This paper develops a unified model of that problem. We formulate "follow the expert" as delegated allocation by a dual-role platform. A unit mass of users arrives in each of two periods, each user indexed by query difficulty. The platform owns an incumbent model. An outside expert offers superior gross answer quality, and its advantage widens with difficulty, but the platform must pay an access price to use it. The outside expert also chooses a quality level through costly investment. After period 1 routing, the outside expert's period-2 quality increases with period-1 routed demand. This feedback captures a broad family of economically relevant mechanisms: data accumulation, learning-by-serving, domain adaptation, reputation building, or a larger installed base over which the expert can amortize improvement.

The model deliberately stays one-dimensional. We do not model the internal architecture of an LLM, user-side search over a menu, or horizontal differentiation across many experts. The value of the abstraction is that it lets us isolate the industrial-organization margins that matter most for routing governance. There are four of them. First, the platform allocates traffic across competing technologies. Second, the outside expert sets an access price. Third, the outside expert invests in quality in anticipation of traffic. Fourth, current traffic shapes future quality. Those four margins already generate a rich theory of market power at the routing layer.

Two features deserve emphasis at the outset. The first concerns the interpretation of comparative advantage. In our specification, the outside expert is not worse in gross answer quality on easy queries. Rather, it is absolutely superior, and its advantage expands with difficulty. The reason a cutoff nonetheless emerges is that the platform faces an access price for the outside expert and may additionally obtain a private benefit from using its own incumbent model. The cutoff therefore reflects routing costs and integrated-platform incentives, not a literal crossing of production frontiers in gross quality. This interpretation is natural for many contemporary deployments in which the stronger external model is technically better across the board but too expensive or strategically inconvenient to invoke for every query.

The second feature concerns dynamics. In many digital settings, current demand affects future capability. In AI, that channel is especially salient. More traffic can mean more labeled outcomes, more comparative feedback, more opportunities for fine-tuning, more observed failure cases, and stronger incentives to invest in domain-specific quality. Our dynamic extension does not replace the static model with a different one. It extends the same primitives. The same quality choice, the same wholesale access price, and the same platform bias govern both periods. Period-1 routing determines period-2 quality through a transparent law of motion. That unified structure is central because it allows us to study how pricing, self-preferencing, and learning interact inside one equilibrium rather than across loosely connected submodels.

The paper relates to several literatures. The closest economic foundations come from platform economics and biased intermediation. Classic work on two-sided and multisided platforms emphasizes that intermediaries shape market outcomes not only through prices, but also through the access conditions and allocation rules they impose on participants [4–6]. The literature on intermediary bias and dual-role platforms shows that vertically integrated gatekeepers may divert traffic toward affiliated sellers or first-party offerings [7–10]. Our setting fits squarely in that tradition, but with a distinct AI twist: routing does not only redirect current trade; it also governs the future quality path of specialized experts.

The paper is also related to the literature on search, ranking, and recommendation. Search engines, marketplaces, and recommender systems do not merely reveal information; they shape what users see and what suppliers can profitably offer. Athey and Ellison [11] and de Cornière [12] study search environments in which intermediary design affects market outcomes. Che and Hörner [13] show that recommender systems may need to distort current recommendations in order to facilitate socially valuable learning. We take a parallel idea into AI routing, but the learning object is not only user

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

beliefs or platform information. It is the outside expert's own quality trajectory. When routing to the expert raises future expert quality, the planner values referral more than a myopic or self-preferring platform does.

A third connection is to innovation and information. Arrow [14] made precise the idea that current production can increase future capability through learning-by-doing. Akcigit and Liu [15] show how informational frictions shape innovative effort and market structure. In our model, routed traffic is the analog of productive experience. If the platform restricts access to hard or high-value queries, it reduces the scale on which outside experts can recover current costs and improve future quality. The market structure of AI intermediation therefore affects the direction and level of capability investment.

Against that background, our contribution is fourfold.

First, we provide a tractable industrial-organization model of AI routing in which the routing rule, access pricing, quality investment, and learning-by-serving are jointly determined. For any given expert quality and access price, the platform routes by a cutoff rule in query difficulty. This structure turns "follow the expert" into a delegated screening problem: easy queries stay in-house, and difficult queries are escalated outward. Because the expert's gross advantage grows with difficulty, the geometry is simple and closed form.

Second, we show that self-preferencing acts as a tax on outside expertise. A larger platform bias toward the incumbent raises the routing threshold, reduces the outside expert's demand in period 1, and thereby lowers future expert quality as well. The dynamic effect is not an add-on. It is the product of the same demand reduction that already distorts static routing. Once traffic is also a learning input, the harm from bias is amplified.

Third, we characterize the outside expert's pricing and investment problem in closed form and show that data feedback makes the incidence of bias more severe. Stronger data feedback raises the return to outside traffic and therefore increases equilibrium investment under neutral governance. But the same feedback also magnifies the damage from self-preferencing because every lost query is simultaneously lost revenue and lost future capability. In that sense, the industrial-organization consequences of routing bias are larger in environments where traffic and learning are tightly linked.

Fourth, we derive a dynamic first-best benchmark and identify three distinct wedges between decentralized routing and efficient routing. The first is the access-markup wedge: the platform compares the access price to zero, whereas society compares real resource cost to zero. The second is the bias wedge: a self-preferring platform attaches extra private value to using the incumbent. The third is the data-feedback wedge: the platform does not internalize that routing a query outward today raises future expert quality. This decomposition yields a sharp governance implication. Neutrality rules reduce the bias wedge. Access-pricing remedies reduce the markup wedge. But even a neutral platform with marginal-cost access pricing still under-routes relative to the dynamic first best when it fails to internalize future outside learning. Hence neutrality, access pricing, and data-governance instruments are complements rather than substitutes.

Our theory is intentionally parsimonious, but it speaks directly to current debates about AI gateways, enterprise model hubs, regulated escalation systems, and vertically integrated assistants. In those environments, the economically relevant question is often not which model is globally best in the abstract. It is who controls the router that decides which model is used for which query, under what commercial terms, and with what consequences for future competition. Once the router is modeled as a strategic intermediary, platform economics becomes central to AI governance.

The rest of the paper proceeds as follows. Section 2 presents the model. Section 3 solves for equilibrium routing, pricing, and investment and derives the comparative statics of self-preferencing and data feedback. Section 4 studies the dynamic planner's benchmark and the welfare-relevant wedges generated by decentralized routing. Section 5 discusses policy implications, empirical predictions, and extensions. Section 6 concludes.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

2. Model

There are two periods, $t \in \{1, 2\}$, and in each period a unit mass of users arrives on the platform. A user is indexed by query difficulty $\theta \in [0, 1]$, distributed uniformly. Higher θ means that the query is more difficult. The platform can route each query to one of two AI systems.

The first system, denoted I for incumbent, is owned by the platform. The second, denoted E for expert, is supplied by an outside firm. The incumbent's gross user utility is

$$u_I(\theta) = v - \theta, \tag{1}$$

where $v > 0$ is a common baseline. In period t, the outside expert's gross user utility is

$$u_{E,t}(\theta) = v + q_t - \gamma\theta, \tag{2}$$

where $q_t \geq 0$ is expert quality in period t and $\gamma \in (0, 1)$. Define

$$\delta \equiv 1 - \gamma \in (0, 1). \tag{3}$$

Then the expert's gross advantage over the incumbent is

$$u_{E,t}(\theta) - u_I(\theta) = q_t + \delta\theta. \tag{4}$$

Thus the expert is grossly superior for every θ whenever $q_t > 0$, and that superiority grows with difficulty.

The expert has marginal inference cost $m > 0$ per routed query. Before routing begins, the expert chooses two variables. First, it chooses an initial quality level $q \equiv q_1 \geq 0$ at convex cost

$$\frac{\kappa}{2}q^2, \tag{5}$$

where $\kappa > 0$ is the investment-cost parameter. Second, it chooses a constant wholesale access price $w \geq m$ paid by the platform for each query routed to the expert in either period. We treat the constant wholesale price as a long-term access contract. Allowing period-specific prices would add algebra but not alter the basic logic that the platform faces an access wedge relative to real cost.

The platform is a dual-role intermediary. If it routes a query to its own incumbent model, it obtains a private benefit $b \geq 0$. This reduced-form term captures self-preferencing, internal accounting advantages, traffic-retention motives, data capture, brand control, or ecosystem complementarity. Hence, in period t, the platform compares

$$\Pi_I(\theta) = u_I(\theta) + b \tag{6}$$

with

$$\Pi_{E,t}(\theta) = u_{E,t}(\theta) - w. \tag{7}$$

The platform therefore routes to the expert in period t whenever

$$q_t + \delta\theta \geq w + b. \tag{8}$$

The dynamic link is a data-feedback or learning-by-serving channel. If the expert receives demand D_1 in period 1, then period-2 expert quality becomes

$$q_2 = q + \lambda D_1, \tag{9}$$

where $\lambda \geq 0$ measures the strength of data feedback. The parameter λ can be interpreted broadly. More routed traffic may generate more outcome labels, more comparative user feedback, more observations

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

of failure modes, more opportunities to specialize, or a larger installed base over which improvement can be amortized. Our law of motion is deliberately simple: one more unit of period-1 traffic raises period-2 quality by λ.

The timing is as follows.

1. Parameters $(m, \delta, \kappa, \lambda, b, \beta)$ are given, where $\beta \in (0, 1]$ is the discount factor on period 2.
2. The outside expert chooses initial quality q and wholesale access price w.
3. The platform observes each user's difficulty θ and routes period-1 queries to I or E.
4. Period-1 demand D_1 updates expert quality to $q_2 = q + \lambda D_1$.
5. The platform routes period-2 queries using the same access price w and bias parameter b.
6. Payoffs are realized.

We maintain the following parameter restriction.

Assumption 1. *Define*

$$A(\lambda) \equiv 1 + \beta\left(1 + \frac{\lambda}{\delta}\right). \tag{10}$$

Parameters satisfy

$$2\kappa\delta > A(\lambda), \qquad \beta\lambda^2 < \delta^2, \tag{11}$$

and

$$0 < \delta - m - b < \frac{2\kappa\delta - A(\lambda)}{\kappa(1 + \lambda/\delta)}. \tag{12}$$

The first inequality guarantees strict concavity of the expert's optimization problem. The second guarantees strict concavity of the planner's dynamic routing problem in Section 4. The third ensures an interior equilibrium: the expert serves a positive but not full measure of queries in both periods. These restrictions are transparent economically. The expert must be valuable enough on difficult queries to attract some traffic, but not so valuable that it serves the entire market; investment cannot be arbitrarily cheap; and data feedback cannot be so explosive that a small amount of initial traffic collapses the cutoff immediately.

Because users do not directly choose the model and make no explicit monetary payment in this reduced-form environment, consumer surplus is simply expected user utility. Total welfare is expected user utility minus real inference cost and minus quality-investment cost. The wholesale price w is a transfer between the platform and the expert; it affects equilibrium behavior but not welfare directly.

Before solving the model, it is useful to state the interpretation clearly. The platform controls the allocation of queries across technologies. The expert chooses quality and access terms in anticipation of that control. Current routing determines future quality because routed traffic generates knowledge and incentives. The industrial-organization object of interest is therefore not merely the ranking of existing models, but the governance of the traffic-allocation layer itself.

3. Equilibrium Routing, Pricing, and Investment

We solve the model by backward induction. For any expert quality and access price, the platform chooses routing in each period. Anticipating those routing rules and the period-2 quality feedback they induce, the expert chooses (q, w).

3.1. Cutoff Routing in Both Periods

For a given pair (q, w), expert quality in period 1 is $q_1 = q$. Period-1 routing to the expert occurs whenever

$$q + \delta\theta \geq w + b. \tag{13}$$

Define the period-1 cutoff by

$$t_1(q, w, b) = \frac{w + b - q}{\delta}. \tag{14}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

Let $D_1(q, w, b)$ denote the mass of period-1 queries routed to the expert.

Proposition 1 (Cutoff routing and dynamic propagation). *For any (q, w, b), the platform's routing rule is a cutoff rule in both periods. If the implied cutoffs are interior, then in period 1,*

$$D_1(q, w, b) = 1 - t_1(q, w, b) = \frac{q - w - b + \delta}{\delta}. \tag{15}$$

Period-2 quality is

$$q_2 = q + \lambda D_1(q, w, b), \tag{16}$$

and the period-2 cutoff is

$$t_2(q, w, b) = \frac{w + b - q_2}{\delta} = t_1(q, w, b) - \frac{\lambda}{\delta} D_1(q, w, b). \tag{17}$$

Hence period-2 expert demand is

$$D_2(q, w, b) = 1 - t_2(q, w, b) = \left(1 + \frac{\lambda}{\delta}\right) D_1(q, w, b). \tag{18}$$

Proof. The platform routes to the expert in period 1 whenever $q + \delta\theta \geq w + b$. Since $\delta > 0$, the set of routed types is the upper interval $[t_1, 1] \cap [0, 1]$, where

$$t_1(q, w, b) = \frac{w + b - q}{\delta}. \tag{19}$$

Whenever this cutoff is interior, period-1 expert demand equals the length of that upper interval:

$$D_1(q, w, b) = 1 - t_1(q, w, b) = \frac{q - w - b + \delta}{\delta}. \tag{20}$$

By the law of motion for expert quality,

$$q_2 = q + \lambda D_1(q, w, b). \tag{21}$$

In period 2, the platform routes to the expert whenever $q_2 + \delta\theta \geq w + b$, which implies the cutoff

$$t_2(q, w, b) = \frac{w + b - q_2}{\delta} = \frac{w + b - q}{\delta} - \frac{\lambda}{\delta} D_1(q, w, b) = t_1(q, w, b) - \frac{\lambda}{\delta} D_1(q, w, b). \tag{22}$$

Whenever this period-2 cutoff is interior, period-2 expert demand is

$$D_2(q, w, b) = 1 - t_2(q, w, b) = 1 - t_1(q, w, b) + \frac{\lambda}{\delta} D_1(q, w, b) = \left(1 + \frac{\lambda}{\delta}\right) D_1(q, w, b). \tag{23}$$

□

Proposition 1 gives the basic mechanism-design geometry. "Follow the expert" is implemented by partitioning the query space. Low-difficulty queries are kept in-house, while harder queries are escalated. The dynamic addition is equally transparent: a lower period-1 cutoff raises period-1 expert traffic, which increases period-2 expert quality and lowers the period-2 cutoff as well. Current routing therefore changes future routing through the quality law of motion.

3.2. The Outside Expert's Optimization Problem

Given Proposition 1, the outside expert chooses (q, w) to maximize discounted profit:

$$\pi_E(q, w; b, \lambda) = (w - m) D_1(q, w, b) + \beta(w - m) D_2(q, w, b) - \frac{\kappa}{2} q^2. \tag{24}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

Using Proposition 1, this becomes

$$\pi_E(q, w; b, \lambda) = A(\lambda)(w - m)D_1(q, w, b) - \frac{\kappa}{2}q^2, \tag{25}$$

where $A(\lambda) = 1 + \beta(1 + \lambda/\delta)$. Since

$$D_1(q, w, b) = \frac{q - w - b + \delta}{\delta}, \tag{26}$$

we may rewrite profit as

$$\pi_E(q, w; b, \lambda) = A(\lambda)(w - m)\frac{q - w - b + \delta}{\delta} - \frac{\kappa}{2}q^2. \tag{27}$$

The factor $A(\lambda)$ is a dynamic multiplier. When $\lambda = 0$, it equals $1 + \beta$: one unit of period-1 demand generates one contemporaneous margin and one discounted period-2 margin. When $\lambda > 0$, period-1 demand is more valuable because it also lifts period-2 quality and thus period-2 demand.

Proposition 2 (Unique interior equilibrium)**.** *Under Assumption 1, the expert's problem has a unique interior solution. Equilibrium initial quality and wholesale price are*

$$q^*(b, \lambda) = \frac{A(\lambda)(\delta - m - b)}{2\kappa\delta - A(\lambda)}, \tag{28}$$

$$w^*(b, \lambda) = m + \frac{\kappa\delta(\delta - m - b)}{2\kappa\delta - A(\lambda)}. \tag{29}$$

Equilibrium expert demand in periods 1 and 2 is

$$D_1^*(b, \lambda) = \frac{\kappa(\delta - m - b)}{2\kappa\delta - A(\lambda)}, \tag{30}$$

$$D_2^*(b, \lambda) = \left(1 + \frac{\lambda}{\delta}\right)\frac{\kappa(\delta - m - b)}{2\kappa\delta - A(\lambda)}. \tag{31}$$

The associated cutoffs are $t_1^* = 1 - D_1^*$ *and* $t_2^* = 1 - D_2^*$.

Proof. Differentiate the profit function with respect to w and q:

$$\frac{\partial \pi_E}{\partial w} = A(\lambda)\frac{q - 2w + m - b + \delta}{\delta}, \tag{32}$$

$$\frac{\partial \pi_E}{\partial q} = A(\lambda)\frac{w - m}{\delta} - \kappa q. \tag{33}$$

Setting the first-order conditions equal to zero gives

$$q - 2w + m - b + \delta = 0, \tag{34}$$

$$A(\lambda)(w - m) = \kappa\delta q. \tag{35}$$

Equation (35) implies

$$w = m + \frac{\kappa\delta}{A(\lambda)}q. \tag{36}$$

Substituting into (34) yields

$$q - 2\left(m + \frac{\kappa\delta}{A(\lambda)}q\right) + m - b + \delta = 0, \tag{37}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

which simplifies to

$$(2\kappa\delta - A(\lambda))q = A(\lambda)(\delta - m - b). \tag{38}$$

Therefore

$$q^*(b,\lambda) = \frac{A(\lambda)(\delta - m - b)}{2\kappa\delta - A(\lambda)}. \tag{39}$$

Substituting back gives

$$w^*(b,\lambda) = m + \frac{\kappa\delta(\delta - m - b)}{2\kappa\delta - A(\lambda)}. \tag{40}$$

Using Proposition 1,

$$D_1^*(b,\lambda) = \frac{q^* - w^* - b + \delta}{\delta}. \tag{41}$$

Substitute the formula for w^*:

$$D_1^*(b,\lambda) = \frac{q^*(1 - \kappa\delta/A(\lambda)) + \delta - m - b}{\delta}. \tag{42}$$

Since

$$\delta - m - b = \frac{2\kappa\delta - A(\lambda)}{A(\lambda)} q^*, \tag{43}$$

we obtain

$$D_1^*(b,\lambda) = \frac{\kappa\delta q^*/A(\lambda)}{\delta} = \frac{\kappa q^*}{A(\lambda)} = \frac{\kappa(\delta - m - b)}{2\kappa\delta - A(\lambda)}. \tag{44}$$

Proposition 1 then implies

$$D_2^*(b,\lambda) = \left(1 + \frac{\lambda}{\delta}\right) D_1^*(b,\lambda). \tag{45}$$

By Assumption 1,

$$0 < D_1^*(b,\lambda) = \frac{\kappa(\delta - m - b)}{2\kappa\delta - A(\lambda)} < \frac{1}{1 + \lambda/\delta} < 1, \tag{46}$$

so

$$0 < D_2^*(b,\lambda) = \left(1 + \frac{\lambda}{\delta}\right) D_1^*(b,\lambda) < 1. \tag{47}$$

Hence the candidate indeed lies in the interior region characterized in Proposition 1. It remains to verify uniqueness of the interior solution. The Hessian of π_E is

$$H = \begin{pmatrix} -2A(\lambda)/\delta & A(\lambda)/\delta \\ A(\lambda)/\delta & -\kappa \end{pmatrix}. \tag{48}$$

The leading principal minor is negative. The determinant is

$$\det(H) = \frac{A(\lambda)}{\delta^2}(2\kappa\delta - A(\lambda)) > 0 \tag{49}$$

by Assumption 1. Hence H is negative definite, so the objective is strictly concave on the interior branch and the first-order conditions characterize its unique maximizer there. Under Assumption 1, this interior solution is the equilibrium stated in the proposition. □

Proposition 2 shows that the equilibrium remains closed form despite the dynamic channel. The expert chooses a strictly positive markup,

$$w^*(b,\lambda) - m = \frac{\kappa\delta(\delta - m - b)}{2\kappa\delta - A(\lambda)}, \tag{50}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

so the platform compares the expert to a transfer price above real inference cost even when it is neutral. That is the familiar markup distortion. The novel dynamic element is that $A(\lambda)$ increases the return to expert demand, so stronger data feedback raises equilibrium quality and routed volume.

The next result formalizes the comparative statics of self-preferencing and data feedback.

Proposition 3 (Self-preferencing as a tax on outside expertise)**.** *Under Assumption 1, equilibrium quality, routed demand, and expert profit are all strictly decreasing in the platform bias parameter b. Specifically,*

$$\frac{\partial q^*}{\partial b} = -\frac{A(\lambda)}{2\kappa\delta - A(\lambda)} < 0, \tag{51}$$

$$\frac{\partial D_1^*}{\partial b} = -\frac{\kappa}{2\kappa\delta - A(\lambda)} < 0, \tag{52}$$

$$\frac{\partial D_2^*}{\partial b} = -\left(1 + \frac{\lambda}{\delta}\right)\frac{\kappa}{2\kappa\delta - A(\lambda)} < 0. \tag{53}$$

Equilibrium expert profit is

$$\pi_E^*(b, \lambda) = \frac{A(\lambda)\kappa(\delta - m - b)^2}{2(2\kappa\delta - A(\lambda))}, \tag{54}$$

which is strictly positive and strictly decreasing in b.

Proof. Differentiate the closed-form expressions in Proposition 2. Since $A(\lambda)$ does not depend on b,

$$\frac{\partial q^*}{\partial b} = -\frac{A(\lambda)}{2\kappa\delta - A(\lambda)} < 0, \tag{55}$$

which implies

$$\frac{\partial D_1^*}{\partial b} = -\frac{\kappa}{2\kappa\delta - A(\lambda)} < 0, \tag{56}$$

and, because $D_2^* = (1 + \lambda/\delta)D_1^*$,

$$\frac{\partial D_2^*}{\partial b} = -\left(1 + \frac{\lambda}{\delta}\right)\frac{\kappa}{2\kappa\delta - A(\lambda)} < 0. \tag{57}$$

To compute equilibrium profit, substitute Proposition 2 into

$$\pi_E = A(\lambda)(w - m)D_1 - \frac{\kappa}{2}q^2. \tag{58}$$

Using

$$w^* - m = \frac{\kappa\delta(\delta - m - b)}{2\kappa\delta - A(\lambda)}, \qquad D_1^* = \frac{\kappa(\delta - m - b)}{2\kappa\delta - A(\lambda)}, \tag{59}$$

and

$$q^* = \frac{A(\lambda)(\delta - m - b)}{2\kappa\delta - A(\lambda)}, \tag{60}$$

we obtain

$$\pi_E^*(b, \lambda) = A(\lambda)\frac{\kappa^2\delta(\delta - m - b)^2}{(2\kappa\delta - A(\lambda))^2} - \frac{\kappa}{2}\frac{A(\lambda)^2(\delta - m - b)^2}{(2\kappa\delta - A(\lambda))^2} \tag{61}$$

$$= \frac{A(\lambda)\kappa(\delta - m - b)^2}{2(2\kappa\delta - A(\lambda))} > 0, \tag{62}$$

where positivity follows from Assumption 1. Since the right-hand side is increasing in $(\delta - m - b)^2$ and $\delta - m - b > 0$, it is strictly decreasing in b. □

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

Proposition 3 is the central incidence result. A larger bias does not merely misroute a given stock of quality. It lowers the revenue base over which the outside expert can recover access costs and quality investment. Because routed traffic also improves future quality, the effect propagates intertemporally. In economic terms, self-preferencing taxes the outside expert's scale and thereby taxes outside capability accumulation.

The dynamic environment also lets us identify an amplification effect. Stronger data feedback increases the social and private return to expert demand, but it also makes bias more consequential because the same lost query now matters twice: once for current profit and again for future quality.

Corollary 1 (Data feedback amplifies the harm from bias)**.** *Under Assumption 1,*

$$\frac{\partial q^*}{\partial \lambda} > 0, \qquad \frac{\partial D_1^*}{\partial \lambda} > 0, \qquad \frac{\partial D_2^*}{\partial \lambda} > 0. \tag{63}$$

Moreover, the absolute sensitivity of expert quality and period-1 demand to platform bias is increasing in λ:

$$\frac{\partial}{\partial \lambda}\left|\frac{\partial q^*}{\partial b}\right| > 0, \qquad \frac{\partial}{\partial \lambda}\left|\frac{\partial D_1^*}{\partial b}\right| > 0. \tag{64}$$

Proof. Since $A'(\lambda) = \beta/\delta > 0$, differentiating Proposition 2 with respect to A gives

$$\frac{\partial q^*}{\partial A} = \frac{2\kappa\delta(\delta - m - b)}{(2\kappa\delta - A)^2} > 0, \tag{65}$$

so $\partial q^*/\partial\lambda = (\partial q^*/\partial A)A'(\lambda) > 0$. Next,

$$D_1^* = \frac{\kappa(\delta - m - b)}{2\kappa\delta - A(\lambda)}, \tag{66}$$

so

$$\frac{\partial D_1^*}{\partial \lambda} = \frac{\kappa(\delta - m - b)A'(\lambda)}{(2\kappa\delta - A(\lambda))^2} > 0. \tag{67}$$

Since $D_2^* = (1 + \lambda/\delta)D_1^*$ and both factors are increasing in λ, we also have $\partial D_2^*/\partial\lambda > 0$.

For the amplification claim,

$$\left|\frac{\partial q^*}{\partial b}\right| = \frac{A(\lambda)}{2\kappa\delta - A(\lambda)}. \tag{68}$$

Differentiating with respect to λ yields

$$\frac{\partial}{\partial \lambda}\left|\frac{\partial q^*}{\partial b}\right| = \frac{2\kappa\delta A'(\lambda)}{(2\kappa\delta - A(\lambda))^2} > 0. \tag{69}$$

Similarly,

$$\left|\frac{\partial D_1^*}{\partial b}\right| = \frac{\kappa}{2\kappa\delta - A(\lambda)}, \tag{70}$$

so

$$\frac{\partial}{\partial \lambda}\left|\frac{\partial D_1^*}{\partial b}\right| = \frac{\kappa A'(\lambda)}{(2\kappa\delta - A(\lambda))^2} > 0. \tag{71}$$

□

The corollary clarifies why AI routing is a particularly consequential setting for self-preferencing. In a static product-market model, bias changes current market share. In our environment, where traffic also creates future quality, stronger feedback magnifies the damage from the same amount of bias. This mechanism is exactly what one would expect in markets where access to queries, outcomes, and user interactions is a core input into capability development.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

4. Dynamic First Best and the Wedges in Decentralized Routing

We now compare decentralized routing to a planner benchmark. The goal is not to solve a full regulatory game, but to identify precisely which margins decentralized routing fails to internalize.

4.1. The Planner's Dynamic Benchmark

Fix the expert's initial quality level q. The planner chooses period-1 routing to maximize discounted total welfare, taking as given that period-2 expert quality is $q_2 = q + \lambda D_1$. Because user types are uniformly distributed and the expert's gross advantage is increasing in θ, it is sufficient for the planner to choose the mass $D_1 \in [0,1]$ of hardest period-1 queries routed to the expert.

If the planner routes the hardest D_1 mass of period-1 queries to the expert, the period-1 incremental welfare relative to using the incumbent for all queries is

$$S_1(D_1; q) = \int_{1-D_1}^{1} (q + \delta\theta - m)\, d\theta = D_1(q - m + \delta) - \frac{\delta}{2} D_1^2. \tag{72}$$

Given D_1, period-2 expert quality is $q + \lambda D_1$. In period 2, the planner again routes all and only those queries for which the expert's incremental welfare is nonnegative. On the interior branch where period-2 expert demand lies in $(0,1)$, optimal period-2 demand is

$$D_2^{FB}(q + \lambda D_1) = \frac{q + \lambda D_1 - m + \delta}{\delta}. \tag{73}$$

The corresponding period-2 incremental welfare on this branch is

$$S_2(q + \lambda D_1) = \frac{(q + \lambda D_1 - m + \delta)^2}{2\delta}. \tag{74}$$

Therefore the planner solves

$$\max_{D_1 \in [0,1]} \; S_1(D_1; q) + \beta S_2(q + \lambda D_1) - \frac{\kappa}{2} q^2. \tag{75}$$

Since q is fixed in this benchmark, the investment cost term does not affect the choice of D_1.

Proposition 4 (Conditional dynamic first best)**.** *Fix the expert's initial quality q. Under Assumption 1, suppose the planner's optimum is interior in period 1 and induces interior period-2 demand. Then the planner's unique interior optimum is*

$$D_1^{FB}(q) = \frac{(\delta + \beta\lambda)(q - m + \delta)}{\delta^2 - \beta\lambda^2}. \tag{76}$$

Equivalently, the planner's period-1 cutoff is

$$t_1^{FB}(q) = 1 - D_1^{FB}(q). \tag{77}$$

The planner's marginal condition is

$$q + \delta(1 - D_1^{FB}(q)) - m + \beta\lambda D_2^{FB}(q + \lambda D_1^{FB}(q)) = 0. \tag{78}$$

Proof. The planner's objective as a function of D_1 on this interior branch is

$$\Omega(D_1; q) = D_1(q - m + \delta) - \frac{\delta}{2} D_1^2 + \beta \frac{(q + \lambda D_1 - m + \delta)^2}{2\delta} - \frac{\kappa}{2} q^2. \tag{79}$$

Differentiate with respect to D_1:

$$\Omega_{D_1}(D_1; q) = q - m + \delta - \delta D_1 + \beta\lambda \frac{q + \lambda D_1 - m + \delta}{\delta}. \tag{80}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

The second derivative is

$$\Omega_{D_1D_1}(D_1;q) = -\delta + \frac{\beta\lambda^2}{\delta} = -\frac{\delta^2 - \beta\lambda^2}{\delta} < 0 \tag{81}$$

by Assumption 1, so the objective is strictly concave on the interior branch and the interior optimum is unique. Setting the first derivative equal to zero yields

$$(q - m + \delta)\left(1 + \frac{\beta\lambda}{\delta}\right) + D_1\left(\frac{\beta\lambda^2}{\delta} - \delta\right) = 0. \tag{82}$$

Solving for D_1 gives

$$D_1^{FB}(q) = \frac{(\delta + \beta\lambda)(q - m + \delta)}{\delta^2 - \beta\lambda^2}. \tag{83}$$

The cutoff form follows from the monotonicity of incremental welfare in θ. Finally, the first-order condition may be rewritten as

$$q + \delta(1 - D_1) - m + \beta\lambda\frac{q + \lambda D_1 - m + \delta}{\delta} = 0. \tag{84}$$

Since the fraction on the right is exactly $D_2^{FB}(q + \lambda D_1)$ on the interior branch, the marginal condition becomes

$$q + \delta(1 - D_1^{FB}) - m + \beta\lambda D_2^{FB}(q + \lambda D_1^{FB}) = 0. \tag{85}$$

□

The planner's condition has an intuitive interpretation. At the period-1 margin, routing one more borderline query to the expert has a current payoff equal to its current incremental welfare, $q + \delta\theta - m$. Unlike the platform, however, the planner also values the effect of that extra query on future expert quality. The continuation value is $\beta\lambda D_2^{FB}$ because one more period-1 expert query raises period-2 quality by λ, and a one-unit increase in expert quality raises the payoff on each period-2 expert-routed query by one. The planner therefore sets a lower period-1 cutoff than a static planner would on this interior branch.

Indeed, when $\lambda = 0$, Proposition 4 collapses to the static benchmark demand $(q - m + \delta)/\delta$. When $\lambda > 0$, the planner sends more traffic to the expert in period 1 because current routing also creates future capability. The next proposition shows how decentralized routing falls short.

4.2. Markup, Bias, and Data-Feedback Wedges

For any fixed (q, w, b), the platform's period-1 expert demand is

$$D_1^P(q, w, b) = \frac{q - w - b + \delta}{\delta}. \tag{86}$$

Comparing this expression to Proposition 4 yields the wedges between decentralized routing and efficient routing.

Proposition 5 (Three wedges in decentralized routing)**.** *Fix the expert's initial quality q and suppose the interior characterization in Proposition 4 applies and that platform demand is interior. Then the gap between the planner's period-1 expert demand and the platform's period-1 expert demand is*

$$D_1^{FB}(q) - D_1^P(q, w, b) = \frac{w + b - m}{\delta} + \frac{\beta\lambda(\delta + \lambda)(q - m + \delta)}{\delta(\delta^2 - \beta\lambda^2)}. \tag{87}$$

Hence decentralized routing under-routes to the expert for three distinct reasons:

(i) an access-markup wedge, $\frac{w-m}{\delta}$;
(ii) a bias wedge, $\frac{b}{\delta}$;

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

(iii) a data-feedback wedge, $\frac{\beta\lambda(\delta+\lambda)(q-m+\delta)}{\delta(\delta^2-\beta\lambda^2)}$.

All three wedges are nonnegative, and the data-feedback wedge is strictly positive whenever $\lambda > 0$.

Proof. Using Proposition 4,

$$D_1^{FB}(q) = \frac{(\delta+\beta\lambda)(q-m+\delta)}{\delta^2-\beta\lambda^2}. \tag{88}$$

Write this as

$$D_1^{FB}(q) = \frac{q-m+\delta}{\delta} + \left[\frac{\delta+\beta\lambda}{\delta^2-\beta\lambda^2} - \frac{1}{\delta}\right](q-m+\delta). \tag{89}$$

The bracketed term simplifies to

$$\frac{\beta\lambda(\delta+\lambda)}{\delta(\delta^2-\beta\lambda^2)}. \tag{90}$$

Hence

$$D_1^{FB}(q) = \frac{q-m+\delta}{\delta} + \frac{\beta\lambda(\delta+\lambda)(q-m+\delta)}{\delta(\delta^2-\beta\lambda^2)}. \tag{91}$$

Now subtract the platform's demand,

$$D_1^{P}(q,w,b) = \frac{q-w-b+\delta}{\delta}. \tag{92}$$

We obtain

$$D_1^{FB}(q) - D_1^{P}(q,w,b) = \left[\frac{q-m+\delta}{\delta} - \frac{q-w-b+\delta}{\delta}\right] + \frac{\beta\lambda(\delta+\lambda)(q-m+\delta)}{\delta(\delta^2-\beta\lambda^2)} \tag{93}$$

$$= \frac{w+b-m}{\delta} + \frac{\beta\lambda(\delta+\lambda)(q-m+\delta)}{\delta(\delta^2-\beta\lambda^2)}. \tag{94}$$

The first term may be decomposed into $(w-m)/\delta + b/\delta$. Under the maintained parameter restrictions and interior conditions, each term is nonnegative, and the second fraction is strictly positive whenever $\lambda > 0$. □

Proposition 5 is the paper's main welfare decomposition on the interior branch. The platform under-routes to the outside expert not only because it faces an access price above real inference cost and not only because it may self-preference the incumbent, but also because it fails to internalize the future quality gains that current expert traffic creates. The third wedge is specific to an AI-routing environment in which traffic allocation doubles as data allocation.

The decomposition immediately implies that conduct and pricing remedies are not sufficient on their own.

Corollary 2 (Neutrality and access pricing are complements but not enough)**.** *Fix the expert's initial quality q and suppose the interior benchmark in Proposition 5 applies. If regulation enforces neutrality* ($b = 0$) *and marginal-cost access pricing* ($w = m$)*, then the platform's period-1 expert demand becomes*

$$D_1^{MC}(q) = \frac{q-m+\delta}{\delta}, \tag{95}$$

which is still strictly below the dynamic first best whenever $\lambda > 0$*:*

$$D_1^{FB}(q) - D_1^{MC}(q) = \frac{\beta\lambda(\delta+\lambda)(q-m+\delta)}{\delta(\delta^2-\beta\lambda^2)} > 0. \tag{96}$$

Proof. Set $b = 0$ and $w = m$ in Proposition 5. The first two wedges vanish, leaving only the data-feedback wedge. Because $\lambda > 0$, $q - m + \delta > 0$ on the interior branch, and $\delta^2 - \beta\lambda^2 > 0$ under Assumption 1, the remaining wedge is strictly positive. □

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

This corollary provides a sharp reason why AI-routing governance cannot be reduced to a single nondiscrimination mandate within the interior benchmark. Neutrality corrects the platform's direct incentive to favor the incumbent. Marginal-cost access corrects the static markup distortion. But neither instrument makes the platform internalize how current expert traffic raises future expert capability. If the regulator also cares about dynamic efficiency, some instrument aimed at data sharing, learning access, or mandated experimentation remains relevant.

5. Policy Discussion, Empirical Predictions, and Extensions

The theoretical results carry several implications for the governance of AI gateways and other routing intermediaries.

5.1. Routing Neutrality and Dual-Role Platforms

The first implication is the most immediate. A platform that owns one of the candidate models has a structural incentive to overuse it. In the model, the distortion appears as the reduced-form bias parameter b. The interpretation is broader than overt ranking favoritism. The platform may make the incumbent look artificially cheap through internal transfer pricing, grant it better latency or context length, privilege it in hidden prompts, or favor it because routing inward preserves user data and ecosystem control. All of these mechanisms raise the effective platform-side value of keeping traffic in-house.

The model shows that the effects of such bias are broader in AI routing than in many standard search settings. Bias raises the period-1 threshold, but that is only the first-round distortion. It also lowers the scale on which the outside expert earns margins and, through that channel, reduces initial quality investment. Because period-1 traffic improves period-2 expert quality, the same initial diversion weakens future expert performance as well. In other words, self-preferencing reallocates demand and simultaneously depresses the rival's future capability frontier.

This result matters for competition policy. In a conventional product-market setting, a dominant intermediary may harm a rival by reducing current sales. In AI markets, reducing routed traffic may also deny the rival the data and learning opportunities embedded in those interactions. The harm from self-preferencing is therefore not exhausted by static market-share diversion. It may operate through capability accumulation, which is exactly the margin on which specialized entrants attempt to compete.

The natural policy counterpart is an auditable neutrality rule for routing. Such a rule need not require identical traffic shares across models. It need only require that routing decisions be justified by observable cost, latency, quality, or safety criteria rather than by ownership or hidden strategic preferences. In the language of the model, the objective is to reduce b, not to eliminate all asymmetry.

5.2. Access Pricing and Interoperability

The second implication is that neutrality alone cannot solve the allocation problem. Even when $b = 0$, the expert sets a strictly positive markup over real inference cost. The platform therefore compares the expert to a transfer price that exceeds the real resource cost of using the expert. Proposition 5 shows that this markup produces an under-routing wedge even in the absence of self-preferencing.

In actual AI markets, the analogs are numerous: wholesale API prices, platform commissions, revenue-sharing obligations, technical restrictions on external calls, minimum-spend commitments, and the inability of outside providers to interoperate on equal terms with the platform's in-house model. These are all ways in which the gateway can make third-party expertise costly to invoke.

The model therefore points to a familiar but important conclusion from platform economics: conduct remedies and price-structure remedies are complements. A policy that eliminates discrimination but leaves outside access overpriced corrects only part of the problem. A policy that forces better access terms but allows the platform to tilt routing toward its own model also corrects only part of the problem. The relevant object is the joint design of the routing rule and the access contract.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

One can see this directly from Proposition 5. The platform's period-1 under-routing equals the sum of three nonnegative wedges. A neutrality rule removes only b/δ. A cost-based access rule removes only $(w - m)/\delta$. The remaining distortions survive unless each wedge is addressed on its own terms.

5.3. Data Governance and the Dynamic Wedge

The third implication is specific to AI routing. Even if a regulator perfectly eliminated self-preferencing and forced marginal-cost access, decentralized routing would still remain inefficient because the platform does not internalize the future value of expert learning. This is the data-feedback wedge isolated in Corollary 2.

That result suggests that AI governance should pay attention not only to current referral neutrality but also to the allocation of learning opportunities. In practice, those opportunities may depend on access to labeled outcomes, failure cases, side-by-side comparisons, or user-feedback logs. If the integrated platform controls those assets, it can preserve a dynamic advantage for its own model even under formally neutral routing.

Three types of policy instrument emerge naturally.

First, the regulator may require minimum auditability of routing logs and outcomes. If outside experts can document where they were bypassed and how they would have performed, the platform's control over learning opportunities is reduced.

Second, the regulator may support data portability or data-sharing obligations in narrow, contestable forms. The objective is not unconditional diffusion of all data. It is to prevent the routing intermediary from using traffic control to monopolize the information generated by user interactions. The broader literature on data governance argues that data-sharing institutions can matter materially for innovation and competition in digital markets [16]. Our model gives a simple AI-routing rationale for such instruments.

Third, the regulator may in some contexts mandate limited outward referral or comparative testing on borderline queries. In our framework, the planner values additional expert referral because it raises future expert quality. A practical analog is a rule that requires periodic benchmarking, contested escalation rights, or a minimum external referral share for hard or high-stakes tasks.

The key point is conceptual. In many AI deployments, routing is not only a procurement decision. It is also a learning-allocation decision. Once that is recognized, the design of the router acquires a dynamic industrial-organization dimension that static neutrality rules alone cannot capture.

5.4. High-Stakes Domains and Escalation Design

The results are especially relevant in regulated or high-stakes environments. Consider legal, medical, financial, or safety-critical AI systems in which complex cases are exactly the ones that should be escalated to specialized experts or human review. If the platform self-preferences a cheap in-house model, or if it faces distorted access terms for the stronger outside system, it will set the escalation threshold too high. The resulting errors are concentrated among difficult cases, which are often precisely the cases with the largest downside risk.

The model suggests that such harms may be understated by average-performance metrics. A platform can satisfy broad performance targets while still routing too many borderline or difficult queries to the weaker in-house system. Sectoral governance may therefore need to focus on triage architecture: which cases are escalated, under what conditions, and with what audit trail. In the language of the model, regulation should attend to the cutoff itself, not only to unconditional model quality.

5.5. Empirical Predictions

Although the paper is theoretical, it yields several empirical predictions.

First, holding observable cost and quality constant, ownership links between the router and a candidate model should increase the probability that the owned model is selected. This ownership

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

effect should be largest on intermediate-difficulty queries, where the platform is closest to indifferent and therefore has the greatest room to steer traffic.

Second, platform decisions that worsen outside access terms—higher commissions, weaker interoperability, slower latency guarantees, or more restrictive API conditions—should lower not only current third-party routing shares, but also subsequent third-party quality investment or performance improvements. The dynamic margin is a core prediction of the model and distinguishes it from a purely static diversion story.

Third, the damage from self-preferencing should be greatest in market segments where traffic is particularly important for capability accumulation. In our notation, those are environments with a larger λ. Empirically, one would expect stronger adverse effects in domains where user feedback is rich, labels arrive quickly, or domain adaptation depends heavily on active deployment.

Fourth, if regulators improve data portability, auditability, or contestability at the routing layer, one should observe not only more outward referrals but also greater responsiveness of referrals to measured relative performance. In the model, better governance reduces the non-performance wedges and increases the alignment between routing and comparative quality.

These predictions suggest a natural empirical agenda. The relevant evidence is not confined to cross-sectional traffic shares. It also includes how routing reforms affect subsequent investment, performance growth, and entry by specialized experts. For AI gateways, the dynamic response of outside capability may be the most informative outcome variable.

5.6. Extensions

The model can be extended in several directions without changing its core logic.

A first extension is many-expert routing. Suppose the platform can route across multiple outside experts with different (q_j, γ_j, w_j) tuples. Under single crossing, the query space would be partitioned into intervals. Self-preferencing would then shift several routing boundaries rather than one. The central force would remain unchanged: a dual-role platform would distort the allocation of both demand and learning opportunities across specialists.

A second extension is user-side screening. Some users may be able to pay for escalation, wait longer for better models, or strategically rephrase queries to obtain stronger service. Embedding such behavior would connect AI routing to classical screening and referral design. The platform could then distort not only invisible backend routing, but also the user-facing menu that governs access to outside experts.

A third extension is endogenous entry. In the current model, the presence of an outside expert is taken as given. Yet Proposition 3 already implies the basic entry logic. A larger self-preferencing wedge compresses the revenue base and the learning base available to outsiders. In a richer environment, that would reduce entry into specialized capabilities. The router would then affect not only how demand is split among existing models, but also whether such specialists appear at all.

A fourth extension is richer data governance. Instead of the simple law of motion $q_2 = q + \lambda D_1$, one could allow some fraction of all queries to generate portable data, or allow the platform to choose whether outcomes are disclosed to the expert. Those choices would make data governance an explicit control variable. The present framework already indicates what such an extension would deliver: withholding portable information would become another way for the integrated intermediary to enlarge the dynamic wedge.

These extensions all point in the same direction. Once the routing layer is treated as a strategic intermediary, questions commonly described as AI-governance problems can often be restated as industrial-organization problems about platform conduct, access, and investment incentives.

6. Conclusion

AI deployment increasingly depends on routers, gateways, and assistants that decide which model answers which query. This paper argues that such routing should be analyzed as an industrial-organization problem rather than as a purely algorithmic choice rule. A routing intermediary allocates

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

traffic across models. If it owns one of those models, routing becomes endogenous to self-preferencing. If outside expertise is priced through access contracts and improves through routed traffic, then the platform simultaneously allocates current demand and future capability.

We build a tractable model of that environment. The platform routes queries between an in-house incumbent and an outside expert. The expert chooses an access price and a quality investment. Period-1 expert demand raises period-2 expert quality through data feedback. In equilibrium, routing is a cutoff rule in each period. Self-preferencing raises the cutoff, lowers routed demand, and depresses quality investment. Stronger data feedback increases the value of outside traffic, but it also amplifies the harm from bias because diverted demand now carries a dynamic cost.

Relative to a conditional dynamic first best, decentralized routing features three wedges: an access-markup wedge, a bias wedge, and a data-feedback wedge. The first two are familiar from platform economics. The third is distinctive to AI routing because traffic allocation is also learning allocation. For that reason, neutrality rules and access-pricing rules are complements, but even both together are not enough when the platform does not internalize the future value of outside learning.

The broader lesson is simple. In AI markets, the question is often not merely which model is best. It is who controls the router that decides which model is used, under what commercial terms, and with what consequences for future competition. Once that control layer is made explicit, the economics of platforms, biased intermediation, innovation incentives, and data governance become central to the analysis of AI systems.

References

1. Chen, L.; Zaharia, M.; Zou, J. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *Transactions on Machine Learning Research* **2024**. Published in December 2024.
2. Ong, I.; Almahairi, A.; Wu, V.; Chiang, W.L.; Wu, T.; Gonzalez, J.E.; Kadous, M.W.; Stoica, I. RouteLLM: Learning to Route LLMs with Preference Data. In Proceedings of the Proceedings of the International Conference on Learning Representations, 2025.
3. Panda, P.; Magazine, R.; Devaguptapu, C.; Takemori, S.; Sharma, V. Adaptive LLM Routing under Budget Constraints. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, 2025, pp. 25297–25313. https://doi.org/10.18653/v1/2025.findings-emnlp.1301.
4. Rochet, J.C.; Tirole, J. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association* **2003**, *1*, 990–1029. https://doi.org/10.1162/154247603322493212.
5. Armstrong, M. Competition in Two-Sided Markets. *The RAND Journal of Economics* **2006**, *37*, 668–691. https://doi.org/10.1111/j.1756-2171.2006.tb00037.x.
6. Jullien, B.; Sand-Zantman, W. The Economics of Platforms: A Theory Guide for Competition Policy. *Information Economics and Policy* **2021**, *54*, 100880. https://doi.org/10.1016/j.infoecopol.2020.100880.
7. Hagiu, A.; Jullien, B. Why Do Intermediaries Divert Search? *The RAND Journal of Economics* **2011**, *42*, 337–362. https://doi.org/10.1111/j.1756-2171.2011.00136.x.
8. de Cornière, A.; Taylor, G. A Model of Biased Intermediation. *The RAND Journal of Economics* **2019**, *50*, 854–882. https://doi.org/10.1111/1756-2171.12298.
9. Hagiu, A.; Teh, T.H.; Wright, J. Should Platforms Be Allowed to Sell on Their Own Marketplaces? *The RAND Journal of Economics* **2022**, *53*, 297–327. https://doi.org/10.1111/1756-2171.12408.
10. Etro, F. e-Commerce Platforms and Self-Preferencing. *Journal of Economic Surveys* **2024**, *38*, 1516–1543. https://doi.org/10.1111/joes.12594.
11. Athey, S.; Ellison, G. Position Auctions with Consumer Search. *The Quarterly Journal of Economics* **2011**, *126*, 1213–1270. https://doi.org/10.1093/qje/qjr028.
12. de Cornière, A. Search Advertising. *American Economic Journal: Microeconomics* **2016**, *8*, 156–188. https://doi.org/10.1257/mic.20130138.
13. Che, Y.K.; Hörner, J. Recommender Systems as Mechanisms for Social Learning. *The Quarterly Journal of Economics* **2018**, *133*, 871–925. https://doi.org/10.1093/qje/qjx044.
14. Arrow, K.J. The Economic Implications of Learning by Doing. *The Review of Economic Studies* **1962**, *29*, 155–173. https://doi.org/10.2307/2295952.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2026 doi:10.20944/preprints202603.1069.v1

15. Akcigit, U.; Liu, Q. The Role of Information in Innovation and Competition. *Journal of the European Economic Association* **2016**, *14*, 828–870. https://doi.org/10.1111/jeea.12153.
16. Graef, I.; Prüfer, J. Governance of Data Sharing: A Law & Economics Proposal. *Research Policy* **2021**, *50*, 104330. https://doi.org/10.1016/j.respol.2021.104330.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.