Measuring and Valuing Service Quality with Large Language Models

Bruno Abrahao Olalekan Bello Anton Bobrov

Xingji "Jax" Li James Traina

January 2026

Service quality is unmeasured in official statistics despite services comprising two-thirds of consumer spending. We develop a scalable text-based method to score millions of consumer reviews across five established service quality dimensions. In the cross-section, higher-quality establishments command higher prices: a one-standard-deviation quality increase corresponds to 9% higher spending. Yet aggregate quality declined 5 percentage points from 2019 to 2021, driven by deteriorating reliability and responsiveness—implying a quality-adjustment wedge of 0.2–0.4 percentage points annually that standard price indices miss. In financial markets, the quality-value relationship is more limited: quality is not broadly capitalized into firm valuations, though consumer-facing industries show a positive association. Our approach extends hedonic quality adjustment from goods to services while highlighting measurement challenges for both inflation statistics and asset pricing.

JEL Classification: E31, L81, D12, C81
Keywords: service quality, hedonic pricing, inflation measurement, text analysis, consumer reviews

Abrahao: NYU Shanghai/NYU Stern (abrahao@nyu.edu). Bello: NYU Stern (ob708@stern.nyu.edu). Bobrov: University of Michigan (bobrov@umich.edu). Li: NYU (xl2860@nyu.edu). Traina: NYU Stern Abu Dhabi (james.traina@nyu.edu). We thank colleagues for helpful comments.

1. Introduction

Services account for the majority of consumer spending, employment, and firm value in modern economies, yet their quality remains largely unmeasured in official statistics and corporate balance sheets. Unlike goods, whose prices are routinely adjusted for quality changes using hedonic methods, most services are included in inflation statistics without corrections for changes in consumer outcomes. This asymmetry creates two fundamental frictions. For macroeconomists, if service quality systematically improves or deteriorates, measured inflation and real wage growth are meaningfully biased. For financial markets, the inability to observe service quality obscures the actual stock of "intangible capital"—the brand equity and customer loyalty that drive future cash flows but appear nowhere in financial statements. Despite the centrality of services to the modern economy, there is no scalable method for measuring service quality itself.

This measurement gap is especially consequential during periods of economic disruption. In the post-pandemic period, service prices rose sharply as the service experience plausibly changed through multiple dimensions, including staffing levels, wait times, hours of operation, and available amenities. In such episodes, observed service inflation may conflate not only price changes but also shifts in quality. At the same time, these shifts represent real shocks to customer capital that accounting metrics, which track expenditures rather than asset accumulation, fail to capture.

Statistical agencies routinely apply hedonic methods to adjust for quality changes in goods, exploiting detailed information on observable product characteristics such as memory speed or horsepower (Rosen 1974; Triplett 2006). The Boskin (1996) Commission identified unmeasured quality change in goods as a primary source of bias in the Consumer Price Index. Extending this approach to services has proven difficult because service quality is inherently multidimensional and subjective. Attributes such as reliability,

responsiveness, and empathy (Parasuraman, Zeithaml, and Berry 1988) are central to the consumer experience but are difficult to observe consistently at scale.

The lack of observability creates a parallel challenge in corporate finance. A growing wedge between market values and book assets has been attributed to intangible capital—including brand equity, customer relationships, and organizational processes (Eisfeldt and Papanikolaou 2013). For service firms, the quality of the customer experience is a key driver of these intangibles, yet it remains unobserved. Standard empirical approaches proxy for intangible capital using accounting inputs such as Selling, General, and Administrative (SG&A) expenses (Peters and Taylor 2017). This input-based approach is inherently noisy: a firm that spends heavily to fix service failures appears identical to one that invests to prevent them. Consistent with this concern, Mizik and Jacobson (2008) show that perceptual brand attributes drive firm value beyond what accounting inputs capture.

This paper develops a scalable method for measuring service quality using online consumer reviews and studies its implications for both price measurement and firm valuation. We classify 7 million Yelp reviews across five dimensions of service quality–reliability, responsiveness, assurance, empathy, and tangibles–using modern NLP methods. We then merge these measures with transaction-level data from SafeGraph covering over 38,000 matched establishments, of which 8,330 have sufficient review coverage for our hedonic analysis. Across establishments, higher service quality is associated with higher consumer spending: a one standard deviation increase in quality corresponds to 8.7 percent higher spending (t = 13.3). This positive cross-sectional relationship is consistent with hedonic theory, in which consumers pay more for higher-quality services. Within establishments, quality changes are associated with modest spending adjustments (0.8 percent per standard deviation, t = 3.3), consistent with short-run price and behavioral stickiness. We return to the interpretation of this contrast in Section 4.2.

Turning to aggregate patterns, we find that quality declined 5.0 percentage points between 2019 and 2021, with the sharpest declines in reliability ("the business does what it

promises") and responsiveness ("staff are helpful and attentive"). These dimensions map directly to pandemic labor shortages: fewer workers means longer waits, more errors, and reduced attentiveness. Hotels and limited-service restaurants where service delivery is labor–intensive experienced the greatest deterioration, compared to retail categories such as grocery stores where product quality matters more than service encounter.

We interpret our hedonic coefficient as an upper bound on unmeasured inflation. The cross-sectional estimate of $\beta = 0.087$ reflects long-run equilibrium differences across establishments, while within-establishment adjustment is slower due to price and behavioral stickiness. The true inflation bias therefore lies between zero (if prices fully absorb quality changes) and the cross-sectional gradient (if they do not adjust at all). Our estimates of 0.2–0.4 percentage points annually are best understood as a plausible range rather than a point estimate, comparable in magnitude to the quality bias the Boskin (1996) Commission identified for goods.

Service quality is also a form of customer capital that should be reflected in valuations. Linking establishment-level quality to Compustat via brand name matching, we find the quality-valuation relationship varies with sample composition. When pooling all 143 public firms, the relationship is essentially zero ($\beta = 0.017$, $t = 0.59$). Restricting to consumer-facing industries yields a marginally positive coefficient ($\beta = 0.050$, $p = 0.05$). The quality-valuation relationship also shifted during the COVID crisis: the interaction of quality with the crisis period is marginally significant ($\beta = -0.049$, $p = 0.078$), indicating that the quality premium essentially disappeared when operational disruptions made service delivery difficult to maintain.

Our approach builds on two strands of the economics literature. Online reviews have transformed how consumers learn about quality (Anderson and Magruder 2012; Luca 2016), but this literature relies on aggregate star ratings—a single number that collapses multiple quality dimensions. Prior work demonstrates the value of disaggregating quality from text. Archak, Ghose, and Ipeirotis (2011) extract multi-attribute evaluations from

consumer electronics reviews and show that aspect-level quality measures explain price variation better than aggregate ratings. Lawani et al. (2019) document similar gains in the Airbnb setting, where sentiment-derived quality measures outperform star ratings. These findings, combined with the well-documented bimodality of online ratings (Hu, Zhang, and Pavlou 2009), motivate our approach: extracting *what* consumers say, not just how many stars they give. More broadly, text-as-data methods have enabled economists to measure otherwise unobservable constructs at scale (Gentzkow, Kelly, and Taddy 2019)—from policy uncertainty (Baker, Bloom, and Davis 2016) to firm-level political risk (Hassan et al. 2019). We apply these methods to measure service quality, constructing establishment-level scores that we use to estimate hedonic price adjustments and test whether perceived quality is reflected in firm values.

The remainder of the paper proceeds as follows. Section 2 presents our conceptual framework. Section 3 describes our data and measurement approach. Section 4 presents hedonic price results. Section 5 examines the relationship between quality and firm value. Section 6 concludes.

2. Conceptual Framework

2.1. Hedonic Price Adjustment

Statistical agencies routinely apply hedonic methods to adjust goods prices for quality changes, exploiting observable characteristics like processor speed or horsepower. The Bureau of Labor Statistics applies hedonic quality adjustment to 34 CPI entry-level items, of which only four are services (primarily telecommunications and cable), with the remainder in goods categories (U.S. Bureau of Labor Statistics 2025). Yet, services now account for over 60% of consumer expenditure. If service quality systematically improves or deteriorates while prices hold constant, measured inflation is biased.

Our approach demonstrates the feasibility of text-based quality measurement for services, though we view the resulting estimates as a research diagnostic rather than a blueprint for official statistics. Yelp reviewers skew younger and more urban than the general population, our categories cover only a subset of services, and quality dimensions are inherently subjective. These limitations reflect the central challenge facing quality adjustment in services rather than idiosyncrasies of our data.

Service quality is inherently multidimensional and subjective: attributes such as reliability, responsiveness, and empathy are difficult to observe and measure consistently. Text-based measures extracted from consumer reviews provide a way to operationalize these otherwise unobserved characteristics. Following Rosen (1974), we model the price of a service encounter as a function of its quality attributes:

$$\log(P_{it}) = \alpha + \beta \cdot \text{Quality}_{it} + \gamma \cdot X_{it} + \delta_t + \mu_i + \varepsilon_{it} \tag{1}$$

where P_{it} is the median transaction amount at establishment i in period t, Quality_{it} is our text-derived quality measure, X_{it} are controls for transaction composition, δ_t are time fixed effects, and μ_i are establishment fixed effects.

The coefficient β represents the implicit price of quality—consumers' marginal willingness to pay for better service. With β and time-series variation in quality, we can decompose observed price changes into inflation versus quality change. If quality deteriorates while nominal prices hold constant, the hedonic model registers this as hidden inflation.

The within-establishment specification (with μ_i) provides the key test: if prices adjust when quality changes at a given location, consumers are not systematically paying more for less. If prices are sticky while quality declines, consumers receive less value per dollar—inflation that standard indices miss. We estimate both specifications to distinguish the

cross-sectional relationship (do high-quality establishments have higher prices?) from the dynamic response (do prices move when quality changes?).

The two specifications identify different objects. The cross-sectional coefficient captures the long-run equilibrium relationship: consumers who want higher quality sort into establishments that provide it and charge accordingly. The within-establishment coefficient captures short-run adjustment: when quality changes at a given location, do prices respond? If the within-establishment coefficient is zero while the cross-sectional coefficient is positive, prices are sticky—quality can deteriorate without immediate price adjustment. The CPI-relevant inflation bias lies between these estimates. If prices eventually adjust to match the cross-sectional gradient, our estimate provides an upper bound on cumulative bias. If prices never adjust, the bias takes a different form: consumers receive less value per dollar indefinitely, but measured inflation is not technically wrong—it simply fails to capture the welfare loss from quality decline.

2.2. Service Quality and Firm Value

The customer capital literature models customer relationships as a state variable affecting firm value (Gourio and Rudanko 2014). Rust, Lemon, and Zeithaml (2004) formalize how service quality drives customer loyalty, which in turn generates future cash flows. Empirically, survey-based measures of customer satisfaction predict stock returns (Fornell et al. 2006), correlate with firm value (Luo, Homburg, and Wieseke 2010), and reduce cost of capital (Truong, Nguyen, and Huynh 2021). Service quality is a key antecedent of satisfaction and loyalty (Leisen Pollack 2009). Our text-derived quality measure proxies for this "customer capital"—an intangible asset that reduces demand elasticity and increases lifetime customer value (He, Mostrom, and Sufi 2024). If markets fully incorporate quality information, the relationship between quality and valuations should be contemporaneous. However, if markets are slow to process quality signals—particularly those not reported in

standard earnings releases—high-quality firms may report positive earnings surprises and earn abnormal returns. We test the first channel directly: whether quality is reflected in Tobin's Q after controlling for standard firm characteristics. Testing return predictability would require daily stock returns and factor portfolios beyond the scope of this paper.

2.3. Identification and Limitations

We are explicit about what our analysis can and cannot establish.

Hedonic coefficients capture correlations, not causal effects—unobserved establishment characteristics may drive both quality and pricing. We include establishment fixed effects to absorb time-invariant heterogeneity, but cannot rule out time-varying confounds such as management changes or local competition that affect both quality and prices.

We measure perceived quality among review-writers, not objective service delivery. Review-writers skew younger and more urban than the general population. Our quality measure reflects the experience of customers who choose to write reviews, which may differ from average customer experience.

We document correlations and predictive relationships. We do not estimate structural parameters of production functions or depreciation rates for customer capital. Our contribution is measurement and reduced-form evidence, not structural estimation.

Our training data derives from a single large language model (Gemini 2.5 Pro), which we use to generate initial quality labels that train our classifiers. We validate these labels against star ratings and test face validity on known service leaders, but do not test robustness to alternative models (e.g., GPT-4, Claude). Prior work suggests high agreement across models for structured labeling tasks, but model-specific biases could affect dimension scores in ways we cannot detect with single-model validation.

3. Data and Measurement

3.1. Data Sources

Our analysis combines three primary data sources: consumer reviews from Yelp, transaction-level spending from SafeGraph, and firm financials from CRSP/Compustat.

Yelp Academic Dataset. The Yelp Academic Dataset provides roughly 7 million reviews covering 150,000 establishments across 11 metropolitan areas in the United States and Canada from 2005–2022. Our analysis uses the 10 U.S. metros where we have SafeGraph spending coverage. Establishments concentrate in restaurants (52,000), retail (34,000), and personal services (11,000). The median establishment has 15 reviews, with substantial right-skew (mean of 45 reviews). Star ratings range from 1 to 5, with 46% of reviews awarding 5 stars and 15% awarding 1 star.

SafeGraph Spend. SafeGraph Spend provides monthly transaction aggregates—median transaction amount, total spend, transaction counts—for 1.6 million points of interest from January 2019 onward. SafeGraph aggregates anonymized credit and debit card transactions from a consumer panel covering establishments across all 50 states. To assess representativeness, we compare aggregate SafeGraph spending to Personal Consumption Expenditures on services from the Bureau of Economic Analysis. Year-over-year growth rates correlate at 0.88 for our matched sample (Figure 1), indicating that SafeGraph captures aggregate spending dynamics well.

CRSP/Compustat. For valuation tests, we match establishments to parent firms using CRSP/Compustat via WRDS. We focus on national chains (McDonald's, Starbucks, Marriott) where brand-to-ticker mapping is unambiguous. We apply standard screens, excluding financials (SIC 6000–6999) and utilities (SIC 4900–4999), while retaining franchise operations and REITs with consumer-facing businesses.

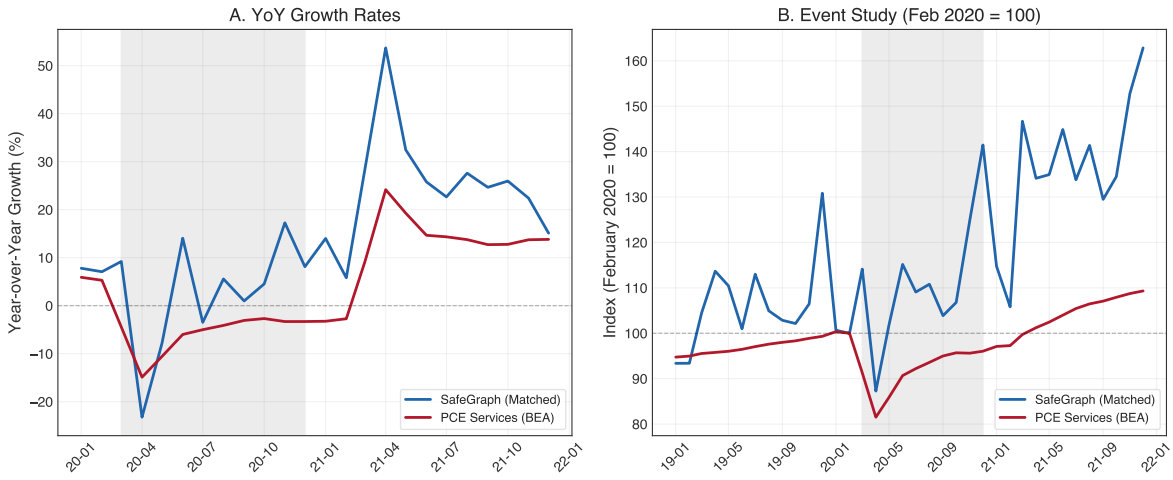


FIGURE 1. SafeGraph Spending Co-Moves with PCE Services

Notes: This figure validates SafeGraph as a high-frequency proxy for service consumption. Panel A compares year-over-year growth rates of SafeGraph spending (matched sample) and BEA Personal Consumption Expenditures on services; the correlation is 0.88. Panel B shows an event study indexed to February 2020, demonstrating that both series capture the COVID shock similarly. The matched sample includes SafeGraph establishments linked to our Yelp-derived quality measures.

3.2. Sample Construction

We link Yelp establishments to SafeGraph locations using ZIP-code blocking, spatial matching (100-meter tolerance), and name or brand similarity (Jaro-Winkler ≥ 0.80). Of 145,000 U.S. Yelp businesses with valid coordinates, our matching procedure yields 38,894 high-confidence matches, or 27% of eligible businesses (Table 1). Match quality is high: median geographic distance is 8 meters and median name similarity is 1.0 (exact match). For the hedonic sample, we construct an establishment-month panel with quality computed from a 6-month rolling window of reviews. We require at least 10 reviews in the rolling window to ensure stable quality estimates; results are qualitatively similar at thresholds of 5 and 20 reviews.

Sample periods differ across analyses due to data availability. Hedonic analysis covers 2019–2021, constrained by the overlap of SafeGraph (2019–2025) and our Yelp review sample (2019–2021 for matched establishments). Valuation analysis extends back to 2015–2021 using Yelp reviews merged with Compustat—sacrificing transaction data to maximize the time series for firm-level tests. Quality trend analysis extends to 2012–2024 using reviews alone, though early years have sparser coverage.

For sector-level analysis, we use SafeGraph's NAICS codes directly for matched establishments, aggregating to 4-digit codes except for restaurants where we preserve 6-digit codes to distinguish full-service (722511), limited-service (722513), and snack bars (722515).

Geographic Coverage. The Yelp Academic Dataset covers 11 metropolitan areas, 10 of which are in the United States: Philadelphia, Tampa, Nashville, Indianapolis, St. Louis, New Orleans, Tucson, Reno, Boise, and Santa Barbara. Our matched sample spans these metros across 13 states (because metros cross state lines) and 773 cities. All four Census regions are represented: South (35%), Northeast (31%), Midwest (17%), and West (17%). Three of our metros—Philadelphia, Tampa, and St. Louis—are among the 23 metropolitan areas in the BLS Consumer Price Index program. While not nationally representative, our

sample captures geographic variation across local service markets. We validate aggregate spending patterns against BEA Personal Consumption Expenditures (correlation 0.88, Figure 1), confirming that our sample tracks broader consumption dynamics.

TABLE 1. Summary Statistics

	Mean	Std. Dev.	Median	N
Panel A: SafeGraph Spending Data				
Median spend per transaction ($)	91.45	389.59	32.93	1,618,082
Transactions per month	116	468	13	1,618,082
Unique establishments		—		1,618,082
Panel B: Yelp Review Data				
Star rating (1–5)	3.62	1.00	4.0	119,698
Reviews per establishment	46.7	130.8	15	119,698
Unique establishments (open)		—		119,698
Panel C: Matched Sample				
Matched establishments		—		38,894
Match rate (Yelp, US)		26.9%		—
States / Cities		—		13 / 773

Notes: This table reports summary statistics for our primary data sources and the matched sample. Panel A summarizes SafeGraph spending data for 2019–2021. Panel B summarizes Yelp review data. Panel C describes the matched sample resulting from our geographic and name-based matching procedure.

TABLE 2. Sample Flow and Attrition

Stage	N	% of Prior	Selection Criterion
Reviews with quality scores	6,990,280	—	All reviews processed by NLP
Matched establishments	38,894	—	SafeGraph-Yelp spatial match
Hedonic panel (estab-months)	110,679	100%	≥10 reviews in 6-mo window
Valuation panel (firm-quarters)	2,419	3%	Public firm with Yelp match

Notes: This table documents sample attrition through the analysis pipeline. The hedonic panel is at the establishment-month level with quality computed from a 6-month rolling window of reviews; 8,330 unique establishments contribute 110,679 observations over 37 months. The valuation panel aggregates to firm-quarters for 143 public firms with identifiable brand-to-ticker mappings.

3.3. Measuring Service Quality

We measure service quality using a two-stage approach: large language model (LLM) labeling to generate training data, followed by efficient classifiers for full-corpus prediction. This design reflects a cost-accuracy tradeoff. LLMs can label reviews with high accuracy and without manual annotation, but applying them to millions of reviews is prohibitively expensive. SetFit classifiers, by contrast, achieve high accuracy from small labeled datasets and can process millions of reviews in hours on standard hardware (Tunstall et al. 2022). Combining these approaches—LLM labels for training, SetFit for inference—yields accurate measures at low cost, and the modular design allows researchers to substitute alternative LLMs or classifiers as these technologies evolve. We validate classifier performance in Section 3.7.

We use organic consumer reviews rather than structured surveys. Reviews offer three advantages for economic measurement: scale (7 million observations vs. expensive survey sampling), frequency (continuous vs. annual), and ecological validity (revealed preferences from actual customers vs. stated preferences from recruited respondents). The cost is representativeness: review-writers skew younger and more urban than the general population, and selection into review-writing may correlate with extreme experiences. We address these concerns in Section 3.7.

We ground our quality measures in the SERVQUAL framework (Parasuraman, Zeithaml, and Berry 1988), which identifies five dimensions of service quality: *Reliability* (does the firm deliver what it promised?), *Responsiveness* (do employees help promptly?), *Assurance* (do employees inspire trust and confidence?), *Empathy* (does the firm provide individualized attention?), and *Tangibles* (do facilities and equipment appear professional?). This framework, widely adopted in marketing and operations research, provides established theoretical grounding for our quality taxonomy.

We draw a stratified sample of 10,000 reviews balanced across 33 business categories and 5 star ratings (60 reviews per cell). For each review, we prompt Gemini 2.5 Pro to score each SERVQUAL dimension on a 5-point scale (1 = Very Negative, 3 = Neutral/Not Mentioned, 5 = Very Positive). The prompt includes verbatim SERVQUAL definitions from Kotler and Keller (2012) Table 13.5, instructs the model to score dimensions as neutral (3) if not mentioned, and to ignore non-service attributes such as price, location, and parking. We use temperature 0 for deterministic outputs and parse structured JSON responses. Of 10,000 reviews, 8,992 (90%) receive valid labels across all five dimensions.

We train five SetFit classifiers, one per SERVQUAL dimension (Tunstall et al. 2022). SetFit uses contrastive pre-training to learn accurate classifiers from small labeled datasets. We use `all-MiniLM-L6-v2` as the sentence encoder, representing each review as a 384-dimensional vector. The classification head uses ordinal logistic regression to respect the natural ordering of scores ($1 < 2 < 3 < 4 < 5$). We train on 8,000 reviews and evaluate on a held-out set of 992 reviews. The full-scale prediction applies these classifiers to all 7 million reviews using distributed GPU inference.

Each review receives five dimension scores (1–5). We aggregate to establishment-month using a 6-month rolling window. For each dimension, we compute the *mean-centered score*: the average score minus 3, divided by 2, ranging from −1 to +1 with 0 representing neutral. We require a minimum of 5 reviews in the rolling window; results are robust to thresholds of 10 and 20 reviews. For firm-level analysis, we aggregate establishment scores weighted by review count.

3.4. Econometric Framework: Time-Dummy Hedonic Model

In the construction of price indices for goods, statistical agencies typically rely on matched-model methods to track identical or closely comparable items over time. Services differ fundamentally in this respect. A service provided by a given establishment is a “continuing

good": while its quality may evolve, the underlying service and production unit persist, and there is no analogue of frequent model replacement as in consumer electronics. This feature allows us to work with a fixed panel of establishments without requiring explicit product matching. We therefore adopt a stratified time-dummy hedonic regression framework. Following the high-dimensional implementation of Ehrlich et al. (2023), this approach enables us to disentangle pure price inflation from changes in service quality and shifts in consumer expenditure behavior.

To isolate the implicit price of service quality, we define the service encounter as the unit of consumption. Unlike goods, where price is defined by a distinct unit, the price of a service experience is the total cost incurred to consume the bundle of amenities, food, and labor provided during the visit. Consequently, we treat the median transaction expenditure per transaction in establishment i as the effective price (P_{it}) of the service bundle.

Recent scholarship validates the use of these aggregates for tracking the service economy. Aladangady et al. (2019) demonstrate that credit card transaction samples track Census Bureau Retail Sales with high precision. For our identification, the decomposition of spending is critical. Chetty, Friedman, and Stepner (2024) show that volatility in aggregate service spending is driven primarily by the extensive margin (visit frequency) rather than the intensive margin (ticket size). This observed stability in the intensive margin supports our treatment of the "service encounter" as a fixed unit of consumption. Consequently, we interpret the growth in the median transaction size as price inflation rather than quantity expansion. We emphasize that the median transaction is an *equilibrium* price—reflecting both menu prices and consumer choices over bundle composition—rather than a posted menu price. Our hedonic regression therefore prices the realized consumption bundle. Adhering to the hedonic framework established by Rosen (1974), we follow the OECD Handbook on Hedonic Indexes (Triplett 2006) and for each category c estimate the following

semi-log regression.[1]

$$\ln(P_{it}) = \delta_{tc} + \sum_{k=1}^{5} \beta_{kc} S_{itk} + \gamma_c X_{it} + \mu_i + \epsilon_{it} \tag{2}$$

where $\ln(P_{it})$ is the natural log of the median spend per transaction at establishment i at time t; δ_{tc} is the category-specific time dummy, capturing the pure price inflation for category c after adjusting for quality changes; S_{itk} is the normalized score for the k-th SERVQUAL attribute; β_{kc} is the category-specific shadow price for attribute k; X_{it} represents transaction controls; and μ_i denotes establishment fixed effects. We estimate this specification both with and without μ_i to isolate the cross-sectional relationship (higher-quality establishments charge more) from the within-establishment dynamics (whether prices adjust when quality changes at a given location).

We exclude consumer demographics from the hedonic specification. In the Rosen (1974) framework, the hedonic function prices characteristics of the good, not characteristics of the buyer. Consumer demographics at an establishment reflect equilibrium sorting: higher-income consumers select into higher-quality establishments. Controlling for demographics would partial out the quality variation that generates this sorting, mechanically attenuating the estimated shadow prices of service attributes. We do, however, control for shifts in the mode of consumption that mechanically affect transaction size: (i) weekend intensity—the share of expenditure occurring Friday through Sunday—which captures differences between weekday "utilitarian" visits and weekend "leisure" consumption; and (ii) digital transaction share—the share of expenditure from online orders—which accounts for the distinct pricing of digital channels.

[1]We specify quality in levels rather than logs because our normalized scores range from −1 to +1 with zero representing neutral quality; the logarithmic transformation is undefined for non-positive values. This semi-log form is standard for hedonic regressions with bounded quality indices.

3.5. Index Construction

The coefficient $\hat{\delta}_{tc}$ provides an estimate of the logarithm of the category price index. We construct aggregate price indices in two stages: first estimating price indices for each category, and second aggregating these category-specific indices using time-varying consumption weights.

For each business category c, we estimate the stratified hedonic regression (Equation 2) to obtain the time-dummy coefficients $\hat{\delta}_{tc}$. The category-specific quality-adjusted price index in period t is defined as:

$$I_{tc}^{QA} = 100 \cdot \exp(\hat{\delta}_{tc}) \tag{3}$$

Note that we normalize $\hat{\delta}_{tc}$ such that $I_{0c}^{QA} = 100$. To construct the counterfactual Raw Price Index (I_{tc}^{Raw}), we estimate a constrained version of the hedonic model that omits the quality metrics S_{itk} while retaining establishment fixed effects and composition controls.

As derived in Triplett (2006), these indices decompose the raw price change:

$$\exp(\hat{\delta}_{tc}) = \frac{\text{Ratio of Matched-Model Geometric Means}}{\text{Total Characteristic Adjustment}} \tag{4}$$

The total adjustment factor implicit in our regression is:

$$\text{Adjustment}_{tc} = \exp\left[\underbrace{\sum_k \hat{\beta}_{kc}(\bar{S}_{ktc} - \bar{S}_{k0c})}_{\text{Quality Change}} + \underbrace{\sum_j \hat{\gamma}_{jc}(\bar{X}_{jtc} - \bar{X}_{j0c})}_{\text{Composition Shift}} \right] \tag{5}$$

To aggregate category-level indices into an economy-wide measure, we employ a chained Törnqvist index, allowing expenditure weights to vary over time. Let s_{ct} denote the share of total consumer spending allocated to category c in period t. The growth rate

of the aggregate quality-adjusted price index is given by:

$$\Delta \ln I_t^{QA} = \sum_{c=1}^{C} \frac{1}{2}(s_{ct} + s_{c,t-1})(\ln I_{tc}^{QA} - \ln I_{t-1,c}^{QA}) \tag{6}$$

with the index level constructed recursively as $I_t^{QA} = I_{t-1}^{QA} \exp(\Delta \ln I_t^{QA})$. An analogous procedure yields the aggregate Raw Price Index I_t^{Raw}.

We select the Chain Törnqvist index because it is a superlative index that corrects for substitution bias, and it serves as the exact discrete approximation to the Divisia index, allowing for a consistent aggregation of the logarithmic price changes derived from our semi-log hedonic specification.

3.6. Firm-Level Aggregation

To evaluate whether service quality correlates with firm valuation, we map our establishment-level quality measures to publicly traded firms. We link Yelp business names directly to Compustat company names using normalized string matching, supplemented by manual mappings for major franchise systems (e.g., YUM Brands subsidiaries Taco Bell, KFC, and Pizza Hut). To mitigate scope mismatch—where a conglomerate's non-service revenue is unobserved in our establishment data—we exclude financial firms (SIC 6000–6299) and require that matched firms operate primarily in consumer-facing sectors. Approximately 6% of Yelp establishments in our sample can be linked to a firm in the CRSP/Compustat universe, yielding 143 unique firms.[2] This linkage involves two steps: (1) collapsing the high-dimensional quality vector into a scalar metric of economic value per establishment, and (2) aggregating establishment scores to the firm level.

[2]Linkage requires (1) exact or near-exact name match between Yelp business name and Compustat company name, or (2) belonging to a known franchise system with manual parent company mapping.

We define the Establishment Quality Score ($\hat{S}_{it}$) as the linear index of attribute scores weighted by their implicit shadow prices:

$$\hat{S}_{it} = \sum_{k=1}^{K} \hat{\beta}_k S_{itk} \tag{7}$$

where S_{itk} is the score for dimension k and $\hat{\beta}_k$ is the estimated marginal willingness to pay from the hedonic regression. By projecting text-based attributes onto their marginal prices, $\hat{S}_{it}$ is in dollar-equivalent units—that is, predicted log expenditures attributable to service attributes.

We define firm-level quality ($\mathcal{S}_{ft}$) as the review-weighted average of the hedonic quality scores of all establishments belonging to firm f:

$$\mathcal{S}_{ft} = \sum_{i \in \mathcal{I}_f} \left(\frac{n_{it}}{\sum_{j \in \mathcal{I}_f} n_{jt}} \right) \times \hat{S}_{it} \tag{8}$$

where $\mathcal{I}_f$ is the set of all establishments linked to firm f, and n_{it} is the review count for establishment i. Weighting by review count gives more weight to establishments with more signal about quality.

3.7. Validation

We validate our service quality measures using three tests: face validity, incremental validity beyond star ratings, and linguistic analysis.

At minimum, establishments known for service excellence should score well above their category median. We verify that recognized service leaders—Chick-fil-A and In-N-Out for quick service restaurants, Ritz-Carlton and Four Seasons for hotels—rank in the top quartile of their respective categories.

Among quick-service restaurants, In-N-Out ranks at the 88th percentile and Chick-fil-A at the 72nd percentile. Among hotels, Four Seasons ranks at the 87th percentile and Ritz-

Carlton at the 77th percentile. All four brands score well above their category medians, consistent with their reputation for service excellence.

Figure 2 shows that LLM-labeled SERVQUAL scores increase monotonically with star ratings across all five dimensions. This monotonicity confirms that the LLM labeling captures the same underlying quality signal as star ratings while providing finer-grained dimensional information. Notably, Tangibles scores remain relatively flat across star ratings (2.7–3.4 range), suggesting that physical environment contributes less to overall satisfaction than service delivery dimensions.

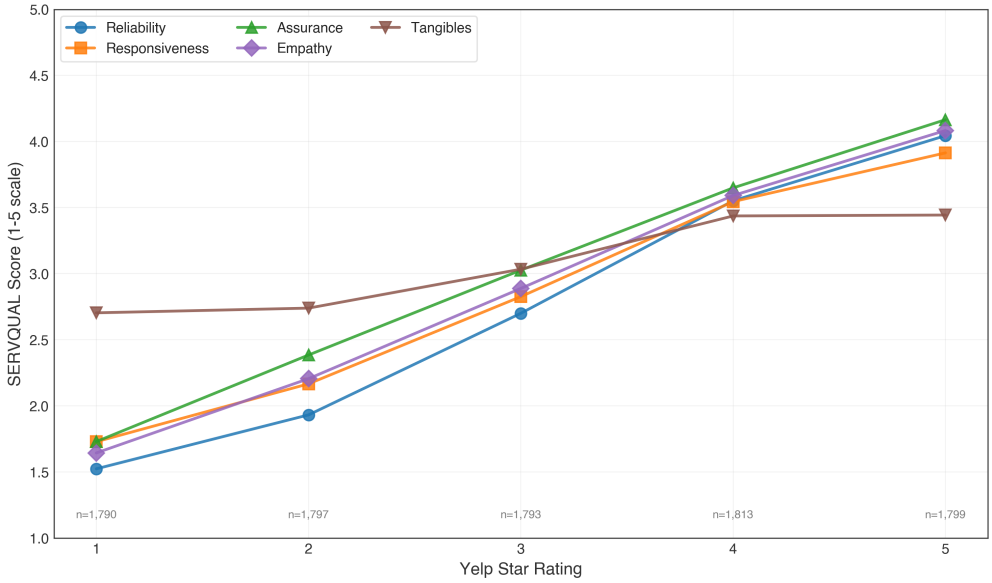


FIGURE 2. Mean SERVQUAL Scores by Yelp Star Rating

Notes: Mean LLM-labeled SERVQUAL scores (1–5 scale) by Yelp star rating. Sample includes 8,992 reviews with valid labels. All dimensions show monotonic increase with stars, confirming face validity. Tangibles shows the weakest relationship, consistent with physical environment being less salient in review text.

Star ratings collapse multiple quality dimensions into a single number. If our text-based measures capture meaningful variation, they should predict consumer behavior even after controlling for star ratings. We confirm this in Section 4.2, where text-derived quality predicts spending beyond what stars alone explain.

To assess semantic coherence, we compare the language used in reviews with extreme predicted scores on each dimension (Appendix Table A2). High-scoring reviews cluster

around staff behavior ("friendly staff", "quick service") and environment ("outdoor seating"). Low-scoring reviews cluster around complaints ("took forever", "place dirty"). The clustering indicates the model captures service-specific content rather than general sentiment. Table 3 reports classification accuracy and mean absolute error (MAE) for each SERVQUAL dimension on the held-out test set of 992 reviews. Accuracy ranges from 61% (Empathy) to 72% (Tangibles and Assurance), with MAE uniformly below 0.5 scale points. The ordinal regression head substantially outperforms standard classification, reducing MAE by exploiting the ordered structure of the 1–5 scale. Performance improves monotonically with training set size (Appendix Figure A7), with diminishing returns beyond 4,000 examples.

TABLE 3. SetFit Classifier Performance by Dimension

Dimension	Accuracy	Balanced Accuracy	MAE
Reliability	0.63	0.59	0.48
Responsiveness	0.70	0.64	0.38
Assurance	0.72	0.69	0.35
Empathy	0.61	0.58	0.48
Tangibles	0.72	0.58	0.33

Notes: Performance metrics on held-out test set (n=992). Training set size is 8,000 reviews. MAE is mean absolute error on the 1–5 scale. Balanced accuracy adjusts for class imbalance.

TABLE 4. Correlation Matrix of SERVQUAL Scores

	Stars	Reliability	Responsiveness	Assurance	Empathy	Tangibles
Stars	1.000					
Reliability	0.711	1.000				
Responsiveness	0.649	0.620	1.000			
Assurance	0.706	0.607	0.582	1.000		
Empathy	0.705	0.560	0.647	0.755	1.000	
Tangibles	0.355	0.232	0.187	0.225	0.204	1.000

Notes: Pairwise correlations computed from 8,992 LLM-labeled reviews. Stars is the Yelp star rating (1–5). SERVQUAL dimensions are scored 1–5 by Gemini 2.5 Pro.

The five SERVQUAL dimensions should be related but distinct. Cross-dimension correlations (Table 4) range from $r = 0.19$ (Tangibles–Responsiveness) to $r = 0.76$ (Empathy–Assurance)—moderate overlap consistent with related but not identical constructs. Tangibles shows the weakest correlations with other dimensions, consistent with physical environment being conceptually distinct from staff-driven service attributes.

4. Service Quality and Prices

4.1. Multi-Dimensional Quality and Spending Gradients

Service experiences are inherently multi-dimensional. A restaurant visit combines food quality, staff behavior, speed, and atmosphere; a hotel stay involves cleanliness, comfort, and front-desk service. If consumers value these dimensions differently and are willing to pay different premia for improvements along each, then understanding price dispersion and inflation requires measuring quality at the aspect level.

TABLE 5. Hedonic Shadow Prices by SERVQUAL Dimension

	Cross-Sectional		Within-Establishment	
SERVQUAL Dimension	$\hat{\beta}$	Std. Error	$\hat{\beta}$	Std. Error
Reliability	−0.036***	(0.009)	−0.002	(0.003)
Responsiveness	−0.001	(0.009)	0.004	(0.003)
Assurance	0.200***	(0.010)	0.008**	(0.003)
Empathy	−0.113***	(0.011)	−0.002	(0.003)
Tangibles	0.076***	(0.007)	0.003	(0.002)
NAICS × Month FE	Yes		Yes	
Establishment FE	No		Yes	
Observations	110,679		110,679	
Establishments	8,330		8,330	

Notes: This table reports hedonic regression coefficients from regressing log median spend on all five SERVQUAL dimensions jointly (standardized, winsorized at 1/99). Cross-sectional specification includes NAICS-by-month fixed effects. Within-establishment specification adds establishment fixed effects. Standard errors clustered at establishment level. *, **, *** denote significance at 10%, 5%, 1%.

Table 5 reports hedonic shadow prices—the marginal value consumers place on each quality dimension, holding the others constant. Assurance dominates: a one-standard-deviation increase in Assurance is associated with 20% higher spending (t = 19.1). Tangibles also commands a significant premium (7.6%, t = 10.3). In contrast, Empathy shows a negative gradient (–11%, t = –10.5), likely reflecting sorting: establishments emphasizing personalized attention may serve lower-spending customers. Reliability shows a modest negative coefficient (–3.6%, t = –4.2), while Responsiveness is near zero. Consumers primarily pay for trust and competence (Assurance) and physical environment (Tangibles). Assurance—employee knowledge, courtesy, and ability to inspire confidence—is the hardest dimension to replicate and the most costly to sustain at scale, consistent with its elevated shadow price. Within-establishment estimates are small, with only Assurance showing a significant positive coefficient (β = 0.008, t = 2.3).

4.2. Hedonic Price Regressions

We estimate the hedonic price regression specified in Equation 2, relating transaction size to service quality while controlling for transaction composition.

Following Chetty, Friedman, and Stepner (2024), who show that volatility in service spending is driven primarily by visit frequency rather than ticket size, we treat the median transaction as a stable unit of consumption. We include controls for weekend intensity (share of spending Friday–Sunday) and digital transaction share (online orders) to separate price changes from quantity or channel shifts.

Table 6 presents hedonic price regressions under two specifications. Column (1) includes NAICS and month fixed effects, isolating the cross-sectional relationship: do higher-quality establishments have higher transaction amounts? Column (2) adds establishment fixed effects, isolating within-establishment variation: when measured quality changes over time, do prices move? Figure 3 visualizes the cross-sectional relationship non-parametrically.

TABLE 6. Hedonic Price Regressions

	(1)	(2)
	$\ln(\text{Median Spend}_{it})$	
Quality Index (std)	0.087***	0.008***
	(0.007)	(0.003)
NAICS × Month FE	Yes	Yes
Establishment FE	No	Yes
Observations	110,679	110,679
Establishments	8,330	8,330

Notes: This table reports panel regressions of log median spend per transaction on the standardized quality index. All continuous variables winsorized at 1st and 99th percentiles. Column (1) includes NAICS-by-month fixed effects, identifying the cross-sectional relationship. Column (2) adds establishment fixed effects, identifying within-establishment variation. Standard errors clustered at the establishment level in parentheses. *, **, *** denote significance at 10%, 5%, 1% levels.

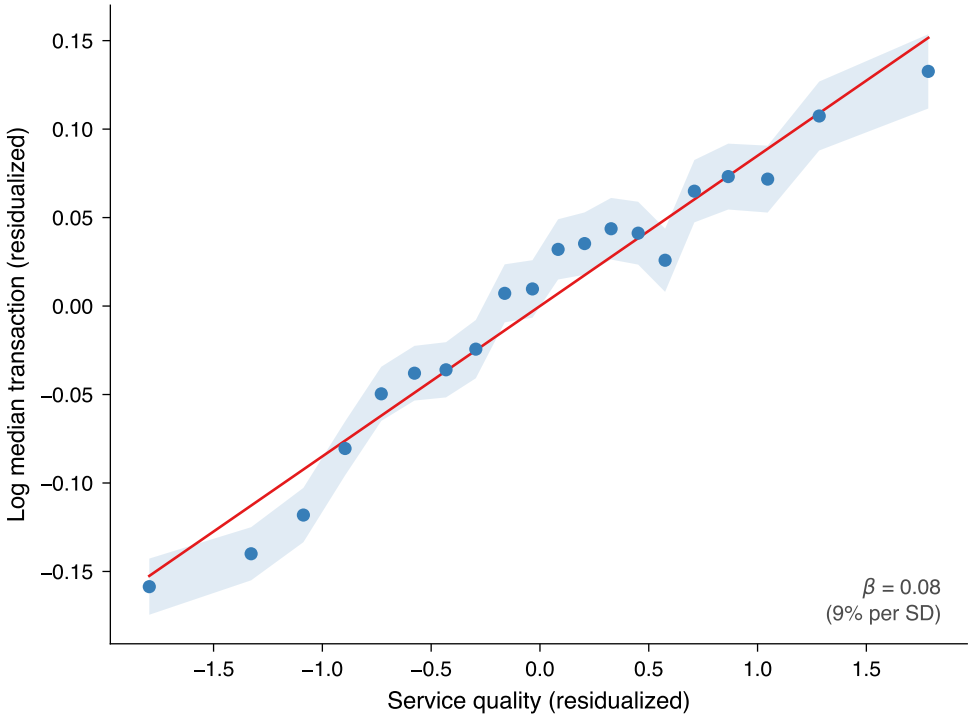


FIGURE 3. Service Quality and Consumer Spending

Notes: Binscatter of log median transaction on service quality, both residualized by NAICS and month fixed effects. Each point averages ~5,500 establishment-months. Shaded region: 95% CI. N = 110,679.

The cross-sectional estimates (Column 1) show that quality is associated with a substantial price premium. A one-standard-deviation increase in our quality index is associated with 8.7% higher median transaction amounts (t = 13.3). This is consistent with the basic premise underlying quality-adjusted price measurement: consumers value service quality, and the market prices it.

In contrast, the within-establishment estimates (Column 2) are economically smaller but statistically significant. The coefficient on quality is 0.008 (SE: 0.003, t = 3.3), indicating modest short-run adjustment of transaction amounts to changes in quality within establishments. The muted within-establishment response—an order of magnitude smaller than the cross-sectional gradient—is consistent with evidence that service prices adjust infrequently (Nakamura and Steinsson 2008; Bils and Klenow 2004; Klenow and Kryvtsov 2008) and that consumers exhibit inertia in their spending behavior (Handel 2013; Dubé, Hitsch, and Rossi 2010; Hortaçsu, Madanizadeh, and Puller 2017). Even sizable quality shocks, such as those experienced during the pandemic, may not translate immediately into proportionate price changes.

These two estimates—the cross-sectional hedonic gradient and the within-establishment response—capture different economic objects. The cross-sectional coefficient reflects long-run equilibrium differences across establishments, in which prices have adjusted and consumers have sorted in response to persistent quality differences. By contrast, the within-establishment coefficient reflects short-run adjustment, when prices and consumer behavior may be constrained by nominal rigidities and inertia. From the perspective of inflation measurement, neither estimate alone identifies the quality-adjustment bias in official price indices. Instead, they provide natural bounds. If prices fully and immediately adjust to quality changes, quality-adjusted inflation would coincide with measured inflation. At the other extreme, if prices do not adjust at all, the cross-sectional hedonic gradient provides an upper bound on the magnitude of unmeasured quality change embedded in observed price movements.

Our baseline uses 6-month rolling windows to aggregate quality scores. We verify that these results are robust to alternative window specifications. Using 3-month windows ($\beta = 0.097$, $N = 34{,}205$), 9-month windows ($\beta = 0.098$, $N = 136{,}587$), or 12-month windows ($\beta = 0.093$, $N = 185{,}299$), we find cross-sectional quality-price gradients that are stable across aggregation windows. The coefficient varies by less than 10% across specifications, suggesting that our baseline estimate is not sensitive to the particular review window chosen.

Our estimates therefore identify a range of plausible counterfactual quality adjustments rather than a point estimate of CPI bias. Reconciling the near-zero within-establishment response with the sizable cross-sectional valuation would require implausibly large attenuation from measurement error alone, suggesting that the contrast primarily reflects differences between short-run adjustment and long-run equilibrium pricing. In this sense, the relevant object for inflation measurement lies between these two estimates and is not point-identified. We use this logic in the next section to construct upper bounds on quality-adjusted service inflation during the post-pandemic period.

The aggregate coefficient masks substantial heterogeneity. Limited-service restaurants show the strongest quality-price gradient ($\beta = 0.15$), followed by full-service restaurants ($\beta = 0.09$). In contrast, snack bars, hotels, and grocery stores show no significant relationship. The source of this heterogeneity is not fully understood; it may reflect differences in within-category price dispersion, the salience of service quality to purchase decisions, or the representativeness of Yelp reviewers across business types.

For subsequent analysis, we construct the establishment quality score $\hat{S}_{it}$ defined in Equation 7, weighting dimension scores by the hedonic shadow prices reported in Table 5. Category-specific shadow prices are reported in Appendix Figure A3.

4.3. Quality Trends

Service quality declined over our sample period. Aggregate quality fell from −0.081 in 2019 to −0.130 in 2021, a decline of 5.0 percentage points on the mean-centered $[-1, +1]$ scale. The decline was not monotonic: quality dipped modestly during the COVID shutdown (2020Q2–Q4), but the largest deterioration occurred during the recovery period (2021Q3–Q4), consistent with staffing and supply constraints during the demand recovery.

By dimension, Reliability (−0.053) and Responsiveness (−0.043) declined most, suggesting deteriorating wait times and service consistency. Assurance and Empathy also fell. Tangibles—the only physical (non-labor) dimension—remained stable until late 2021. This pattern matches documented labor disruptions: the quits rate in accommodation and food services rose from 4.8% in 2019 to 6.9% in late 2021 (U.S. Bureau of Labor Statistics 2022), and 78% of restaurant operators reported understaffing (National Restaurant Association 2021).

The decline varied substantially across categories. Traveler Accommodation experienced the largest quality deterioration (−0.082), followed by Limited-Service Restaurants (−0.063) and Snack Bars (−0.049). Full-Service Restaurants showed a more modest decline (−0.020), while Grocery Stores were essentially stable (−0.013). This pattern is consistent with quality deteriorating most in service categories that are both highly labor-intensive and rely heavily on discretionary service margins—such as attentiveness, speed, and reliability—that are particularly sensitive to staffing constraints during the recovery.

Figure 4 visualizes the divergence between quality and spending. Panel A shows the monthly evolution of service quality, indexed to 100 in January 2019. Panel B shows consumer spending over the same period. The contrast is stark: by December 2021, the quality index had fallen to 90.4 (a 9.6 point decline from the 100 baseline) while spending had risen 21%. This divergence—consumers paying more for less—is the unmeasured inflation our hedonic framework quantifies.



FIGURE 4. Service Quality and Consumer Spending: The Divergence

Notes: Panel A plots monthly average service quality (SERVQUAL composite score), indexed to January 2019. Panel B plots mean transaction spending (winsorized at 5th/95th percentiles), indexed to January 2019. Both panels use simple (unweighted) means across establishments. Gray shading indicates the COVID period (March–December 2020). Sample: 110,679 establishment-months from hedonic panel.

4.4. Constant-Quality Price Indices

Combining hedonic coefficients with quality trends yields constant-quality price indices. We construct two series: (1) the *raw price index,* which tracks nominal transaction prices without quality adjustment, and (2) the *quality-adjusted index,* which holds service quality constant at its base-period level.

The quality-adjusted index I_{tc}^{QA} is constructed from hedonic time dummies as defined in Equation 3, and aggregated to an economy-wide measure using Törnqvist weights (Equation 6).

4.5. The Quality-Adjustment Wedge

We define the quality-adjustment wedge as the gap between measured prices and constant-quality prices:

$$\Delta_t = I_t^{Raw} - I_t^{QA} \tag{9}$$

A positive wedge indicates that standard inflation measures understate the true cost of living because consumers pay similar prices for degraded service.

Figure 5 presents our primary macroeconomic result: the comparison between the raw transaction price index and our quality-adjusted hedonic price index.

By late 2021, the quality-adjusted index sits approximately 0.6 percentage points above the raw index. This wedge reflects quality degradation: consumers paid roughly the same nominal prices but received measurably worse service. Annualized over the 2019–2021 period, this amounts to approximately 0.2 percentage points per year of quality-adjusted price growth beyond the raw index. This estimate relies on the cross-sectional hedonic coefficient; using an intermediate value between the cross-sectional estimate and the within-establishment estimate would yield smaller adjustments. The magnitude is smaller

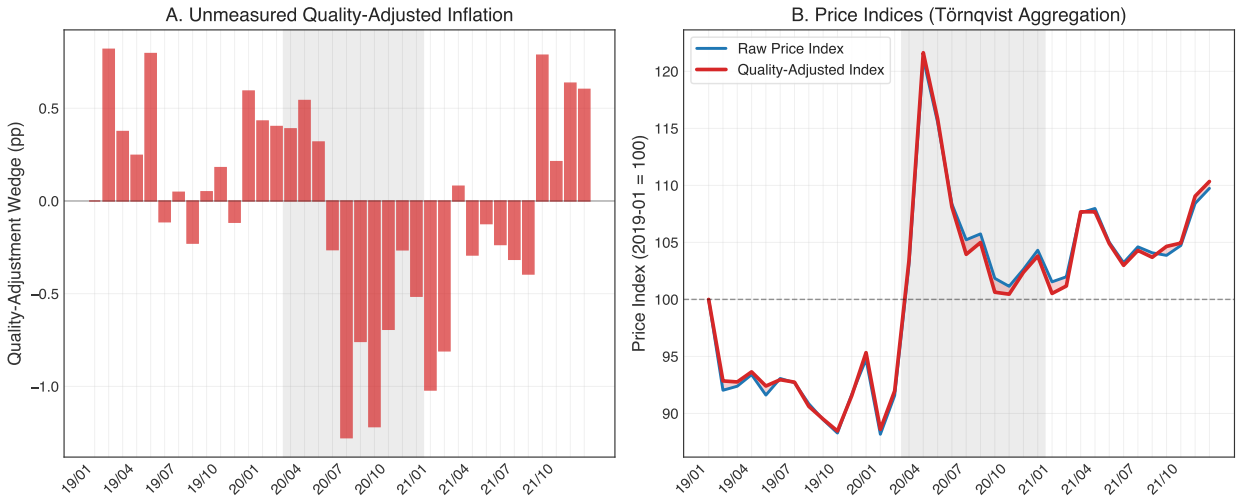


FIGURE 5. The Quality-Adjustment Wedge: Raw vs. Quality-Adjusted Price Indices

Notes: This figure plots monthly price indices from January 2019 to December 2021. The *Raw Price Index* tracks the geometric mean of transaction sizes. The *Quality-Adjusted Index* is derived from the time-dummy hedonic regression, correcting for changes in service attributes. The shaded area represents the "quality-adjustment wedge"—the difference attributable to measured quality changes. Gray shading indicates the COVID period (March–December 2020).

than the Boskin Commission's estimated 0.6 percentage points annual quality bias for goods (Boskin 1996), though concentrated in a period of unusual supply constraints.

Appendix Figures A5–A6 provide additional detail: the dimension-level decomposition shows Reliability and Responsiveness drove the aggregate decline, while word shift analysis reveals increased frequency of terms like "wait," "slow," and "understaffed" in 2021 reviews. Quality trends are broadly similar across metropolitan areas (Appendix Figure A8).

4.6. Implied CPI Bias

Table 7 combines category-specific quality trends with cross-sectional hedonic coefficients to estimate *upper bounds* on implied CPI bias. We use category-specific $\hat{\beta}_c$ estimates rather than the pooled coefficient because quality-price gradients vary across business types. These are upper bounds because, as noted in Section 4.2, the relevant counterfactual lies between the cross-sectional gradient and the near-zero within-establishment response. Limited-service restaurants show the largest implied bias (0.9% annually), reflecting both substantial quality deterioration and a moderate cross-sectional quality premium. Hotels show a null quality-price relationship, suggesting our quality measure may not capture the dimensions valued in that category.

TABLE 7. Upper Bounds on Implied CPI Bias by Category

Category	Nominal Δ	Quality Δ	$\hat{\beta}_c$	Implied Bias
Full-service restaurants	10.8%	−0.020	0.09	0.18%
Limited-service restaurants	9.6%	−0.063	0.15	0.94%
Snack bars	12.3%	−0.049	—	—
Hotels/motels	5.7%	−0.082	—	—
Grocery stores	11.1%	−0.013	—	—

Notes: Nominal Δ is average annual price growth (2019–2021). Quality Δ is average annual change in quality score. $\hat{\beta}_c$ is the cross-sectional hedonic coefficient from category-specific regressions; these estimates represent upper bounds on bias since within-establishment price adjustment is slower. Implied bias equals $\hat{\beta}_c \times |\text{Quality } \Delta|$. Categories with "—" show statistically insignificant quality-price relationships.

These estimates quantify a specific channel of CPI bias: unmeasured changes in service quality. Limited-service restaurants show the largest implied bias (0.9% annually), reflecting both substantial quality deterioration and a strong cross-sectional quality premium. Several categories—snack bars, hotels, and grocery stores—show no significant quality-price relationship, suggesting our quality measure may not capture the dimensions consumers value in those settings. We emphasize that these are estimates, not proposed adjustments to official statistics. Implementation would require addressing representativeness of review-writers, coverage of service categories, and integration with BLS methodology.

5. Service Quality and Firm Value

5.1. Valuation and Firm Performance

The preceding analysis establishes that consumers pay more for higher-quality service in the cross section, consistent with standard hedonic valuation. Whether this quality premium extends beyond prices and is reflected in firm valuations, however, is less clear. We therefore turn from product markets to financial markets to examine whether short-run variation in realized service quality is capitalized into firm value.

Prior work emphasizes customer relationships and brand-related intangibles as potentially important determinants of valuation (Gourio and Rudanko 2014; He, Mostrom, and Sufi 2024). Whether high-frequency fluctuations in the quality of the service encounter—the realized outcome of these investments—contain incremental valuation-relevant information, however, remains an open question. Unlike physical assets recorded on the balance sheet or marketing expenditures observed in financial statements, service quality itself is an unobserved and potentially transient state variable.

This observability friction motivates a natural empirical test. If short-run variation in realized service quality contains incremental information for firm value, it should be reflected in Tobin's Q after controlling for standard accounting proxies. Accordingly, we examine whether the market values the outcome of the service experience (our quality measure) over and above the inputs used to produce it, as proxied by SG&A expenditures.

To test this hypothesis, we link establishment-level quality scores to publicly traded firms via brand name matching (see Section 3.3), then merge with Compustat and CRSP. We aggregate establishment quality to the firm level using transaction-weighted averages (Equation 8), constructing firm-quarter panels for 143 firms with identifiable brand-to-ticker mappings.

Our panel regression follows Mitton (2022). We use the natural log of Tobin's Q as the dependent variable. To isolate the value of service quality from general operating expenditures, we control for Intangible Inputs (SG&A/Assets), in addition to standard controls for firm size (log assets), profitability (ROA), and leverage. To mitigate simultaneity, all control variables are lagged by one quarter. Our sample spans 2015–2021, yielding 143 firms and approximately 2,400 firm-quarters. We use quarter and industry fixed effects rather than firm fixed effects to preserve cross-sectional variation. We winsorize all continuous variables at the 1st and 99th percentiles.

5.1.1. Tobin's Q Regressions

We estimate the following panel regression:

$$\ln(Q_{ft}) = \alpha + \beta_1 \hat{\mathbb{S}}_{ft} + \beta_2 \frac{\text{SG\&A}_{f,t-1}}{\text{Assets}_{f,t-1}} + \Gamma \mathbf{X}_{f,t-1} + \delta_j + \tau_t + \epsilon_{ft} \tag{10}$$

where $\ln(Q_{ft})$ is the natural log of Tobin's Q, $\hat{\mathbb{S}}_{ft}$ is the firm-level hedonic quality score, and the vector of controls includes firm size (ln(Assets)), profitability (ROA), and market

leverage. We include industry fixed effects (δ_j) to control for sector-level differences in valuations and quarter fixed effects (τ_t) to absorb aggregate market shocks.[3]

TABLE 8. Service Quality and Firm Value (Tobin's Q)

	(1) $\ln(\text{Tobin's Q}_{ft})$	(2) $\ln(\text{Tobin's Q}_{ft})$	(3) $\ln(\text{Tobin's Q}_{ft})$
Service Quality (std)	−0.013	0.015	0.017
	(0.044)	(0.035)	(0.029)
ln(Assets)		Yes	Yes
ROA		Yes	Yes
Leverage		Yes	Yes
Intangibles/Assets			Yes
SG&A/Assets			Yes
Quarter FE	Yes	Yes	Yes
Industry FE	No	No	Yes
Observations	2,419	1,879	1,879
Firms	143	143	143
R^2	0.03	0.36	0.54

Notes: This table reports regressions of the natural log of Tobin's Q on firm-level service quality, aggregated from establishment-level scores. Column (1) includes quarter fixed effects only. Column (2) adds standard controls (size, ROA, leverage; all lagged one quarter). Column (3) adds 2-digit SIC industry fixed effects and intangible asset measures (book intangibles, SG&A) to test whether quality predicts valuations beyond existing accounting proxies. Sample drops from 2,419 to 1,879 firm-quarters in Columns (2)–(3) due to missing control variables. Standard errors clustered at the firm level in parentheses. *, **, *** denote significance at 10%, 5%, 1% levels.

Table 8 presents the results. Column (1) reports the unconditional relationship with quarter fixed effects only. Column (2) adds standard firm controls. Column (3) tests whether quality predicts valuations beyond existing intangible measures.

[3]We do not include firm fixed effects. With an average of 17 quarters per firm and limited within-firm variation in quality (quality is relatively persistent), firm FE would absorb most identifying variation while substantially reducing power. Our specification identifies the cross-sectional relationship: do high-quality firms have higher valuations?

In the full sample, the relationship between service quality and Tobin's Q is weak (Column 3: $\beta = 0.017$, $t = 0.59$). But our 143 firms include FedEx, UPS, Southern Company, and Staples—firms where Yelp reviews reflect a sliver of customer interactions. A package delivery company's Yelp reviews come from residential customers, not the B2B contracts that drive firm value.

Restricting to consumer-facing industries—excluding logistics, utilities, wholesale, and financials—yields 130 firms. The coefficient increases to $\beta = 0.050$ ($t = 1.95$, $p = 0.05$). Table 9 shows the pattern holds across definitions: coefficients are uniformly positive and grow as we focus on firms where customer experience plausibly matters.

For a restaurant chain, Yelp reviews capture the interactions that determine repeat business. For a utility, they capture billing complaints. The pattern is consistent with a measure that captures real variation in service quality where it is economically relevant.

These results warrant caution: 130 firms and marginal significance ($p = 0.05$) leave substantial uncertainty. The subsample pattern is consistent with a measure that captures economically relevant variation, but broader data—Google Maps reviews, for instance—could provide sharper estimates.

TABLE 9. Quality and Firm Value: Sample Composition

Sample Definition	Firms	Obs	Quality β	t-stat
Full sample	143	2,419	0.017	0.59
Retail & service SIC only	108	1,936	0.022	0.81
Exclude non-consumer industries	130	2,219	0.050*	1.95
Require Yelp or SafeGraph coverage	123	2,352	0.022	1.12

Notes: This table reports the coefficient on service quality from Tobin's Q regressions under alternative sample definitions. All specifications include controls (size, leverage, ROA), industry fixed effects, and quarter fixed effects. "Non-consumer industries" excludes logistics (SIC 40–47), utilities (49xx), wholesale (50–51), and financials (60–64). Standard errors clustered by firm. * $p < 0.10$.

5.1.2. Visualizing the Relationship

Figure 6 provides a non-parametric view of the consumer-facing subsample. We plot a binned scatterplot of residualized Tobin's Q against residualized service quality scores. After controlling for firm characteristics, firms with higher service quality have modestly higher valuations, consistent with the coefficient in Table 9.

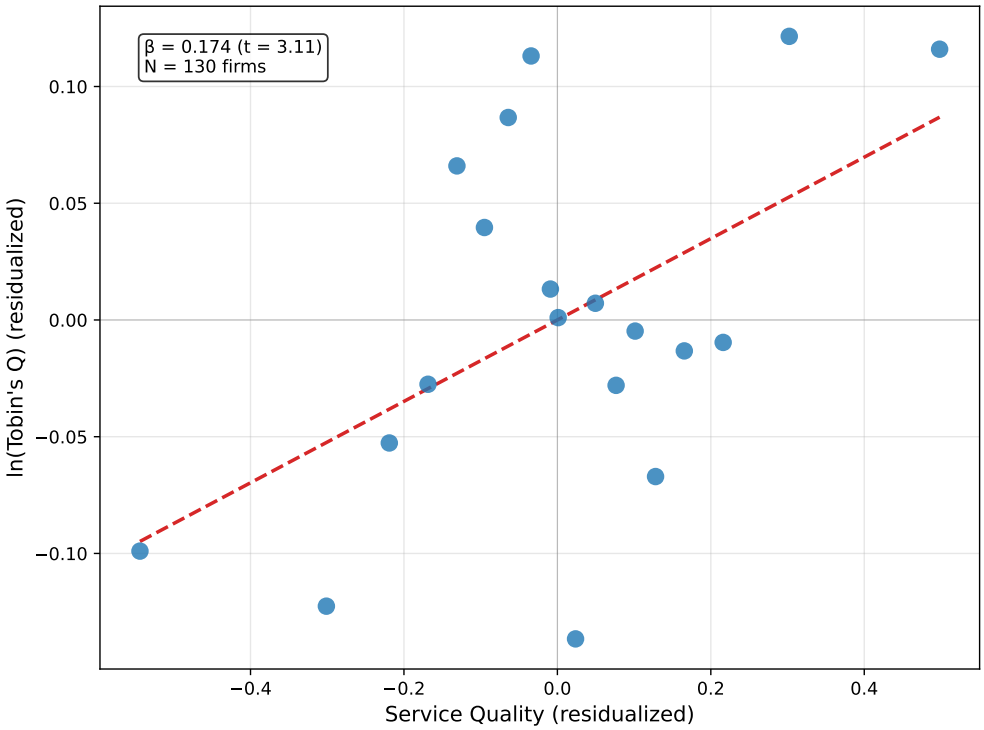


FIGURE 6. Binscatter of Tobin's Q vs. Service Quality (Consumer-Facing Firms)

Notes: This figure presents a binned scatterplot of ln(Tobin's Q) against the firm-level service quality score for consumer-facing industries (excluding logistics, utilities, wholesale, and financials). Both variables are residualized with respect to industry fixed effects, quarter fixed effects, and firm size. Sample: 130 publicly traded firms, 2015–2021. The positive slope is consistent with the marginally significant coefficient ($\beta = 0.050$, $p = 0.05$) in the consumer-facing subsample.

5.1.3. The Quality Premium During the COVID Crisis

Recent theoretical work models customer capital as providing insurance against demand shocks (He, Mostrom, and Sufi 2024). High-quality service relationships should reduce demand elasticity and provide revenue stability during downturns. The COVID-19 pandemic

provides a natural test: if service quality functions as customer capital, high-quality firms should have experienced smaller valuation declines during the crisis.

We test this hypothesis by interacting quality with an indicator for the COVID period (2020 Q1–Q2):

$$\ln(Q_{ft}) = \alpha + \beta_1 \hat{\mathbb{S}}_{ft} + \beta_2 \text{COVID}_t + \beta_3 (\hat{\mathbb{S}}_{ft} \times \text{COVID}_t) + \Gamma \mathbf{X}_{f,t-1} + \delta_j + \tau_t + \epsilon_{ft} \tag{11}$$

The interaction coefficient $\beta_3 = -0.049$ ($t = -1.77$, $p = 0.078$) is negative and marginally significant. This indicates that the quality-valuation relationship *weakened* during the COVID crisis: the cross-sectional quality premium that exists in normal times essentially disappeared during the crisis quarters. A one standard deviation increase in quality, which is associated with modestly higher valuations outside the crisis, showed no such relationship during COVID.

This finding is inconsistent with the simple "customer capital as insurance" hypothesis. Rather than high-quality firms being insulated from the shock, the quality premium disappeared precisely when service delivery became difficult to maintain. One interpretation is that markets anticipated quality deterioration: firms known for high service quality faced greater operational challenges in sustaining that quality during staffing shortages, leading investors to discount the value of pre-crisis quality.

We interpret this result cautiously. The COVID period provides only two quarters of treatment variation, and the interaction relies on cross-sectional differences in quality. The finding suggests that service quality may be fragile rather than resilient during crises—a form of intangible capital that is difficult to maintain when operational stress is highest.

5.1.4. Service Quality and Intangible Capital

Brand quality is an important component of organizational and intangible capital (Vomberg, Homburg, and Bornemann 2015; Mizik and Jacobson 2008; Ewens, Peters, and Wang 2025),

and service quality plausibly contributes to brand value by shaping consumers' perceptions and repeat-purchase behavior. We examine this relationship by testing whether firms with higher perceived service quality also exhibit higher levels of firm-level intangible capital.

We construct firm-level intangible capital using the capitalization framework of Eisfeldt and Papanikolaou (2013, 2014). Capital stocks are formed via a perpetual-inventory method applied to R&D and SG&A expenditures, using standard parameters from the literature: 15% annual depreciation for both R&D and organizational capital, with 30% of SG&A treated as long-lived investment.

We measure *total intangible capital* by capitalizing both R&D expenditure and SG&A from Compustat. We also construct *organizational capital* separately by capitalizing SG&A only, following the distinction that organizational capital captures customer-facing processes and internal structures, while total intangible capital additionally includes innovation-oriented assets. Distinguishing between these measures allows us to test whether service quality is more closely related to organizational and brand capital than to broader intangible investments.

We find a weak positive correlation between service quality and book intangibles ($r = 0.08$), consistent with the intuition that quality contributes to brand value. However, the correlation with organizational capital—constructed by capitalizing SG&A expenditures—is negative ($r = -0.35$). The negative correlation indicates that firms with high customer-facing investment do not necessarily deliver superior service quality; spending more on SG&A does not guarantee better outcomes. Regressing book intangibles on quality (controlling for size, leverage, and ROA) yields a coefficient indistinguishable from zero. We also tested whether quality matters more for intangible-intensive firms by interacting quality with organizational capital intensity; the interaction coefficient is essentially zero ($\beta = 0.003$, $p = 0.91$).

Three factors may explain the weak relationship. Intangible capital stocks reflect cumulative past investments, whereas our quality measure captures contemporaneous service delivery—a stock-flow mismatch that attenuates correlations. Accumulated intangibles also encompass marketing, training, and information systems that need not map tightly to the service dimensions observed in reviews. And firms with substantial intangible capital may compete on different margins, limiting the informational content of short-run quality variation.

Because intangible capital stocks reflect cumulative past investments, while our service quality measures are based on contemporaneous consumer reviews, these results should be interpreted as descriptive rather than causal.

Our mixed valuation results contrast with prior work finding that customer satisfaction predicts stock returns (Fornell et al. 2006; Fornell, Morgeson III, and Hult 2016). Three differences may explain this contrast. First, we measure service quality, not satisfaction—quality is an input to satisfaction, but satisfaction also reflects price, convenience, and expectations. Second, the ACSI covers large firms with broad brand recognition; our Yelp-based measure captures local service delivery, which may be less valuation-relevant for diversified firms. Third, markets may have become more efficient at incorporating customer sentiment since the original ACSI studies. Our consumer-facing result ($\beta = 0.050$, $p = 0.05$) suggests that the relationship emerges when the measure is economically relevant, consistent with Edmans (2011), who finds employee satisfaction predicts returns only in certain periods and samples.

The weak relationship in the full sample, combined with the negative quality-organizational capital correlation, suggests several directions for future research. With access to daily CRSP returns and Fama-French factors, researchers could test whether quality predicts risk-adjusted returns or equity volatility more broadly. Linking quality to analyst forecasts would test whether quality information is incorporated into earnings expectations. Expanding beyond our sample of 143 publicly traded chains to include private firms or

international markets would improve generalizability. We leave these extensions for future work, noting that our COVID crisis analysis provides preliminary evidence that quality may matter most during periods of stress.

6. Conclusion

This paper develops a scalable method for measuring service quality from consumer review text and applies it to two foundational problems: adjusting price indices for quality change and testing whether quality predicts firm value.

We validate our text-based quality measures using face validity tests, incremental validity over star ratings, and linguistic analysis. Known service leaders rank highly; text predicts prices beyond what stars capture; and word patterns confirm the model distinguishes service quality from product quality.

Applying our quality measures to 110,679 establishment-months, we find that service quality commands a hedonic premium: a one standard deviation increase in quality is associated with 8.7% higher transaction spending (cross-sectional). Aggregate quality declined by 5.0 percentage points between 2019 and 2021, with the sharpest deterioration during the economic recovery. This quality decline implies a potential quality-adjustment wedge of 0.2–0.4 percentage points annually, depending on whether the cross-sectional or within-establishment hedonic coefficient better captures the relevant price response.

Analysis of 143 public firms (approximately 2,400 firm-quarters) finds no significant relationship between service quality and Tobin's Q after controlling for firm characteristics. The unconditional coefficient is essentially zero ($\beta = -0.01$, SE = 0.04), and coefficients remain null with controls for size, profitability, leverage, and intangible assets. This is consistent with markets already pricing quality through existing information channels, though alternative interpretations (measurement error in firm-level aggregation, sample selection) cannot be ruled out.

First, our analysis establishes correlations, not causal effects. The cross-sectional hedonic coefficient identifies the equilibrium quality-price relationship but cannot distinguish demand-side willingness-to-pay from supply-side cost differences. The within-establishment specification is subject to attenuation bias from measurement error in quality changes. Second, we measure perceived quality among review-writers, who skew younger and more urban than the general population. Third, only 27% of Yelp establishments match to SafeGraph spending data, and these are predominantly chain restaurants in urban areas. Results may not generalize to independent establishments or non-urban markets. Fourth, our valuation sample of 143 publicly-traded firms with adequate Yelp coverage is highly selected; effects in this sample may differ from the broader service economy.

For macroeconomists and statistical agencies, our approach demonstrates the feasibility of hedonic quality adjustment for services. The methodology could extend BLS quality adjustment from goods to services, though implementation would require addressing representativeness of review-writers and integration with existing procedures.

For investors and corporate managers, our null Tobin's Q result suggests that markets may already incorporate quality information through existing channels. Whether systematic quality measurement could identify mispricing in less-followed firms—or provide operational diagnostics independent of valuation—remains an open question.

For researchers, we provide validated quality measures at the establishment-month level that can be merged with other data sources for further study. We release our code and aggregated quality indices for replication and extension.

Several directions remain. First, extending quality measurement to services beyond food and hospitality—healthcare, automotive, personal services—would test whether the methodology scales to the broader service economy. Second, event-study designs around operational changes (menu simplification, self-checkout adoption, staffing reductions) could provide sharper identification of how quality adjusts to cost pressures. Third, survey

validation comparing review-based quality measures to perceptions of non-reviewers would address the representativeness concerns central to statistical agency adoption.

References

Aladangady, Aditya, Shifrah Aron-Dine, Wendy Dunn, Laura Feiveson, Paul Lengermann, and Claudia Sahm. 2019. "From transactions data to economic statistics: Constructing real-time, high-frequency, geographic measures of consumer spending." In *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Anderson, Michael and Jeremy Magruder. 2012. "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database." *The Economic Journal* 122 (563): 957–989. 10.1111/j.1468-0297.2012.02512.x.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews." *Management Science* 57 (8): 1485–1509. 10.1287/mnsc.1110.1370.

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring economic policy uncertainty." *The quarterly journal of economics* 131 (4): 1593–1636.

Bils, Mark and Peter J Klenow. 2004. "Some evidence on the importance of sticky prices." *Journal of Political Economy* 112 (5): 947–985.

Boskin, Michael J. 1996. *Toward a more accurate measure of the cost of living*. United States Senate, Committee on Finance.

Chetty, Raj, John N Friedman, and Michael Stepner. 2024. "The economic impacts of COVID-19: Evidence from a new public database built using private sector data." *The Quarterly Journal of Economics* 139 (2): 829–889.

Dubé, Jean-Pierre, Günter J Hitsch, and Peter E Rossi. 2010. "State dependence and alternative explanations for consumer inertia." *The RAND Journal of Economics* 41 (3): 417–445.

Edmans, Alex. 2011. "Does the stock market fully value intangibles? Employee satisfaction and equity prices." *Journal of Financial economics* 101 (3): 621–640.

Ehrlich, Gabriel, John C Haltiwanger, Ron S Jarmin, David Johnson, Ed Olivares, Luke W Pardue, Matthew D Shapiro, and Laura Zhao. 2023. "Quality adjustment at scale: Hedonic vs. exact demand-based price indices." Tech. rep., National Bureau of Economic Research.

Eisfeldt, Andrea L and Dimitris Papanikolaou. 2013. "Organization capital and the cross-section of expected returns." *The Journal of Finance* 68 (4): 1365–1406.

Eisfeldt, Andrea L and Dimitris Papanikolaou. 2014. "The value and ownership of intangible capital." *American Economic Review* 104 (5): 189–194.

Ewens, Michael, Ryan H. Peters, and Sean Wang. 2025. "Measuring Intangible Capital with Market Prices." *Management Science* 71 (1): 407–427. 10.1287/mnsc.2021.02058.

Fornell, Claes, Sunil Mithas, Forrest V Morgeson III, and Mayuram S Krishnan. 2006. "Customer satisfaction and stock prices: High returns, low risk." *Journal of marketing* 70 (1): 3–14.

Fornell, Claes, Forrest V Morgeson III, and G Tomas M Hult. 2016. "Stock returns on customer satisfaction do beat the market: Gauging the effect of a marketing intangible." *Journal of marketing* 80 (5): 92–107.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as data." *Journal of Economic Literature* 57 (3): 535–574.

Gourio, Francois and Leena Rudanko. 2014. "Customer capital." *Review of Economic Studies* 81 (3): 1102–1136.

Handel, Benjamin R. 2013. "Adverse selection and inertia in health insurance markets: When nudging hurts." *American Economic Review* 103 (7): 2643–2682.

Hassan, Tarek A, Stephan Hollander, Laurence Van Lent, and Ahmed Tahoun. 2019. "Firm-level political risk: Measurement and effects." *The quarterly journal of economics* 134 (4): 2135–2202.

He, Bianca, Lauren I Mostrom, and Amir Sufi. 2024. "Investing in Customer Capital." Tech. rep., National Bureau of Economic Research.

Hortaçsu, Ali, Seyed Ali Madanizadeh, and Steven L Puller. 2017. "Power to choose? An analysis of consumer inertia in the residential electricity market." *American Economic Journal: Economic Policy* 9 (4): 192–226.

Hu, Nan, Jie Zhang, and Paul A Pavlou. 2009. "Overcoming the J-shaped distribution of product reviews." *Communications of the ACM* 52 (10): 144–147.

Klenow, Peter J and Oleksiy Kryvtsov. 2008. "State-dependent or time-dependent pricing: Does it matter for recent US inflation?" *The Quarterly Journal of Economics* 123 (3): 863–904.

Kotler, Philip and Kevin Lane Keller. 2012. *Marketing Management*. Upper Saddle River, NJ: Prentice Hall, 14th ed.

Lawani, Abdelaziz, Michael R. Reed, Tyler Mark, and Yuqing Zheng. 2019. "Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston." *Regional Science and Urban Economics* 75: 22–34. https://doi.org/10.1016/j.regsciurbeco.2018.11.003.

Leisen Pollack, Birgit. 2009. "Linking the hierarchical service quality model to customer satisfaction and loyalty." *Journal of services marketing* 23 (1): 42–50.

Luca, Michael. 2016. "Reviews, reputation, and revenue: The case of Yelp. com." *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper* (12-016).

Luo, Xueming, Christian Homburg, and Jan Wieseke. 2010. "Customer satisfaction, analyst stock recommendations, and firm value." *Journal of Marketing Research* 47 (6): 1041–1058.

Mitton, Todd. 2022. "Methodological variation in empirical corporate finance." *The Review of Financial Studies* 35 (2): 527–575.

Mizik, Natalie and Robert Jacobson. 2008. "The Financial Value Impact of Perceptual Brand Attributes." *Journal of Marketing Research* 45 (1): 15–32. 10.1509/jmkr.45.1.015.

Nakamura, Emi and Jón Steinsson. 2008. "Five facts about prices: A reevaluation of menu cost models." *The Quarterly Journal of Economics* 123 (4): 1415–1464.

National Restaurant Association. 2021. "2021 Mid-Year State of the Restaurant Industry Update." Tech. rep., National Restaurant Association. Survey finding: 75% of operators said recruitment/retention is their top challenge; 78% do not have enough employees to support customer demand.

Parasuraman, ABLL, Valarie A Zeithaml, and Leonard Berry. 1988. "SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality." *1988* 64 (1): 12–40.

Peters, Ryan H and Lucian A Taylor. 2017. "Intangible capital and the investment-q relation." *Journal of Financial Economics* 123 (2): 251–272.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82 (1): 34–55. 10.1086/260169.

Rust, Roland T, Katherine N Lemon, and Valarie A Zeithaml. 2004. "Return on marketing: Using customer equity to focus marketing strategy." *Journal of marketing* 68 (1): 109–127.

Triplett, Jack E. 2006. *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products*. Paris: Organisation for Economic Co-operation and Development.

Truong, Cameron, Thu Ha Nguyen, and Thanh Huynh. 2021. "Customer satisfaction and the cost of capital." *Review of accounting studies* 26 (1): 293–342.

Tunstall, Lewis, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. "Efficient few-shot learning without prompts." *arXiv preprint arXiv:2209.11055* .

U.S. Bureau of Labor Statistics. 2022. "Job Openings and Labor Turnover Survey: Accommodation and Food Services." Tech. rep., U.S. Bureau of Labor Statistics. Quits rate in accommodation and food services increased from 4.8% (2019) to 6.9% (Nov 2021). Data from JOLTS Table 4.

U.S. Bureau of Labor Statistics. 2025. "CPI Quality Adjustment Methodology List." https://www.bls.gov/cpi/quality-adjustment/qa-methodology-list.xlsx. List of CPI entry-level items and their quality-adjustment methods.

Vomberg, Arnd, Christian Homburg, and Torsten Bornemann. 2015. "Talented people and strong brands: The contribution of human capital and brand equity to firm value." *Strategic Management Journal* 36 (13): 2122–2131. https://doi.org/10.1002/smj.2328.

Appendix A. Data Matching and Sample Construction

A.1. SafeGraph–Yelp Matching Algorithm

We match SafeGraph establishments to Yelp businesses using ZIP-code blocking combined with spatial and name filtering. The algorithm requires three conditions:

a. **ZIP-code match**: Establishments must share the same 5-digit ZIP code. This blocking step reduces computational cost and ensures geographic coherence; 99.5% of Yelp businesses are in ZIP codes with SafeGraph coverage.
b. **Spatial proximity**: Within each ZIP code, we compute haversine distances and retain candidates within 100 meters.
c. **Name or brand similarity**: We compute Jaro-Winkler similarity between establishment names (after standardization: lowercasing, removing punctuation). A match requires similarity ≥ 0.80 on either the business name or the brand name. Brand matching captures franchises where the SafeGraph brand field (e.g., "Starbucks") may differ from the location name (e.g., "Starbucks Coffee #12345").

When multiple Yelp businesses match a single SafeGraph location, we keep the match with the highest name similarity; ties are broken by distance.

The final matched sample contains 38,894 unique establishments with high match quality: median geographic distance is 8 meters, and median name similarity is 1.0 (exact match). Brand-based matches account for 12% of the sample, primarily national chains.

A.2. NAICS Classification

Yelp's business categories are self-reported and not mutually exclusive, complicating sector-level analysis. We map establishments to standardized NAICS codes using a three-tier approach:

a. **Chain matching** (16.5% of businesses): For 688 national chains appearing in both datasets, we apply the SafeGraph NAICS code directly, with 33 manual corrections for misclassified chains.
b. **Exact name matching** (1.5%): For non-chains, we check whether the exact business name appears elsewhere in SafeGraph with a consistent NAICS code.
c. **Pattern matching** (82%): Remaining businesses are classified by matching Yelp category strings to NAICS codes. Priority ordering resolves ambiguities: gas stations are evaluated before restaurants (many operate food service), convenience stores before snack bars, and car rentals before hotels.

Validation against a hand-coded sample yields 92% accuracy. We aggregate to 4-digit NAICS codes for sample size, except for restaurants (40% of establishments), where we preserve 6-digit codes to distinguish full-service (722511), limited-service (722513), and snack bars (722515).

Appendix B. NLP Methodology

We measure service quality from Yelp review text using the SERVQUAL framework (Parasuraman, Zeithaml, and Berry 1988). We first label a stratified sample of reviews using an LLM, then train classifiers to replicate those labels. The trained classifiers score all 7 million reviews, which we aggregate to establishment-quarter quality measures.

B.1. LLM Labeling

We label 10,000 reviews using Gemini 2.5 Pro with the following prompt:

> You are labeling Yelp reviews with SERVQUAL service quality scores.
>
> Use the canonical SERVQUAL definitions from Kotler & Keller (2012), Marketing Management, 14th edition, Table 13.5.

SERVQUAL DIMENSIONS:

1. *Reliability*: The ability to perform the promised service dependably and accurately.

2. *Responsiveness*: Willingness to help customers and provide prompt service.

3. *Assurance*: The knowledge and courtesy of employees and their ability to convey trust and confidence.

4. *Empathy*: The provision of caring, individualized attention to customers.

5. *Tangibles*: The appearance of physical facilities, equipment, personnel, and communication materials.

SCORING: 1=Very negative, 2=Mostly negative, 3=Neutral/not mentioned, 4=Mostly positive, 5=Very positive

RULES:

- Score ONLY based on evidence in the review
- If a dimension is NOT mentioned, assign 3
- Ignore price, value-for-money, or unrelated product features
- Return ONLY valid JSON

We use temperature 0 for reproducibility—the same review yields identical scores on repeated queries. Reviews are sampled with stratification by business type (restaurants, hotels, retail, healthcare, automotive, other) and star rating (1–5) to ensure coverage of negative reviews and non-restaurant businesses.

B.2. SetFit Classifiers

We train five SetFit classifiers (Tunstall et al. 2022), one per SERVQUAL dimension, using the LLM-labeled reviews as training data. SetFit is a few-shot learning framework that fine-tunes sentence transformers via contrastive learning, then trains a classification head on the resulting embeddings.

Two architecture choices are notable. First, we use `all-MiniLM-L6-v2` as the base encoder, which produces 384-dimensional embeddings with fast inference. Second, we replace the standard classification head with ordinal logistic regression. Standard classification treats all errors equally: predicting 2 when the true label is 1 costs the same as predicting 5. Ordinal regression respects the ordering, modeling cumulative probabilities $P(Y > k)$ for each threshold k. We train on 8,000 reviews and evaluate on 992 held-out reviews. Accuracy ranges from 61% (Empathy) to 72% (Tangibles), with mean absolute error uniformly below 0.5 on the 1–5 scale. Performance improves monotonically with training set size, with diminishing returns beyond 4,000 examples.

B.3. Aggregation

We apply the trained classifiers to all 7 million Yelp reviews, producing five scores per review. Reviews are truncated to 256 tokens to fit the encoder's context window; because reviewers typically front-load their main points, truncation retains the most salient content.

We aggregate review-level scores to establishment-quarters using mean-centered scores: $(\bar{s} - 3)/2$, which ranges from -1 (all scores = 1) to $+1$ (all scores = 5). This preserves distributional information lost by net approval metrics and provides continuous variation for regression analysis.

B.4. Limitations

LLM labels are not ground truth—we replicate Gemini's interpretation of SERVQUAL, which may contain systematic biases. Truncating reviews to 256 tokens loses some information, and a 5-point scale cannot capture fine-grained sentiment. Yelp reviewers skew toward extreme experiences and are not representative of all customers. SERVQUAL was developed

for traditional service businesses; its applicability to retail or healthcare may differ from restaurants.

Appendix C. Robustness Checks

C.1. COVID Period Exclusion

The COVID-19 pandemic coincided with sharp changes in both service quality and consumer behavior. To ensure our results are not driven by pandemic-specific dynamics, we re-estimate the hedonic specifications excluding observations from 2020 Q2–Q4.

For this robustness check, we aggregate to establishment-quarters rather than establishment-months, yielding 28,299 observations in the full sample and 22,816 after excluding COVID quarters. The quarterly cross-sectional coefficient ($\beta = 0.099$) is modestly larger than the monthly baseline in Table 6 ($\beta = 0.087$), reflecting reduced measurement error from temporal aggregation; both estimates indicate an 8–10% spending premium per standard deviation of quality.

Table A1 reports the results. Excluding the COVID period has negligible impact: the cross-sectional coefficient is $\beta = 0.100$ (SE = 0.008) versus $\beta = 0.099$ (SE = 0.008) in the full quarterly sample, a difference of 1%. The within-establishment coefficient remains economically small and statistically indistinguishable from zero in both samples, consistent with the main results.

Appendix D. Additional Exhibits

D.1. Spending Patterns

Figure A1 plots the distribution of median spend per transaction across the ten largest business categories. Substantial heterogeneity emerges both across and within categories.

TABLE A1. Hedonic Results: COVID Period Exclusion

	Full Sample		Excl. COVID	
	β	SE	β	SE
Cross-sectional	0.099	0.008	0.100	0.008
Within-establishment	−0.003	0.005	−0.002	0.006
N	28,299		22,816	

Notes: Cross-sectional specification includes NAICS × quarter fixed effects. Within-establishment specification adds establishment fixed effects. COVID period defined as 2020 Q2–Q4. Standard errors clustered by establishment.

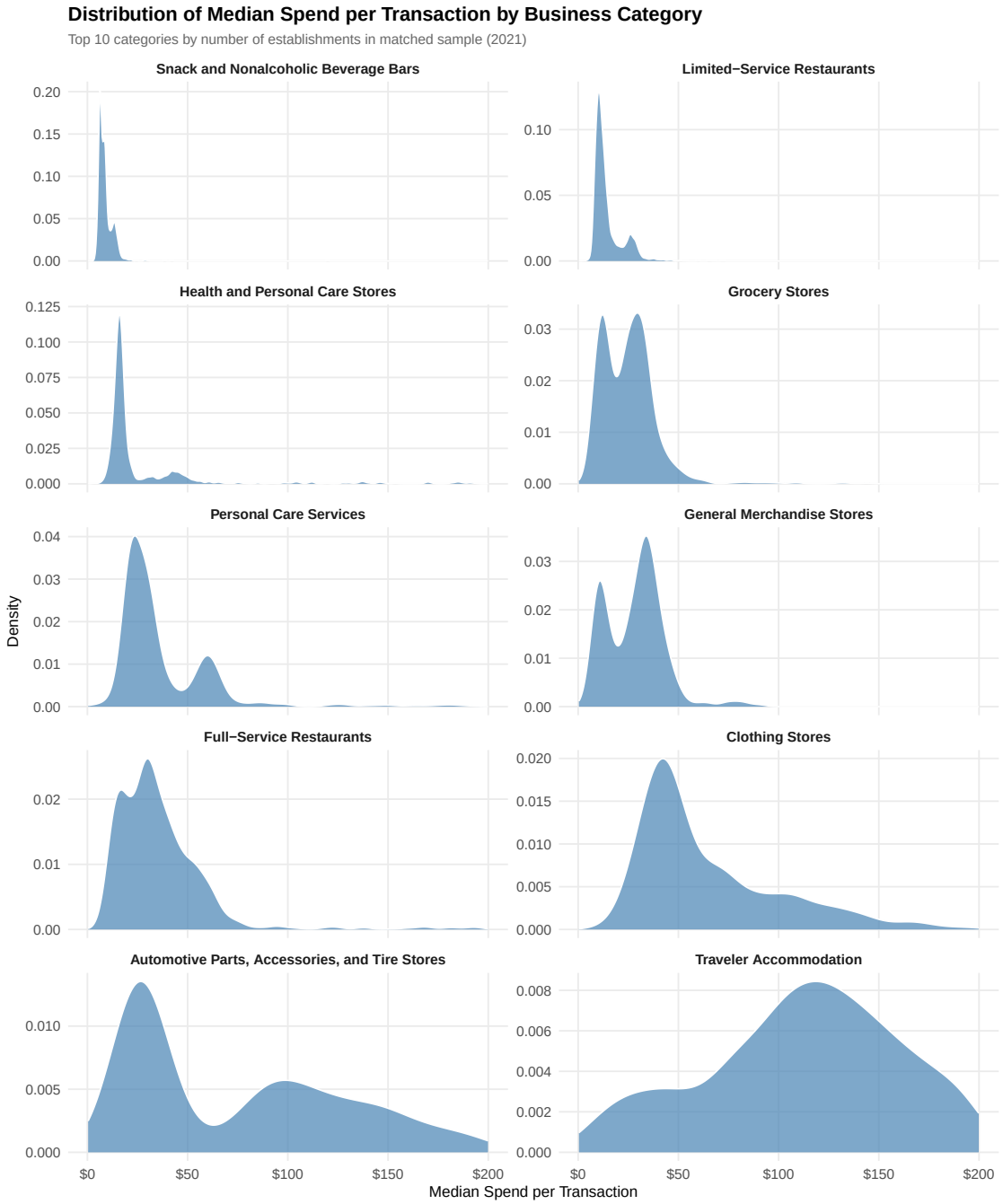


FIGURE A1. Distribution of Median Spend per Transaction by Business Category

Notes: This figure shows the distribution of median spend per transaction for the ten largest business categories in our matched sample. Each panel displays a density plot for one category. Categories are ordered by median spending. The sample includes all matched establishments with complete spending data in 2021.

Figure A2 plots the monthly evolution of median spend per transaction from January 2019 through December 2021, separately for the five largest business categories.

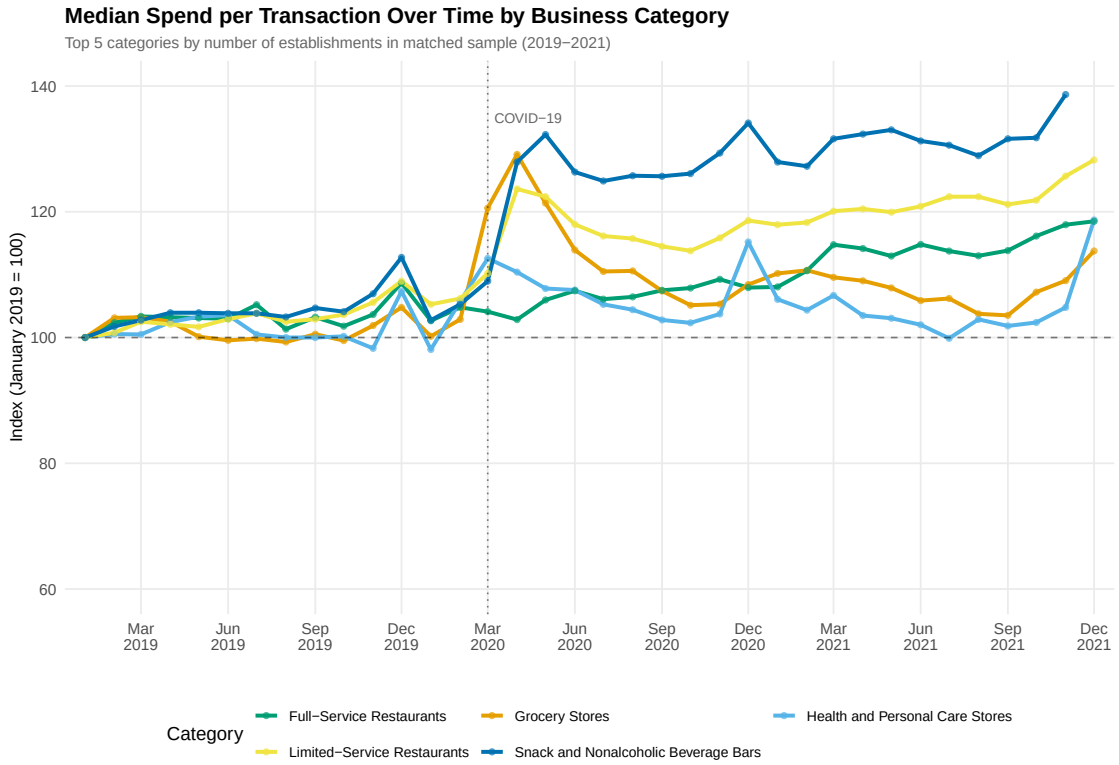


FIGURE A2. Median Spend per Transaction Over Time by Business Category

Notes: This figure plots the monthly median spend per transaction for the five largest business categories in our matched sample from January 2019 to December 2021. Each series is normalized to 100 in January 2019.

D.2. Hedonic Detail

Figure A3 visualizes the estimated shadow prices for service attributes across the largest categories. The heatmap reveals substantial heterogeneity in how consumers value different aspects of the service experience.

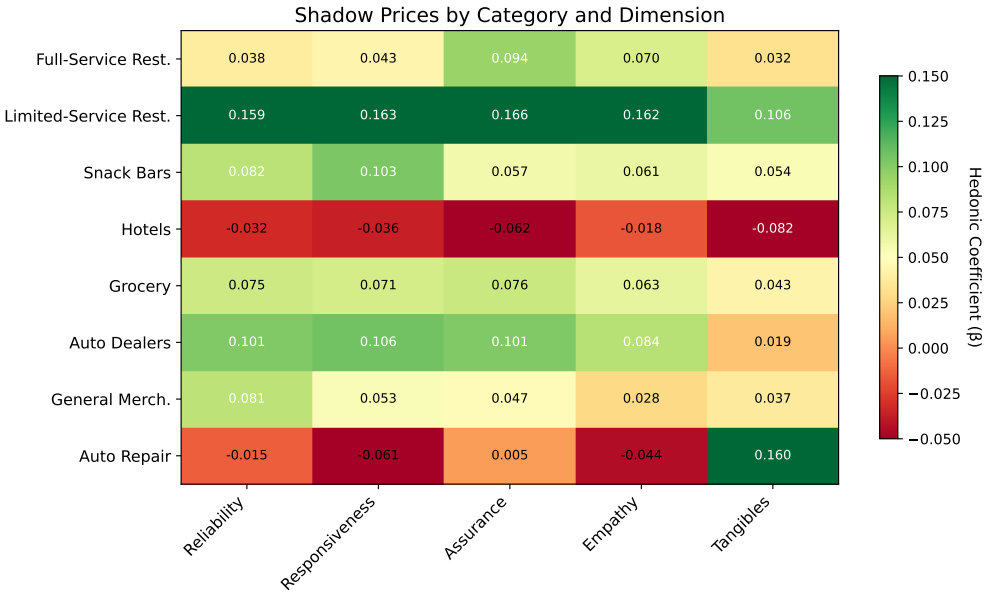


FIGURE A3. Heatmap of Hedonic Shadow Prices by Category

Notes: This figure displays the estimated hedonic coefficients (β_k) for each SERVQUAL dimension across major business categories. Darker shades indicate a larger positive marginal willingness-to-pay for a one-standard-deviation improvement in that attribute. Coefficients are estimated using the hedonic specification, interacting aspect scores with category indicators.

D.3. Quality Trends Over Time

Figure A4 plots the monthly sentiment for each SERVQUAL dimension from January 2019 through December 2021, indexed to 100 at the start of the period to facilitate comparison of relative declines across dimensions with different baseline levels.

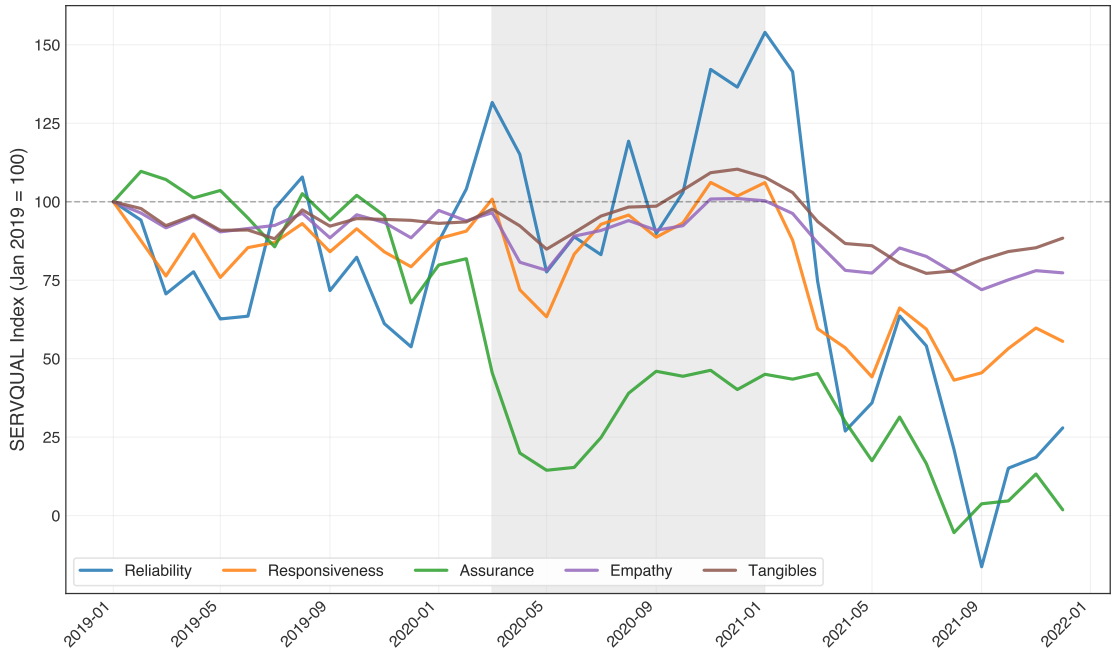


FIGURE A4. SERVQUAL Dimensions Over Time (Indexed to January 2019 = 100)

Notes: This figure plots monthly average sentiment scores for each SERVQUAL dimension, indexed so that January 2019 = 100. Shaded region indicates the initial COVID period (March–December 2020). Indexing reveals that proportional declines varied substantially: Assurance fell to 6% of baseline (raw: 0.21 to 0.01), Reliability to 23% (raw: 0.28 to 0.06), Responsiveness to 64% (raw: 0.13 to 0.08), Empathy to 52% (raw: 0.16 to 0.08), while Tangibles remained at 89% (raw: 0.13 to 0.12). The divergent patterns suggest that staff-dependent dimensions were disproportionately affected by pandemic labor disruptions.

D.4. Quality Decomposition

Figure A5 decomposes the change in the aggregate quality score into contributions from the distinct SERVQUAL dimensions. The bars represent the weighted contribution of each aspect to the total deviation from the 2019 baseline.

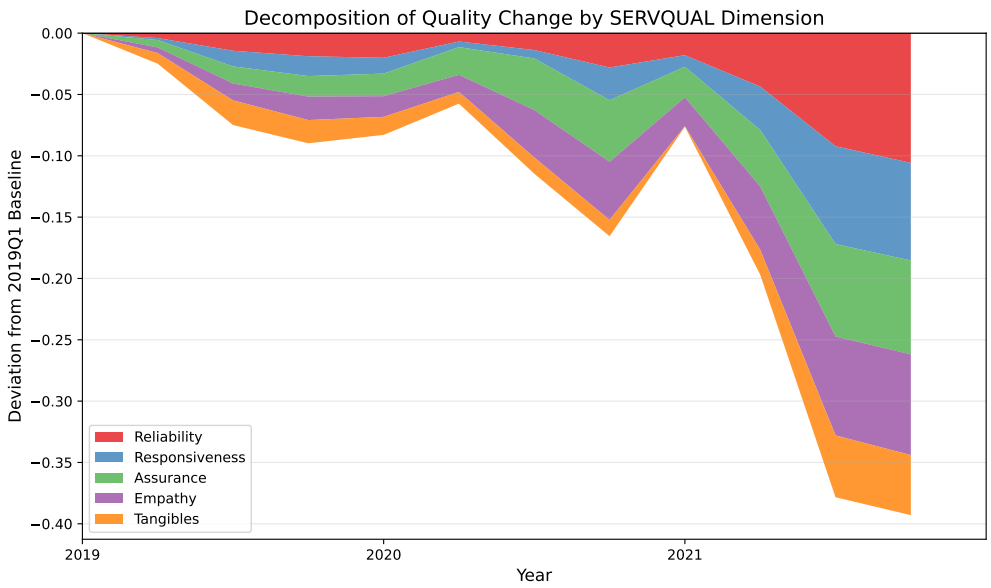


FIGURE A5. Decomposition of Quality Change by SERVQUAL Dimension

Notes: This figure presents a stacked bar chart decomposing the monthly change in the aggregate Service Quality Index relative to the January 2019 baseline. Each color corresponds to one SERVQUAL dimension. The height of each segment represents that dimension's weighted contribution to the total quality deviation.

D.5. Linguistic Evidence

Figure A6 presents a word shift graph comparing the vocabulary of reviews in late 2021 against the 2019 baseline, providing semantic context for the observed quality trends.



FIGURE A6. Word Shift Analysis: 2019 vs. Late 2021

Notes: This figure displays terms that most strongly distinguish review language in 2019 and late 2021 to early 2022, based on differences in average TF–IDF weights across the two years. Terms on the right are more characteristic of reviews written in 2021–2022, while terms on the left are more characteristic of reviews from 2019. The ranking reflects relative distinctiveness in language use rather than raw frequency or sentiment.

D.6. Validation Detail

Table A2 presents the top TF-IDF phrases distinguishing low vs high SERVQUAL scores for each dimension, with other dimensions held at neutral.

Figure A7 shows that SetFit classifier performance improves monotonically with training set size, with diminishing returns beyond 4,000 examples.

D.7. Heterogeneity

Figure A8 presents differential trends in aggregate service quality across the MSAs in the Yelp dataset. Quality trends are broadly similar across geographic areas, with all regions showing declines during 2020–2021.

TABLE A2. Top TF–IDF phrases distinguishing Low vs High SERVQUAL scores (controls held constant)

	Reliability		Responsiveness		Assurance		Empathy		Tangibles	
	Low	High	Low	High	Low	High	Low	High	Low	High
rank										
1	waste money	cooked perfectly	slow service	fast service	stay away	customer service	customer service	feel welcome	place dirty	outdoor seating
2	don know	cooked perfection	20 minutes	service fast	don know	knowledgeable staff	don care	ice cream	parking lot	live music
3	waste time	perfectly cooked	30 minutes	quick service	wearing mask	staff friendly	zero stars	make feel	place filthy	happy hour
4	don waste	food service	service slow	food fast	buyer beware	friendly knowledgeable	turned away	super sweet	absolutely disgusting	really enjoyed
5	food poisoning	fresh delicious	customer service	service quick	don waste	staff knowledgeable	don want	food delicious	stay away	food delicious

Notes: Top 5 phrases shown. Full table includes top 20 phrases per dimension. High-scoring reviews cluster around staff behavior and speed; low-scoring reviews cluster around complaints and delays.

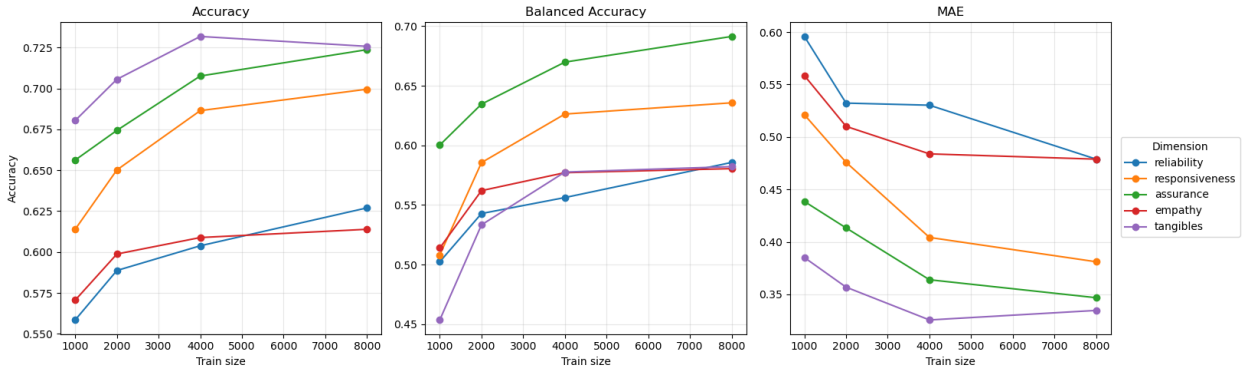


FIGURE A7. SetFit Model Performance Over Sample Size

Notes: Classification accuracy by training set size. Performance improves with more training data but shows diminishing returns beyond 4,000 examples.

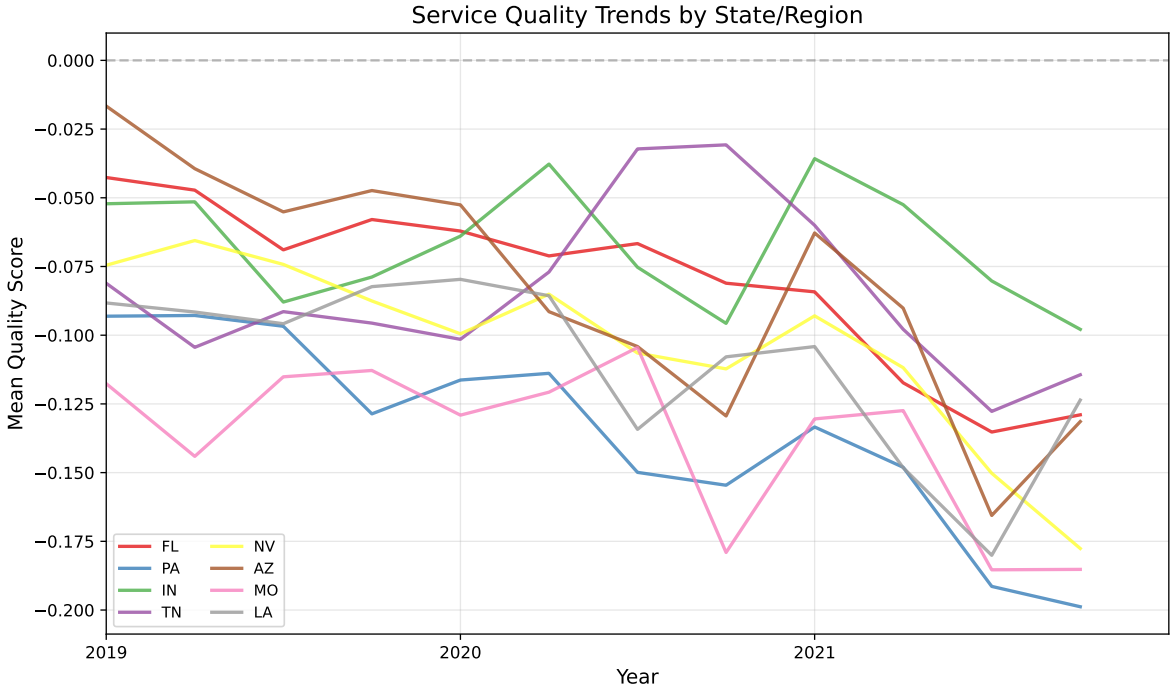


FIGURE A8. Geographic Dispersion in Service Quality Perceptions

Notes: Figure showing differential trends in aggregate service quality across metropolitan areas in the sample.