Substitution Across LLM API Products in the OpenRouter Market

David Lisk
Advisor: Felix Montag
April 21, 2026

Abstract

Using weekly OpenRouter data from 2025-08-24 to 2026-02-22, this study estimates LLM demand and substitution in the OpenRouter LLM API market. OpenRouter is a useful setting because many competing models are offered through a common API environment, with directly comparable prices and relatively low technical switching frictions. The main empirical challenge is that posted prices at the individual-model level barely move over time, making model-level demand estimation with product fixed effects weakly identified. To address that problem, the analysis aggregates the data to provider × bucket composite products and estimates a nested-logit IV demand model with OSS/proprietary nests.

The main result is that substitution is much stronger within the OSS/proprietary partition than across it. Averaged across the 10 products with the largest full-sample token volume over all 27 weeks, the mean within-nest cross-elasticity is about 0.126, while the mean across-nest cross-elasticity is about 0.008, implying a within/across ratio of about 14.9. The preferred model also implies an average own-price elasticity of about -0.78, indicating relatively modest own-price sensitivity at the composite-product level. This within-versus-across substitution result remains stable under the two main robustness checks: nearby outside-good calibrations around the baseline and a stable-roster restriction. The claim is correspondingly specific: competition appears meaningfully segmented between OSS and proprietary models within the OpenRouter-routed market.

1 Introduction

Are LLM API products becoming commodities, or does meaningful differentiation still structure competition in this market? This study examines that question using observed weekly usage data from OpenRouter, a routed API platform that places many competing language models in a common interface. The setting is useful because developers can access many models through a single API environment, compare prices directly, and switch across providers with relatively low technical friction. At the same time, the OpenRouter market observed here is only one part of the broader LLM economy. The analysis therefore focuses

on substitution within the OpenRouter API market, not demand for all LLM inference everywhere.

This question matters because discussion about LLM competition has moved faster than direct evidence. Rapid price declines, the spread of open-source models, and falling inference costs have all encouraged the view that LLM access may be converging toward a commodity service. But lower prices alone do not imply undifferentiated competition. Recent work by Demirer et al. (2025) documents rapid growth, major price declines, and substantial heterogeneity across AI API markets, while also finding evidence consistent with meaningful differentiation rather than a single flat pool of models. The analysis here focuses on one narrower but important piece of that broader question: realized substitution across LLM products within a routed API marketplace. The empirical framework builds on standard differentiated-products demand tools from industrial organization, especially the logit and nested-logit approaches developed in Berry (1994), Berry et al. (1995), and Nevo (2000).

The main empirical challenge is that posted prices at the individual-model level barely move over time in the sample. Once product fixed effects are included to absorb persistent product-level demand shocks, very little within-model price variation remains to identify price sensitivity. The analysis therefore aggregates the data to provider $\times$ bucket composite products observed in weekly markets and estimates a nested-logit IV demand model with OSS/proprietary nests. This aggregation creates enough within-product price movement to make estimation feasible, but it also changes the object being estimated from demand for individual named models to demand for composite products. The outside option is calibrated rather than directly observed. That is standard in this class of models, since total market size is usually not observed directly; here, the calibration reflects the fact that OpenRouter captures only one channel of LLM inference consumption.

Empirically, substitution in this observed market is structured rather than flat. Under the preferred specification, users substitute much more strongly within the OSS/proprietary partition than across it. Averaged across the 10 products with the largest full-sample token volume over all 27 weeks, the mean within-nest cross-elasticity is about 0.126, while the mean across-nest cross-elasticity is about 0.008, implying a within/across ratio of about 14.9. The preferred model also implies an average own-price elasticity of about -0.78 at the composite-product level, indicating relatively modest own-price sensitivity for a provider $\times$ bucket product. That elasticity is more sensitive to the outside-good calibration than the within-versus-across substitution contrast: for example, changing the market-size multiplier from $M = 2.0$ to $M = 1.5$ moves the average own-price elasticity from about -0.78 to about -1.10, while leaving the within/across substitution ratio nearly unchanged. The main

substitution result also survives the stable-roster restriction that removes visible composition changes inside the aggregated products.

The contribution is empirical rather than methodological. The study applies standard IO demand tools to a new digital market in which competitive structure is economically important but still poorly studied. The claim is correspondingly specific: the estimates do not identify demand for individual named models in the full LLM market, but they do show that competition within the OpenRouter-routed market is meaningfully segmented along the OSS/proprietary boundary rather than spread broadly across the full product set. The rest of the study proceeds as follows. Section 2 describes the data, Section 3 explains product construction and aggregation, Section 4 presents the demand framework, Section 5 reports the results and robustness checks, and Sections 6 and 7 discuss interpretation and conclude.

2 Data

2.1 Market Setting

The empirical setting is text-to-text LLM usage within the OpenRouter-routed API market over the period 2025-08-24 to 2026-02-22. OpenRouter is a third-party AI gateway that gives developers access to many models through a common API interface. The observed usage therefore comes from one routed marketplace rather than the full market for LLM inference.

2.2 Dataset Construction

The raw usage and price data were collected directly from publicly available OpenRouter model activity pages. For every OpenRouter model whose input and output modalities included text, a scraper periodically recorded daily prompt-token, completion-token, and, when applicable, reasoning-token usage. The scraper also collected each model's posted pricing information, including input-token and output-token costs. Because the activity pages retain recent history rather than only the current day, the scraper did not need to run continuously to recover the full sample window.

These daily observations were then assembled into a weekly model-level panel. Daily usage was aggregated to weeks, and the working unit of observation became model $\times$ week. To keep the analysis focused on the economically meaningful part of the market, models were dropped if they did not enter the weekly top 30 by token usage at least once during the sample. This selection rule retains about 95% of observed tokens in the raw weekly data while removing a large number of very low-usage models.

To proxy for model quality, the panel merges in the LM Arena Overall Text ELO score for each model. LM Arena is a public evaluation platform that ranks models using human preference comparisons over text responses rather than a single static benchmark. I collected the overall text leaderboard at the end of the sample period and assigned one score to each model in the panel. In the resulting dataset, ELO is time-invariant within model and should be interpreted as a cross-sectional quality proxy rather than a time-varying measure of quality.

Because model names do not always line up perfectly across OpenRouter and LM Arena, aliases were reconciled through a manually curated model map before the two sources were merged. That same mapping step was also used to assign the categorical labels needed later in the analysis. Each model was assigned a provider label, an open-source versus proprietary indicator, a generation-status label distinguishing current from legacy models, and a researcher-defined bucket label. The bucket label places models into Flagship, Balanced, or Value categories and is used below to form the baseline aggregated products.

The resulting raw weekly panel is the starting point for the estimation pipeline. Section 3 explains the additional filtering and aggregation used to move from the raw model panel to the final estimation dataset.

2.3 Sample Overview and Descriptive Statistics

The raw weekly panel covers 53 models from 11 providers over 27 weeks, from 2025-08-24 to 2026-02-22. The final estimation sample aggregates these data to 20 provider-bucket products and contains 456 product-week observations. Usage is concentrated across providers: Google, xAI, and Anthropic are the three largest providers by routed token volume, and the five largest providers—Google, xAI, Anthropic, OpenAI, and DeepSeek—together account for 87.1% of observed tokens in the raw weekly panel.

Prices vary widely across models. In the raw weekly panel, blended price ranges from \$0.025 to \$45.0 per million tokens, with a median of \$0.95. LM Arena Overall Text ELO ranges from 1211 to 1486, with a median of 1411. In the merged panel, ELO varies across models but not within model over time.

The market also changes rapidly over the sample. Weekly observed tokens rise from 3.23 trillion in the first week to 9.89 trillion in the last, a 3.06× increase. Over the same period, the number of active models rises from 31 to 51. The sample therefore describes a fast-growing market with substantial entry rather than a stable mature panel.

Table 1: Descriptive statistics for raw and estimation panels

Statistic	Raw weekly panel	Estimation panel
Unit of observation	model × week	provider × bucket × week
Observations	1,431	456
Models / products	53 models	20 products
Providers	11	11
Weeks	27	27
Sample window	2025-08-24 to 2026-02-22	2025-08-24 to 2026-02-22
Total tokens	154.5T	154.0T
OSS token share	25.0%	21.3%
Median blended price	0.95 USD/M tokens	0.84 USD/M tokens
Blended price range	0.025–45.00 USD/M tokens	0.025–45.00 USD/M tokens
Median ELO	1411	1402
ELO range	1211–1486	1211–1470

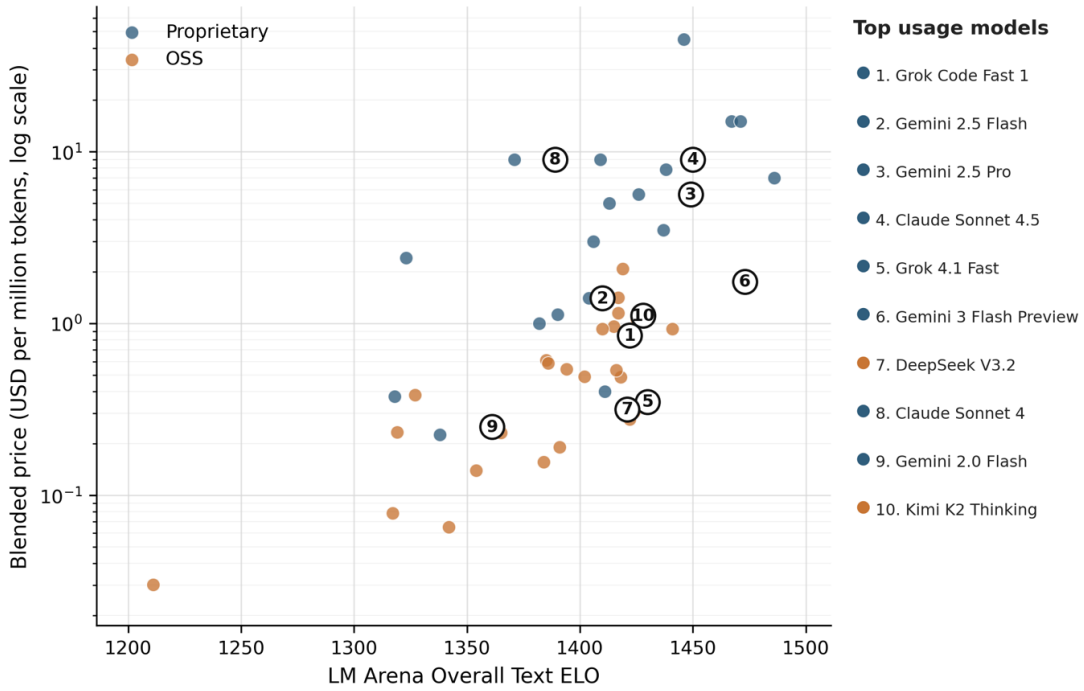

Figure 1: Price-quality scatter, colored by OSS/proprietary status

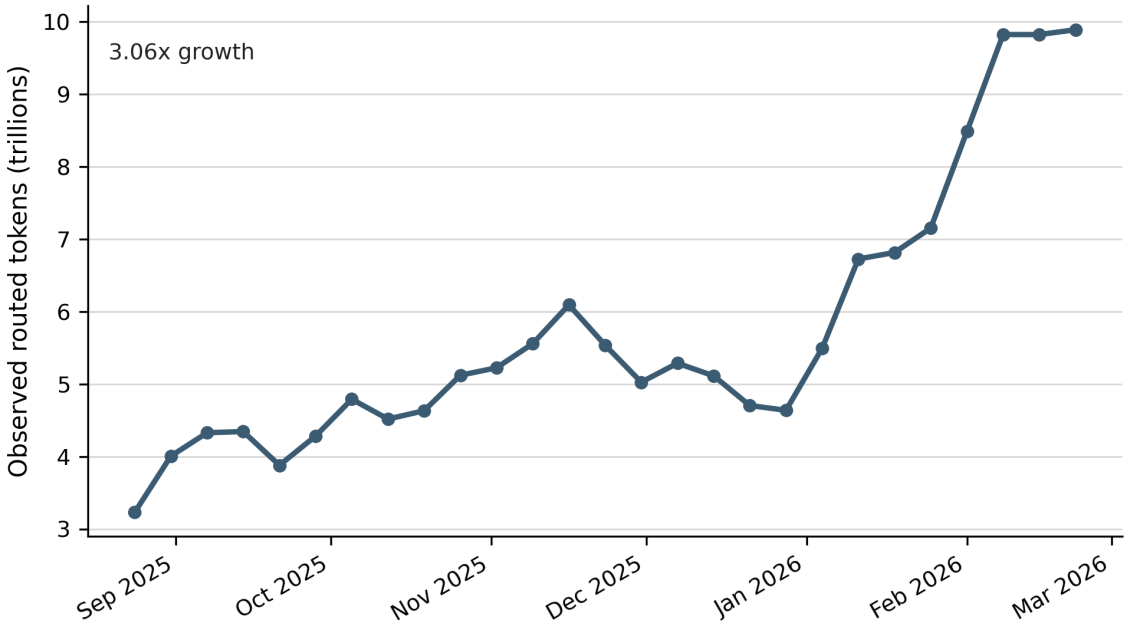

Figure 2: Weekly total routed tokens over time

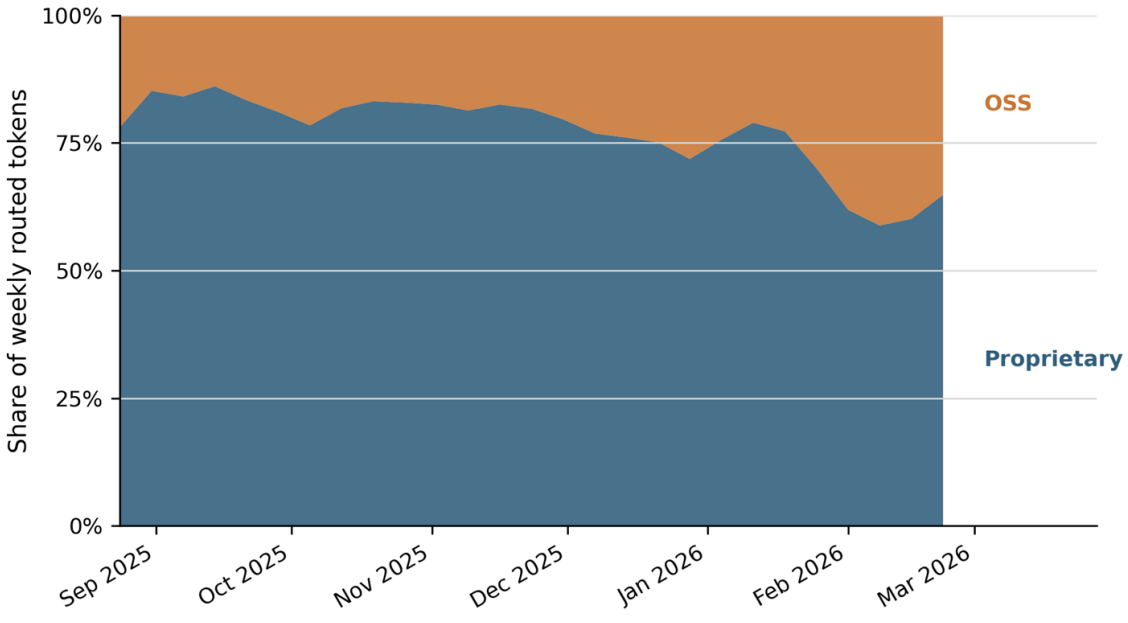

Figure 3: Market share of OSS vs. proprietary models over time

3 Product Construction and the Aggregation Tradeoff

The most direct demand product to analyze is the individual model. OpenRouter reports usage and price at the model level, so model-level demand estimation is the natural starting point. In practice, however, model-level prices barely change over time. Most models are introduced at a posted price that then remains fixed over the sample, so almost none of the observed price variation comes from within-model movement over time. In fact, the within-model share of total price variation is only about 0.02%. As a result, model-level demand estimation is too weak to support a credible estimate of price sensitivity.

For that reason, the analysis estimates demand on an aggregated panel rather than at the individual-model level. In the baseline specification, active model-week observations are aggregated into provider $\times$ bucket $\times$ week composites. This is a practical choice, not a claim that provider $\times$ bucket is the uniquely correct economic product definition. The point of the aggregation is narrower: it creates a coarser product definition in which price moves enough over time to make estimation feasible.

After aggregating to provider $\times$ bucket composites, the within-product share of total price variation rises from about 0.02% to about 23%. Put differently, a much larger share of observed price variation now comes from a product's own price moving over time rather than from fixed price differences across products.

Note. Formally, the within-product share is computed as

$$\frac{\sum_{jt}(p_{jt} - \bar{p}_j)^2}{\sum_{jt}(p_{jt} - \bar{p})^2},$$

where p_{jt} is the price of product j in week t, $\bar{p}_j$ is that product's average price over time, and $\bar{p}$ is the overall average price across all products and weeks.

This aggregation changes the object being estimated. The analysis no longer estimates demand for an individual model. Instead, it estimates demand for composite provider $\times$ bucket products observed on OpenRouter. A composite product's price can move because constituent model prices themselves change, because the active set of models inside the cell changes over time, or because users shift usage across models within the cell. For example, if users reallocate toward a cheaper model within the same cell, the token-weighted average price falls even if no individual model changes price. The resulting elasticity should therefore be interpreted as a composite-product elasticity rather than a model-level own-price elasticity.

Aggregation also creates an additional complication for interpretation. Because price is measured as a token-weighted average within each composite product, some observed price

movement can reflect changes in usage allocation inside the cell rather than only independent price changes. Price would be endogenous even with model-level data, so this is not the fundamental source of the identification problem. Still, in the aggregated panel, observed price changes can partly reflect composition shifts as well as underlying pricing decisions. The next section addresses these issues using instrumental variables.

4 Empirical Framework

4.1 Market Definition and Estimating Equation

Demand is estimated at the weekly product level, where each observation is a provider $\times$ bucket product in week t. The preferred specification uses a nested-logit framework because it allows substitution to be stronger within the two broad groups of open-source and proprietary models than across them. The estimating equation is:

$$\ln(s_{jt}) - \ln(s_{0t}) = \alpha \ln(p_{jt}) + \sigma \ln(s_{j|g,t}) + X_{jt}\beta + \delta_t + \xi_{jt}. \tag{1}$$

The dependent variable is the log share of product j relative to the outside option,

$$\ln(s_{jt}) - \ln(s_{0t}),$$

where s_{jt} is the observed market share of product j in week t, and s_{0t} is the outside share in that same week. Intuitively, the regression explains why one composite product gets more demand than another in a given week, relative to the option of not using one of the inside products. The log-share transformation is the standard dependent variable implied by the logit demand framework.

On the right-hand side of equation (1), p_{jt} is the product's token-weighted price, $s_{j|g,t}$ is its share within nest g, X_{jt} is a vector of observed product characteristics, δ_t denotes week fixed effects, and ξ_{jt} captures unobserved demand factors. In the preferred specification, the nests are OSS and proprietary.

The remaining terms have straightforward roles. The price term captures how demand changes when the token-weighted average price of a composite product changes. Price is log transformed because prices span a 1,800$\times$ range ($0.025 to $45.00 per million tokens), making a proportional price specification more natural than a linear one. The within-nest-share term allows products in the same nest to be closer substitutes than products in different nests. The control vector is composed of LM Arena ELO, open-source status, and whether the model is the most recent release in its series. Week fixed effects absorb shocks that are common to all

products in a week, such as overall market growth or other week-specific changes in platform usage.

4.2 Outside Option and Identification

OpenRouter usage only reveals demand for the inside goods, meaning the products observed on the platform. The model therefore requires a defined outside option. In this analysis, I set weekly market size equal to twice observed inside usage, so the implied outside share is 0.5 in every week. This is a calibrated choice rather than a known market fact. It provides a benchmark for estimation, and later sections test how sensitive the main results are to alternative market-size assumptions.

A central identification problem is that price is endogenous. Even at the individual-model level, observed prices may be correlated with unobserved demand shocks, which would make a simple regression of demand on price problematic. Aggregation adds an additional complication: because price is measured as a token-weighted average within each composite product, some observed price movement can also reflect changes in usage allocation inside the cell rather than only underlying posted price changes. The nested-logit term creates a similar issue, because a product's share within its nest is built from the same demand data the regression is trying to explain. For those reasons, the specification does not rely on a simple regression of demand on observed price and nest share alone.

Instead, the preferred specification uses instrumental variables. The instruments are designed to capture the competitive environment facing each product-week: the number of rival products in the market (`n_rivals`), the total ELO of the provider's other products (`sum_own_elo`), the number of rival products in the same nest (`n_nest`), and the total ELO of same-nest rivals (`sum_nest_rival_elo`). These variables help predict price and within-nest share. The identifying assumption is that, once observed product characteristics and week fixed effects are held constant, they affect demand only through those channels rather than through an additional direct effect on demand. In the preferred specification, the first-stage F-statistics are 180.9 for price and 74.9 for the within-nest-share term, indicating that the excluded instruments have substantial predictive power.

4.3 Scope and Limits of the Framework

Overall, this framework is designed to study substitution across composite LLM products within OpenRouter. More specifically, it asks whether substitution appears more concentrated within the OSS/proprietary split than across the market as a whole.

At the same time, the framework simplifies the market in several ways. The product definition is based on composite products rather than individual models, the nest structure is limited to two broad groups, and the outside option is calibrated rather than directly observed. The instrumental-variables strategy also assumes that the competitive-environment variables affect demand through price and within-nest position rather than through some additional direct channel. For those reasons, the model should be read as an empirical framework for this platform-specific market, not as a complete model of all LLM demand. Section 5 reports the main estimates and robustness checks, and Section 6 discusses their interpretation.

5 Main Results

5.1 Main Results Table

The empirical results progress from logit OLS, to logit IV, to nested-logit IV. Table 2 reports this progression. Logit OLS is a simple benchmark that treats price as exogenous. Logit IV addresses price endogeneity using instrumental variables, but it still keeps the standard multinomial-logit substitution pattern, in which consumers who switch away from one product reallocate toward other products roughly in proportion to those products' market shares. That pattern can be too restrictive when some products are much closer substitutes than others. Nested-logit IV is the preferred specification. It retains the IV correction and adds the OSS/proprietary nesting structure, which allows substitution to be stronger within those two broad groups than across them.

This progression is useful because each step changes what the model can explain. In the logit specifications, instrumenting price changes the estimated price coefficient and confirms that the endogeneity correction is doing real work. Moving from logit IV to nested-logit IV then changes the economic interpretation again by allowing for differential closeness of competition. In the preferred model, the price coefficient is smaller in magnitude than in the plain-logit specifications because some of the substitution pattern is now captured by the nesting structure rather than being forced onto price alone. The nesting parameter is also estimated comfortably inside the admissible range. The excluded instruments are strong in the preferred specification, with first-stage F-statistics of 180.9 for price and 74.9 for the within-nest-share term.

Table 2: Main demand-estimation results

	Logit OLS	Logit IV	Nested Logit IV
Alpha	-0.775	-0.534	-0.211
SE (alpha)	0.075	0.109	0.036
95% CI alpha	[-0.923, -0.628]	[-0.747, -0.321]	[-0.282, -0.140]
Sigma	–	–	0.78
SE (sigma)	–	–	0.046
95% CI sigma	–	–	[0.691, 0.869]
Avg OPE	-0.721	-0.496	-0.784
N	456	456	456
FS price	–	180.92	180.92
FS nest share	–	–	74.87

5.2 Preferred Specification and Parameter Estimates

Column 3 of Table 2 is the preferred specification: the nested-logit IV model with log price and OSS/proprietary nests. In this specification, the estimated price coefficient is $\alpha = -0.211$ with a robust standard error of 0.036 and a 95% confidence interval of [-0.282, -0.140]. The estimated nesting parameter is $\sigma = 0.780$ with a robust standard error of 0.046 and a 95% confidence interval of [0.691, 0.869]. The negative price coefficient implies that higher composite-product prices are associated with lower demand, while the positive nesting parameter is consistent with substitution being more concentrated within the OSS/proprietary partition than across it.

The preferred model also implies an average own-price elasticity of -0.784. This is a token-weighted average across the 456 provider $\times$ bucket $\times$ week observations in the estimation sample, so it should be interpreted as an average composite-product elasticity rather than an elasticity for any single named model. The next subsection turns to the central substitution result: the contrast between within-nest and across-nest substitution.

5.3 Within-Nest Versus Across-Nest Substitution

The central result is the contrast between within-nest and across-nest substitution under the preferred specification. Cross-elasticities are computed for the 10 products with the largest full-sample token volume and then averaged across the 27 weeks of data. The mean same-nest cross-elasticity is 0.126, while the mean different-nest cross-elasticity is 0.00845. This implies a within/across ratio of 14.91. In other words, when the price of one product changes, demand shifts much more strongly toward products in the same OSS/proprietary group than toward products in the other group.

The within/across contrast also remains large under bootstrap resampling. Using a 500-replication week-block bootstrap that re-estimates the preferred model in each replicate, the 95% percentile confidence interval for the within/across ratio is [10.29, 49.28]. The ratio therefore remains comfortably above 1 under resampling, which supports the conclusion that within-nest substitution is substantially stronger than across-nest substitution in this setting.

Within the OpenRouter-routed market and under the preferred composite-product specification, the OSS/proprietary split matters more for substitution patterns than for average own-price elasticity. Products compete much more strongly within that partition than across it.

5.4 Robustness

5.4.1 Outside-Good Sensitivity

The size of the outside option is one of the most important assumptions in the analysis, since OpenRouter only reveals demand for the inside goods. In the baseline specification, weekly market size is set to twice observed inside usage, which implies an outside share of 0.5. To test how much the main result depends on that calibration, the preferred specification is re-estimated under alternative values of M, the market-size multiplier.

Table 3: Outside-good sensitivity under alternative values of the market-size multiplier M

M	**Outside share**	α	σ	**Avg. own-price elasticity**	**Within/across ratio**	**Admissibility / note**
1.5	0.333	-0.242	0.8228	-1.104	13.243	Yes; close to baseline and still economically coherent
2.0	0.500	-0.211	0.7798	-0.784	13.448	Yes; canonical baseline
2.5	0.600	-0.119	0.9216	-1.207	52.451	Formally yes, but the ratio becomes unstable as σ approaches 1 and different-nest substitution shrinks toward zero
3.0	0.667	-0.029	1.0306	0.749	-175.857	No; $\sigma > 1$, so the nested-logit object is inadmissible and the ratio is not structurally meaningful

The main pattern is that the within-versus-across substitution result is reasonably stable near the baseline, while the exact elasticity levels are more sensitive. Moving from $M = 2.0$ to

$M = 1.5$ changes the average own-price elasticity from -0.784 to -1.104, while the within/across substitution ratio remains very similar, moving only from 13.448 to 13.243. This suggests that the qualitative substitution pattern is more robust than the exact magnitude of price sensitivity.

At larger values, however, the model becomes unstable. By $M = 2.5$, the nesting parameter rises toward 1 and the within/across ratio begins to blow up as different-nest substitution shrinks toward zero. At $M = 3.0$, the nesting parameter exceeds 1, making the model inadmissible. The takeaway is that $M = 2$ should be treated as a benchmark calibration rather than a deeply identified market-size parameter, and that the central substitution result is most credible in a local range around that baseline.

A plain-logit market-size sensitivity check gives a similar result. In that simpler model, the price coefficient remains negative and economically plausible across alternative values of M, but it is not perfectly stable. This reinforces the broader point that outside-good calibration matters more for exact elasticity magnitudes than for the study's main qualitative substitution result.

5.4.2 Stable-Roster Restriction

A natural concern is that the main result may be driven by changes in which models are active inside each provider × bucket product over time. To check that possibility, the preferred specification is re-estimated on a stable-roster sample. In this version, a product-week is kept only if the set of active models inside that composite matches the prior observed week for the same product.

Table 4: Preferred-specification estimates under the stable-roster restriction

Specification	**Product-weeks**	α	σ	**Avg. own-price elasticity**	**Within/across ratio**
Baseline preferred specification	456	-0.211	0.780	-0.784	13.448
Stable-roster restriction	419	-0.210	0.795	-0.837	14.311

The main estimates move very little under this restriction. The price coefficient remains essentially unchanged, the nesting parameter rises only slightly, and the average own-price elasticity becomes somewhat more negative. The within/across substitution ratio also remains large. These results suggest that the headline finding is not driven only by visible roster churn inside the aggregated products.

Taken together, these checks strengthen the study's main qualitative result. The within-versus-across substitution pattern survives both the outside-good sensitivity analysis near the baseline and the stable-roster restriction.

6 Discussion

The results have two implications. First, the OpenRouter market is not behaving like one flat pool of LLM products: substitution is much more concentrated within the OSS/proprietary split than across it. Second, the estimated own-price elasticity should be interpreted cautiously, because it is a composite-product elasticity and depends partly on how the outside good is calibrated. The discussion therefore separates the more robust substitution pattern from the more tentative interpretation of price sensitivity.

The preferred model implies an average own-price elasticity of -0.784. At the composite-product level, that is inelastic. One plausible interpretation is that labs are pricing below what a short-run static profit maximizer would choose. Instead, they may be willing to price below that level in order to attract users, grow revenue, generate more usage data, and build workflow lock-in while the market is still developing. Once developers build applications around a particular model stack, switching can become costly, especially when those workflows depend on prompts, tools, and surrounding infrastructure that have already been tuned to a given provider.

At the same time, the analysis does not separately identify these mechanisms, and the elasticity level is also shaped by the composite-product construction and the outside-good calibration. More broadly, the estimates should be read as evidence about competition among composite products within the OpenRouter-routed market, rather than as a complete model of demand for individual named models across the full LLM industry. The most defensible interpretation is therefore a modest but useful one: OpenRouter competition appears meaningfully segmented, while the exact level of price sensitivity should be treated as suggestive rather than definitive.

7 Conclusion

This study examines whether competition in the OpenRouter LLM API market is flat or meaningfully structured. Because model-level prices barely move over time, the analysis aggregates the data to provider $\times$ bucket composite products and estimates a nested-logit IV demand model with OSS/proprietary nests. The main result is that substitution is much stronger within the OSS/proprietary partition than across it. Averaged over the leading

products in the sample, the preferred model implies a within/across cross-elasticity ratio of about 14.9. The same model also implies an average own-price elasticity of about -0.78 at the composite-product level.

Taken together, these results suggest that this market is not behaving like one undifferentiated pool of LLM products. Within OpenRouter, products appear to compete much more strongly within the OSS/proprietary split than across it, even in an environment with relatively low technical switching frictions. At the same time, the claims are deliberately narrow. The estimates describe composite products within one routed platform rather than individual named models in the full LLM market. Even with those limits, the study provides direct evidence that realized substitution in this market is structured in a meaningful way, not simply spread across the full product set.

References

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

Steven T. Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2):242–262, 1994.

Mert Demirer, Andrey Fradkin, Nadav Tadelis, and Sida Peng. The emerging market for intelligence: Pricing, supply, and demand for llms. NBER Working Paper 34608, National Bureau of Economic Research, 2025. URL https://www.nber.org/papers/w34608.

Aviv Nevo. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548, 2000.