Painting Alpha

Arnaud Castillo[1,2], Douglas Bruce[3], and Louis Houseman[4]

[1]TheArt.fund
[2]Université Paris Dauphine
[3]University College London
[4]University of Oxford

October 2026

Abstract

Art prices are often said to defy explanation, set by taste, emotion and the private judgement of collectors. We test this view on 2,460 artists whose work has sold at leading auction houses, asking how much of an artist's price can be explained by what is already known about them, without ever showing our models a price. Three questions guide the study. Can an AI, asked to judge artists as a museum curator would, outperform a model built on the public record of exhibitions, galleries and sales? It cannot. Do the two see the same things? Largely, although about a quarter of what the AI captures is absent from the record. Does combining them help? Markedly: together they explain around 70% of the price differences between artists. The structure that emerges is best described as legacy, prestige sustained across a long career and many bodies of work. What remains unexplained is the artist's alpha.

1 Introduction

The relationship between fundamental ratios, often referred to as factors, and asset prices has been extensively studied in finance, with hundreds of papers examining the factors that determine stock returns. Modern asset pricing research on stock returns began with the Capital Asset Pricing Model of Sharpe (1964), which proposed that a single market-risk factor explained the cross-section of expected returns. Fama and French (1993) later showed that size and value characteristics captured additional risk beyond the market factor, launching a broader search for return-predicting variables. This search proliferated into what Cochrane (2011) called a "zoo" of factors, making it increasingly difficult to distinguish genuine risk factors from statistical artifacts. Feng et al. (2020) addressed this directly, constructing a library of 150 risk factors and testing their marginal explanatory power using a formal model-selection framework.

Art, by contrast, has received comparatively little academic attention, despite a market that recorded an estimated $57.5 billion in sales in 2024 (Art Basel and UBS, 2025) and an estimated $2.6 trillion in art and collectible wealth held by ultra-high-net-worth individuals (Deloitte and ArtTactic, 2025).

This paper was prepared for TheArt.fund, which invests in artworks. The views expressed are the authors' own and do not constitute investment advice. Corresponding author: info@theart.fund.

Part of this neglect reflects a view, widespread among market participants, that art prices resist systematic explanation altogether, as they are driven primarily by the idiosyncratic tastes and emotional responses of individual collectors and curators, rather than by any identifiable set of fundamentals. This skepticism extends to a broader rejection of the financial approach to art collecting itself, with many collectors, dealers and curators insisting that art should be valued as a cultural and aesthetic object rather than as a financial asset subject to factor-based analysis.

This paper challenges that view. Using a cross-section of 2,460 artists whose works have sold at leading international auction houses, we ask how much of the variation in artist-level prices can be explained by characteristics measured without reference to price. We construct three linear factor models of the log average price per unit area. The first is a quantitative model of thirteen public-record measures of institutional recognition, exhibition activity, geographic reach and secondary-market activity, which explains 49.3% of cross-sectional variation in sample and 48.2% on held-out artists. The second is a subjective model of thirteen scores assigned by a large language model (LLM) instructed to act as a museum curator, which explains 39.7% in sample and 38.4% out of sample. The third is a composite of 68 characteristics that pools both sets with artist-level summaries of each artist's auction record, and explains 72.5% in sample and 70.1% out of sample. By analogy with finance, we refer to the part of an artist's price that these characteristics leave unexplained as the artist's *alpha*. We further document that the quantitative factors are strongly correlated with one another, which we argue reflects the structure of the art world itself: museums, galleries, and curators tend to move in sync, converging on the same artists and works through overlapping networks of reputation and endorsement. The curatorial scores of the language model are more correlated still, for reasons we trace to the way such models represent a named artist.

This apparent coordination is precisely what fuels skepticism about the art market's legitimacy. At the 2012 Whitney Museum Biennial, German filmmaker Werner Herzog ironically criticized the contemporary art market:

> *"What makes me particularly suspicious about the art market is that it seems to operate like a criminal scam as to how prices are fixed: auction houses, galleries, artists, magazines, and museum curators all appear to be part of a deep conspiracy, without ever having contacted one another. That's the strangest part of all."*

Some have taken this suspicion further still, alleging outright collusion. In 2018 the artist Robert Cenedella, known as the "Art Bastard", brought a $100 million antitrust action against five New York museums, alleging that they conspired with galleries, auction houses and collectors to raise the prices of a select group of artists through their exhibition choices (Cenedella v. Metropolitan Museum of Art, 2018). The complaint cited evidence that between 2007 and 2013 nearly one-third of solo museum exhibitions in the United States featured artists represented by one of five commercial galleries, including eleven of the twelve major solo exhibitions at the Guggenheim, and pointed to the rise in the average price of a Mark Grotjahn painting from $322,000 to $1.2 million following his 2014 exhibition at the Museum of Modern Art. The court dismissed the action in December 2018, finding no adequately pleaded agreement between the museums. The outcome is, in its way, Herzog's point: the pattern of convergence is plain to see, while the conspiracy that would explain it cannot be found.

Taken together, these observations suggest that art prices respond systematically to identifiable factors, chief among them the institutional and commercial recognition con-

ferred by museums, galleries and auction houses, in ways that parallel, though do not perfectly mirror, asset pricing patterns observed in traditional financial markets. A natural concern with this interpretation is one of causality: it is possible that galleries and curators, anticipating which artists are likely to command higher prices, actively push those artists toward museum exhibitions in the first place. Under this alternative reading, recognition would be a symptom of anticipated market demand rather than an independent driver of it, a chicken and egg problem that our current empirical strategy cannot fully resolve. Further research using time series of factors and prices could help disentangle this sequencing, identifying whether an artist's ascent through the gallery ladder precedes museum recognition, or whether museum recognition itself is the event that triggers the subsequent rise in gallery standing and prices.

The introduction of language models as raters raises a further question. An LLM has read a large share of the critical, curatorial and journalistic record on which reputations rest, and might therefore be expected to summarise an artist's standing more completely than any set of counts. We organise the analysis around three hypotheses.

H1. A language model, prompted as an institutional curator, explains the cross-section of artist-level prices better than a hedonic model built on the quantitative public record.

H2. The subjective and quantitative information sets are partly linearly independent, so that each carries price-relevant information absent from the other. We seek to quantify the extent of this independence.

H3. A composite specification that pools the information sets explains prices better than either set alone.

We reject H1: on the common sample of 2,278 artists the quantitative model explains 49.3% of the variation in log price, against 40.3% for the subjective model. We find support for H2: of the price variation explained by the language model, 77% is shared with the public record and the remaining 23% is unique to it. We find strong support for H3: the composite raises out-of-sample explained variance from 48.2% to 70.1%.

Our contribution to the literature is threefold. First, we provide, to our knowledge, one of the most comprehensive artist-level factor models of art prices to date, estimated without any input derived from realised prices, and we show that a substantial and stable share of cross-sectional price variation is associated with observable characteristics of an artist's career. Second, we provide a systematic comparison between the curatorial judgement of a language model and the documentary record, quantify the overlap between the two, and offer an explanation, grounded in the literature on language-model representations, for why thirteen distinct curatorial questions collapse onto a single prestige axis. Third, we show that the dominant structure of the composite model admits a natural interpretation as legacy, combining the prestige an artist has accumulated with the length of the career and the breadth of the body of work across which it has been sustained.

2 Previous Research

2.1 Hedonic Pricing of Artworks

The empirical literature on art-market valuation has traditionally treated the individual artwork as the principal unit of analysis. Because artworks are heterogeneous goods

that differ simultaneously in artist, medium, size, subject matter, provenance, condition, date of execution, and sale venue, much of this literature has relied on hedonic regressions to estimate the implicit prices associated with observable characteristics. A parallel literature has used repeat-sales methods, which compare successive transactions of the same object and thereby control directly for time-invariant artwork characteristics, at the cost of discarding the large majority of works that do not reappear at auction (Ashenfelter and Graddy, 2003; Mei and Moses, 2002; Renneboog and Spaenjers, 2013; Li et al., 2022).

Early hedonic studies demonstrated that substantial variation in auction prices can be associated with systematic and observable characteristics of the sale and the work. Using 37,605 paintings by 60 Australian artists, Higgs and Worthington (2005) model prices as a function of artist identity and living status, dimensions, medium, auction house, and year of sale. They find, among other effects, price premia associated with particular artists, oil and acrylic works, and sales through major auction houses. Renneboog and Spaenjers (2013) apply a hedonic regression to over one million auction transactions, controlling for artist, medium, subject and attribution, and Li et al. (2022) extend the approach to almost three million transactions from 1957–2016 with a specification that adds provenance, exhibition history, literature, certification and seasonality.

In each of these studies the hedonic characteristics serve to purify a market-level return. They strip out compositional differences between what is sold in different periods, so that the residual time trend can be read as price appreciation and summarised by an annual return and volatility, as one would for a financial asset class. The price of an individual artwork or artist lies outside the object of study, and the findings are therefore not directly comparable to ours, since we seek to explain cross-sectional price variation between individual artists.

2.2 Reputation, Institutional Validation, and Status

A separate but closely related literature studies how the reputation and status of artists are formed. This work is particularly relevant to the present paper because it shifts attention away from the physical attributes of an individual object and toward the artist's position within a wider institutional and cultural system.

A foundational distinction is made by Lang and Lang (1988), who separate *recognition* from the more durable phenomenon of *renown*. Their analysis argues that long-run artistic reputation depends on preservation, advocacy, access to cultural archives, and the networks through which an artist's work continues to be encountered and interpreted, in addition to the qualities of individual works. In this view, canonical standing is a socially and institutionally mediated outcome that cannot be read directly from an artwork.

The importance of institutional gatekeepers has subsequently been demonstrated empirically. Braden (2009) follows the 308 artists represented in the 1913 Armory Show and examines which subsequently appear in major art-history textbooks. The number and type of exhibitions obtained at the Museum of Modern Art between 1929 and 1967 strongly increase the probability of later textbook inclusion. Museum recognition therefore functions as a potential component of the longer process through which an artist enters the historical canon, in addition to its role as contemporaneous exposure.

This institutional perspective is consistent with the sociological account of Beckert and Rössel (2013). They describe contemporary art valuation as taking place under fundamental uncertainty: because artistic "quality" cannot be measured with the precision available for many conventional goods, collectors rely on reputational signals supplied by galleries, curators, critics, dealers, journalists, collectors, and other recognised inter-

mediaries. Price formation is consequently linked to an intersubjective process in which artistic value is validated, communicated, and reinforced through institutions.

Large-scale network evidence provides further support for this interpretation. Using data covering roughly half a million artists, Fraiberger et al. (2018) construct networks based on artists' institutional exhibition histories. They show strong path dependence in artistic careers: early access to prestigious institutions is associated with substantially more favourable subsequent career trajectories. Institutional position is therefore related to the structure through which reputation and market success accumulate, beyond its role as a contemporaneous descriptor of an artist.

The study most closely related to our treatment of prestige is Braden and Teekens (2019). Using the exhibition histories of 1,148 artists across the Stedelijk Museum, Museum Boijmans Van Beuningen, and Van Abbemuseum from 1930–1989, the authors distinguish individual *reputation* from relational *status*. Reputation is proxied by solo exhibitions and art-historical book inclusion, whereas status is constructed from the prestige of the artists with whom a focal artist exhibits. Their results demonstrate an important distinction between institutional quantity and institutional position. The raw number of museum exhibitions by itself does not significantly predict auction prices in their baseline specification, whereas measures of reputation and, particularly, high-status network connections account for a substantial share of price variation. Their combined model reaches an adjusted R^2 of approximately 0.46. The result suggests that *where*, *with whom*, and *under what reputational context* an artist is exhibited may matter more than exhibition counts alone. Our results are consistent with this distinction: in the quantitative model of Section 4, the number of museums at which an artist has exhibited explains under 8% of log-price variation on its own, well below measures of gallery standing and sale venue.

Braden and Teekens (2020) extend this network perspective from market value to art-historical commemoration. They show that museum exhibitions construct symbolic relationships between artists and that the historical cohorts within which curators place an artist are associated with later commemoration. Artists who bridge different historical cohorts receive particularly strong subsequent recognition. These results are directly relevant to attempts to operationalise canonical standing: institutional recognition is embedded in a network of comparisons, historical associations, and curatorial classifications that a count of appearances cannot capture.

2.3 Recent Computational Approaches

The increasing availability of large auction and cultural datasets has allowed recent work to move beyond conventional static hedonic specifications. Park et al. (2024) distinguish an artist's time-varying popularity from latent artistic ability, arguing that traditional fixed reputation variables cannot adequately represent changes in an artist's standing over time. Their model derives popularity from historical sales information while allowing an artist-specific ability component to evolve, highlighting the broader need to separate different dimensions of what is often compressed into a single "artist reputation" variable.

Machine-learning methods have also been used to compare the importance of visual and social information. Lee et al. (2024) use XGBoost on contemporary art auction data and report that low-level visual characteristics alone explain only about 5.5% of test-set price variation, whereas a model based on artist and market level social variables explains approximately 73%. Their social variables include artist age, exhibitions, public and private acquisitions, auction-house reputation and market characteristics, but also

previous price levels, and the principal feature of the model is the size-weighted median price of the artist's ten previous works. Because the model is supplied with an artist's own past prices in order to predict the next one, the 73% figure embeds the persistence of prices themselves. The contribution of non-price social signals is likely to be a smaller, unreported fraction of it, and the figure is not directly comparable with our models, from which previous prices are deliberately withheld.

The literature on canonization itself is also becoming computational. Efthymiou et al. (2026) analyse 60,011 paintings by 942 artists using visual and textual methods to study the relationship between artistic originality and long-run canonical importance. Their findings provide recent evidence that canonization can be investigated quantitatively, while also demonstrating that its determinants need not coincide exactly with the determinants of contemporaneous auction value. In particular, innovation relative to an artist's stylistic context contributes information about later canonical standing.

2.4 Position of the Present Study

Taken together, the existing literature establishes three important facts. First, auction prices are systematically related to observable characteristics of the artwork, artist, transaction, and sale venue. Second, artistic reputation and institutional position are themselves structured phenomena: museum exhibitions, curatorial associations, historical scholarship, and connections to high-status institutions are associated with both market outcomes and long-run commemoration. Third, recent computational evidence indicates that social and reputational information can contain considerably more pricing information than easily measured visual characteristics.

However, these strands have generally been studied separately. The art-pricing literature typically predicts the value of individual works using hedonic characteristics, artist fixed effects, previous sales, or transaction information. The reputation and canonization literature instead examines institutional careers, networks, and long-run cultural recognition, often without constructing a directly comparable pricing model. Consequently, there remains scope for an artist-level framework in which observable institutional and commercial variables and less directly measurable qualitative reputation are placed on the same statistical footing and evaluated against the same price cross-section.

The present study addresses this intersection. We make the artist the unit of analysis and construct explicit measures of the career and institutional characteristics that may underlie what is usually absorbed into an artist fixed effect. We first form a quantitative factor set from observable public records, including museum recognition, gallery representation, exhibition activity, geographic reach, and secondary-market activity. We then construct a separate qualitative assessment of artistic and institutional standing and examine whether its highly correlated components can be represented by a common latent prestige dimension. Finally, we compare the two information sets and combine them, together with a third set summarising the non-price characteristics of each artist's auction record, in a composite model.

A further distinction from several recent predictive studies is that realised auction prices, previous artist price levels, and variables mechanically constructed from the dependent variable are excluded from the explanatory factor set. The objective is to ask how much of the cross-sectional variation in artist-level auction values can be associated with independently measured career, institutional, market-activity, and qualitative reputational characteristics. We interpret these relationships as explanatory associations. They are not evidence that institutional recognition causes prices, nor that the qualita-

tive model directly observes future canonization; the latter is treated as a measurable proxy for the publicly documented cultural and institutional prestige that prior research associates with durable artistic recognition.

3 Data and Methodology

3.1 Data Collection

The data collection process was conducted in two stages. First, a broad sample of auction houses operating internationally was identified and ranked according to the average realised price of artworks sold. These averages were estimated by reviewing publicly available sales data on each auction house's website and recording a sample of completed transactions. The auction houses with the highest estimated average sale prices were then classified as higher-prestige institutions and selected as the principal sources for the second stage of data collection. Individual auction results from these houses were examined to identify artists and obtain artwork-level observations, including realised sale price, dimensions, medium, and other relevant characteristics. This process also produced a broader list of artists for subsequent evaluation. The sample was then extended by a network approach: starting from these houses and artists as initial nodes, we followed their connections to further artists, galleries and sales, so that the final panel represents a broad cross-section of the art world. All collected observations were entered into a structured spreadsheet to support data cleaning, comparison, and later statistical analysis. The resulting price panel contains 2,460 artists.

A key methodological issue we encountered is that prompting a large language model does not yield consistent results across different batch sizes. Specifically, the same artist received different scores depending on whether it was the only name in the request or one of 10, 100, or up to 1,000 names rated in a single prompt. Batch size therefore introduces a source of variability that must be accounted for when interpreting results. To remove it, every production score used in this paper is obtained from an isolated request that names a single artist, carries no conversation state, and is sampled at temperature zero. Section 5 documents the size of the batch effect and the design of the isolated calls.

3.2 Setup

Unit of analysis and outcome. Each observation is an artist. Museum records and the LLM rubric are artist-level, and the price is constructed the same way: we take the average realised auction price over the artist's recorded lots, scaled to a common size of one square metre ($100 \times 100\,\text{cm}$, USD per m^2). Averaging ensures that no single sale defines the career, and differences between works by the same artist (medium, colour, a trophy lot, early work selling above later work) are left in the residual. Area scaling places works of different sizes on a common basis. Because price rises less than proportionally with size, larger works carry lower prices per unit area, a regularity that the composite model of Section 7 recovers. The main specifications use the natural log of the average, and it is this cross-artist variation that we seek to explain. Lot-level minima and maxima are used neither as outcomes nor as inputs.

Prices are excluded from the inputs. Realised prices, and any statistic built from the level of the artist's own hammer prices (the mean, the minimum or the maximum), are left out of every specification, since those prices already form the target. Counts of

lots, sell-through, and the activity of the secondary market are activity measures and may enter the explanatory set. The LLM scores are produced from the artist's name and a fixed rubric. The model is shown no other artists, no other scores, and no prices.

Information sets. The first set is countable public record: museum exhibitions and their ranks, gallery representation, fairs, countries shown, career length, exhibition intensity, and related activity. The second set is evaluative: thirteen institutional scores from a language model acting as a curator, with no access to the price file. The third, used only in the composite model, summarises the non-price characteristics of each artist's auction record: format, medium, chronology, venue and region. All three are asked to account for the same cross-section of log prices. A high R^2 denotes an association and carries no claim that museums or the model set the price.

Evaluation. In-sample R^2 on the full artist panel is the headline number. It is always reported alongside repeated random holdouts (10% of artists, refit on the rest), so that a flexible model is not credited for fitting particular names. Reported coefficients come from one full-sample fit; parameters are not averaged across splits. Many of the variables move together (museums, galleries and curators backing the same names; LLM categories collapsing onto one prestige axis), so we report both single-variable associations and joint fits, and a coefficient in the joint model is not read as the isolated effect of that category.

3.3 Hedonic Specification

A hedonic model explains a good's price through the characteristics it carries. In our case, the good is a standardised artwork by an artist, and its characteristics are those of the artist. These characteristics are captured by a set of factors, such as museum exhibitions, gallery representation, or a curator-style rating, each entering with an estimated weight. Adding the weighted factors to a common intercept gives the model's estimate of the price, and whatever the factors cannot account for is left over as an error:

$$\ln(P_i) = \beta_0 + \beta_1 F_{i1} + \beta_2 F_{i2} + \cdots + \beta_n F_{in} + \varepsilon_i, \tag{1}$$

where:

- P_i is the average price per square metre of artworks by artist i;
- F_{ik} is the standardised value of factor k for artist i;
- β_0 is the intercept, equal to the mean log price because every factor has mean zero. Its estimate of 9.44 corresponds to a geometric-mean price of roughly \$12,500 per square metre;
- β_k is the coefficient on factor k;
- ε_i is the error term: the part of the price the factors do not capture, whether idiosyncratic (hype, luck, a single record sale) or value not captured by any factor. We refer to the estimated residual as the artist's alpha, recognising that in a single cross-section these two components cannot be separated.

We model the natural log of the price rather than the price itself. Prices range from a few hundred dollars to millions, and on the log scale a difference of a fixed *percentage* counts the same at every price level. This prevents a handful of very expensive artists

from dominating the results. The same specification, estimated by ordinary least squares, is used for the quantitative, subjective and composite models.

The remainder of the paper proceeds as follows. Sections 4 and 5 estimate the quantitative and subjective models. Section 6 compares them and tests H1 and H2. Section 7 presents the composite model and tests H3, and Section 8 concludes.

4 Quantitative Factor Model

4.1 Factors and Sample

The quantitative model is constructed from observable public-record measures of artists' institutional recognition, gallery representation, exhibition history, geographic reach, career duration, and secondary-market activity. All subsequent quantitative results use the same 13-variable specification: reciprocal gallery score together with twelve additional public-record measures, including museum and exhibition variables, geographic measures, career characteristics, auction-market activity, and the share of lots offered through Christie's, Sotheby's, or Phillips. Table 1 defines the thirteen variables, ordered by univariate explanatory power.

An inner join of the 2,460-artist price panel with the institutional dataset yields a final quantitative-model sample of $n = 2{,}278$. Eight documented artist-name aliases are resolved during the merge, while 182 artists in the price panel have no corresponding institutional record and are therefore excluded from this specification.

Gallery representation is measured using a reciprocal-rank score,

$$G_i = \sum_{g \in \mathcal{G}_i} \frac{w_{ig}}{r_g},$$

where $\mathcal{G}_i$ is the set of ranked galleries that have represented or exhibited artist i, r_g denotes the rank of gallery g, with lower numerical ranks corresponding to more prestigious galleries, and w_{ig} weights the strength of the artist's association with that gallery.

The art-fair variable measures the number of recorded fair appearances, without weighting by the prestige of the fair. International presence is measured by the number of countries in which an artist has exhibited, while regional presence is measured by the number of cities. Career length records the duration of the artist's documented exhibition career, and exhibitions per year captures exhibition intensity relative to that career length.

The remaining variables capture secondary-market activity and sale venue. These include sell-through rate, hammer performance relative to the high estimate, the number of lots offered, and the share of observations appearing at Christie's, Sotheby's, or Phillips (the "big-3" auction-house share). These measures characterise market activity and venue concentration and contain no information on the absolute level of the artist's realised auction prices.

Of the 2,278 artists in the joined sample, 159 have a reciprocal gallery score equal to zero, indicating that no ranked gallery is observed for those artists in the underlying extract. The reciprocal gallery distribution is strongly right-skewed. As part of the preprocessing procedure described below, 719 observations are clipped at the MAD-based upper fence, approximately 1,155, compared with a raw maximum of approximately 93,000.

No.	Factor	Description	β	R^2	Unique ΔR^2
1	Top auction houses	Share of lots at Christie's, Sotheby's and Phillips	0.829	0.259	0.076
2	Secondary market depth	Number of lots offered at auction	0.693	0.181	0.096
3	Gallery score	Reciprocal-rank score $\sum w/r$ over ranked galleries	0.687	0.178	0.001
4	International presence	Number of countries where the artist was exhibited	0.638	0.154	0.001
5	Exhibitions per year	Exhibitions / career length	0.609	0.140	0.000
6	Hammer vs. high estimate	Mean hammer price relative to the high estimate	0.569	0.123	0.066
7	Regional presence	Number of cities where the artist was exhibited	0.569	0.122	0.000
8	Exhibition score	Total exhibitions	0.544	0.112	0.000
9	Weighted museum recognition	Museum exhibitions weighted by museum rank	0.479	0.087	0.000
10	Art fair score	Number of art-fair appearances	0.468	0.083	0.002
11	Museum recognition	Number of museums with artist exhibitions	0.452	0.077	0.008
12	Career length	Years since first exhibition	0.199	0.015	0.002
13	Sell-through rate	Share of lots sold at auction	0.164	0.010	0.001

Table 1: Independent OLS of log average USD per m^2 on each quantitative factor ($n = 2{,}278$), ordered by univariate R^2. Each column is median-imputed, winsorised, and z-scored; β and R^2 are from the single-factor regression (β per one standard deviation). Unique ΔR^2 is the drop in in-sample R^2 of the joint 13-factor model when that column is removed.

4.2 Preprocessing and Evaluation

Each quantitative variable is first median-imputed. Extreme observations are then winsorised using the tighter of two thresholds: the empirical 1st and 99th percentile rank fences, and a MAD-based 1%/99% rule defined by

$$\text{median}(X) \pm 2.326\,\hat{\sigma}, \qquad \hat{\sigma} = 1.4826 \times \text{MAD}(X).$$

The resulting variables are standardised to zero mean and unit variance.

Because all regressors enter the model on a common scale, each reported coefficient can be interpreted as the change in log price associated with a one-standard-deviation increase in the corresponding factor. This also places the quantitative coefficients on a scale comparable to that used for the standardised qualitative scores.

Three forms of evidence are reported. First, each factor is entered separately in a univariate OLS regression of log price on that factor. Second, all thirteen factors are entered jointly in a multivariate OLS specification. Third, we report each variable's unique ΔR^2, defined as the reduction in full-model in-sample R^2 when that variable is removed

while all remaining variables are retained. Unique ΔR^2 is interpreted as a measure of incremental explanatory information within the chosen specification, and carries no causal interpretation.

Out-of-sample performance is evaluated using repeated random artist-level holdouts. For each split, 10% of the sample (228 artists) is assigned to the test set and the model is refitted on the remaining observations. We conduct 10,000 such splits using random seed 42 and record the resulting test-set R^2, MAE, and related performance statistics. Reported regression coefficients are estimated once on the full sample.

4.3 Univariate Associations

Table 1 reports a separate regression of log price on each of the thirteen quantitative factors.

Big-3 auction-house share is the strongest single correlate of artist-level log price, with $R^2 = 0.259$ and a standardised coefficient of $\beta = 0.829$. It is followed by the number of lots offered ($R^2 = 0.181$) and reciprocal gallery score ($R^2 = 0.178$). International presence, exhibitions per year, hammer performance relative to the high estimate, and city count form the next group of comparatively strong univariate associations, with R^2 values of approximately 0.12–0.15.

Museum recognition and weighted museum recognition exhibit more modest univariate explanatory power, with $R^2 = 0.077$ and $R^2 = 0.087$, respectively. Career length ($R^2 = 0.015$) and sell-through rate ($R^2 = 0.010$) have comparatively limited univariate explanatory power.

For scale, a one-standard-deviation increase in big-3 auction-house share is associated with an

$$e^{0.829} - 1 \approx 129\%$$

higher average price per unit area in the corresponding univariate regression. A one-standard-deviation increase in reciprocal gallery score is associated with an increase of approximately 99%. These magnitudes describe cross-sectional associations and should not be read as structural or causal effects.

At a sample size of $n = 2{,}278$, all thirteen univariate relationships are statistically detectable. Statistical significance, however, does not imply that the variables represent thirteen independent channels of price formation. As shown below, substantial portions of their explanatory information are shared.

The contrast between univariate and incremental explanatory power is particularly important. Although reciprocal gallery score explains 17.8% of cross-sectional log-price variation when considered alone, its unique contribution within the complete thirteen-factor model is only $\Delta R^2 = 0.001$. Exhibition counts, geographic reach, art-fair activity, and weighted museum recognition display similar reductions once related variables are included simultaneously. By contrast, the number of lots offered ($\Delta R^2 = 0.096$), big-3 auction-house share ($\Delta R^2 = 0.076$), and hammer performance relative to the high estimate ($\Delta R^2 = 0.066$) retain comparatively large unique contributions. The difference between their univariate and incremental results indicates substantial informational overlap among the institutional variables and a more distinct contribution from several measures of secondary-market activity.

4.4 Joint Fit

Taken together, the thirteen quantitative factors explain 49.3% of the cross-sectional variation in artist-level log average price. The adjusted coefficient of determination is

49.0%.

This performance remains stable under repeated holdout evaluation. Across 10,000 random 10% holdouts, the mean out-of-sample R^2 is 48.2%, with a median of 48.5%. The 5th and 95th percentiles are 41.0% and 54.7%, respectively. Mean out-of-sample MAE on log price is 0.90. The close correspondence between the full-sample and holdout results indicates that the reported explanatory power is a substantial and persistent cross-sectional signal, and owes little to the estimation of thirteen coefficients on 2,278 observations.

Figure 1 plots fitted against realised log prices for the full sample. Predicted values increase systematically with actual prices through the central region of the distribution, but occupy a substantially narrower range than the realised outcomes. The resulting compression of fitted values toward the cross-sectional mean indicates that the model captures a considerable component of relative artist-level price ranking while remaining less successful in reproducing observations in the extreme tails. Artwork-specific heterogeneity, unusually important individual lots, and other omitted artist- and transaction-level characteristics consequently remain in the residual.

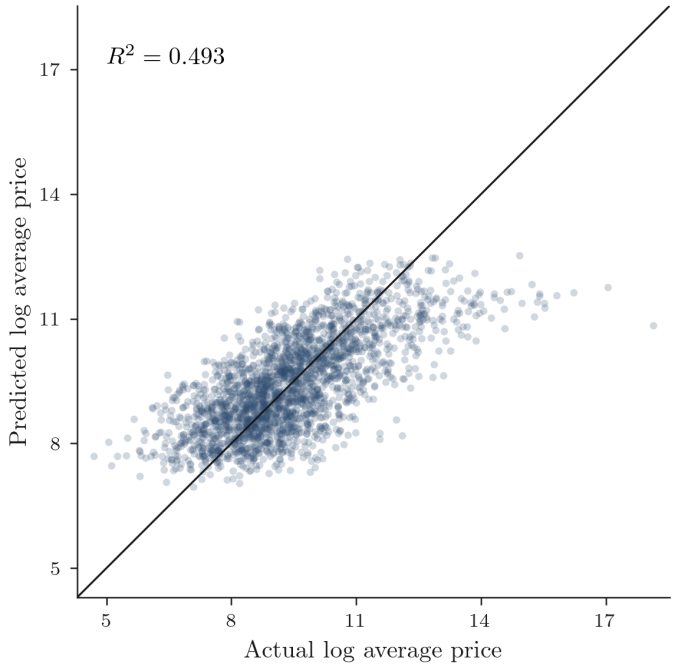


Figure 1: Full-sample OLS of log average USD per m^2 on the 13 quantitative factors ($n = 2{,}278$). The reference line is $y = \hat{y}$. Predicted values are less dispersed than realised values, indicating that the model captures substantial cross-sectional ranking information while reproducing the extreme tails less accurately.

4.5 Correlation Structure and Latent Dimensions

The thirteen quantitative variables encode several closely related groups of observable information, and their correlation structure makes these groups explicit. Figure 2 shows the Pearson correlation matrix of the standardised variables as they enter the OLS specification. The most pronounced cluster concerns exhibition activity and geographic reach. Total exhibitions and city count have a correlation of $r = 0.97$, while country count, art-fair participation, and exhibitions per year are also strongly associated with

this dimension. Reciprocal gallery score and big-3 auction-house share are positively related, while museum recognition and weighted museum recognition form another closely related pair.

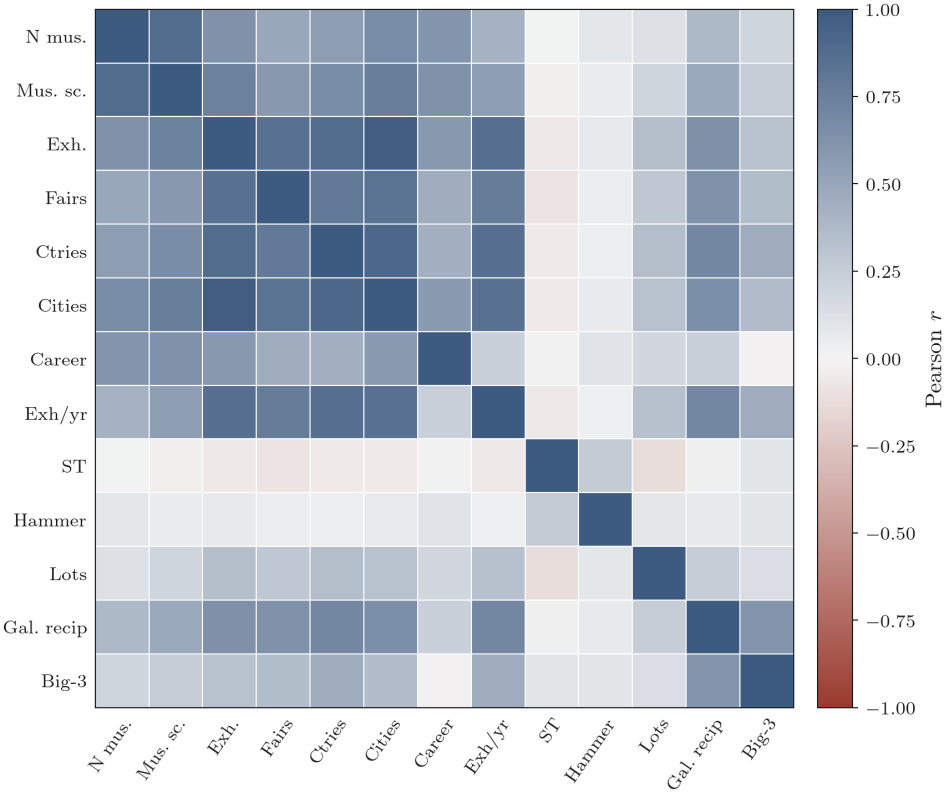


Figure 2: Pearson correlations among the 13 quantitative factors as they enter the OLS specification ($n = 2{,}278$). Exhibition activity and geographic reach form a strongly correlated block, while reciprocal gallery score is positively associated with big-3 auction-house share.

Principal-component analysis provides a complementary representation of this dependence structure (Figure 3). The first principal component accounts for 51.1% of total factor variance and loads most strongly on city count, total exhibitions, country count, art-fair participation, exhibitions per year and weighted museum recognition. We therefore interpret PC1 as a broad dimension of *exhibition activity and geographic reach*. The second component accounts for 11.5% of factor variance. Its loadings contrast career length and museum recognition with concentration in top commercial galleries and big-3 auction houses, and we interpret it as a dimension separating long institutional careers from contemporaneous commercial prestige. The third component (10.1%) is concentrated in sell-through rate and hammer performance relative to the high estimate, and describes auction clearance. The fourth (7.5%) is dominated by the number of lots offered and describes auction volume, while the fifth (5.2%) contrasts sell-through with hammer performance and big-3 share. Together, the five components account for 85.4% of the total variance across the thirteen quantitative variables.

A coefficient in the joint model therefore cannot be read as an isolated estimate of the value the market assigns to that characteristic. The correlation matrix, the principal components and the unique ΔR^2 analysis together indicate that the thirteen observed variables encode a smaller number of underlying dimensions of an artist's observable ca-

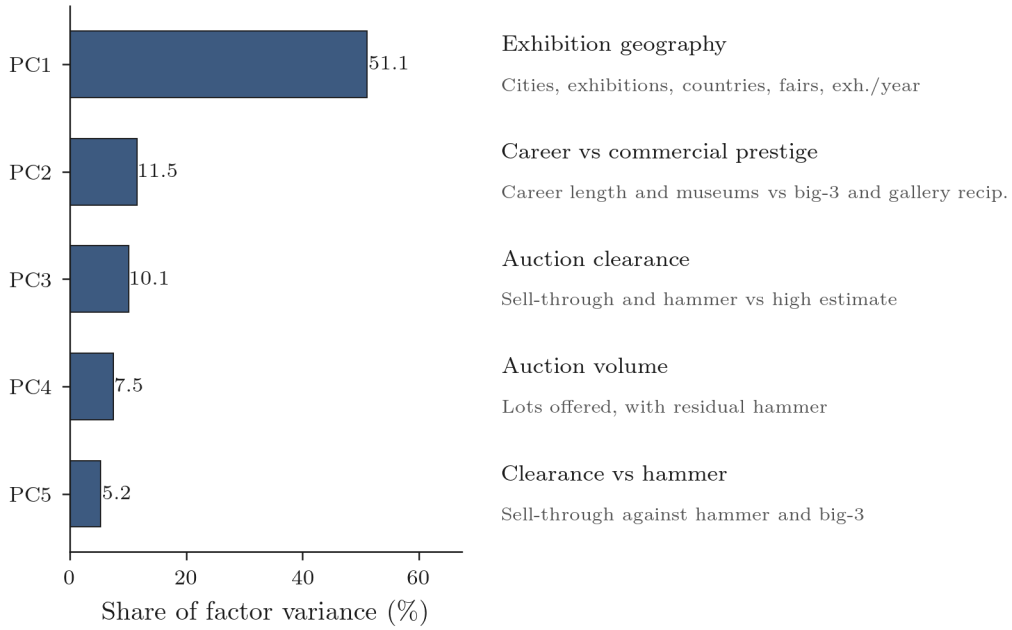


Figure 3: Correlation-based principal-component analysis of the 13 quantitative factors (n = 2,278), with the interpretation of each of the first five components. PC1 (51.1%) is dominated by exhibition activity and geographic reach; PC2 (11.5%) contrasts long institutional careers with commercial prestige; PC3 to PC5 describe auction clearance, auction volume, and the balance between clearance and hammer performance.

reer, most prominently exhibition activity and geographic reach, commercial positioning, and secondary-market activity.

5 Subjective (LLM) Factor Model

5.1 Construction of the Scores

Prompt construction. Every call uses the same system message: an institutional ranking framework in which the model acts as a senior curator scoring a single artist for long-term museum acquisition. The rubric fixes thirteen categories (historical significance, medium innovation, contemporary relevance, underrepresented narratives, aesthetic quality, conceptual depth, market and institutional validation, collection building, conservation, educational engagement, critical reception, mission alignment, and ten-year canonical probability). Each category is scored 0–100 against mandatory bands (80–100 exceptional; 0–19 negligible evidence). The overall score used in the batch-size experiments is the mean of the thirteen scores. The user message names exactly one artist and forbids mentioning any other. The model scores each artist from its own general knowledge, and no price data are supplied in either message.

Model. Institutional scores are produced with Anthropic's Claude Sonnet 4.6 (`claude-sonnet-4-6`) via the Messages API. Sonnet offers the most favourable balance of cost and quality for this task: it follows a long fixed rubric and has usable coverage of contemporary and historical artists, at approximately \$3 / \$15 per million input / output tokens. Opus is several times more expensive for little expected gain on a constrained 0–100 scoring task.

Haiku would reduce cost further but was not used, because weaker art-historical recall would appear as conservative default scores on less famous names. With prompt caching of the rubric, isolated calls cost on the order of $0.004 per artist. The full production run of 2,868 names, of which the 2,460 artists of the price panel are used for estimation, cost $11.77.

Independence. Multi-turn chat interfaces are unsuitable for this step. Chat history would allow earlier scores to leak into later ones, and such interfaces do not expose temperature. Each rating is therefore a fresh HTTP request with no conversation state: the system message is the full institutional rubric; the user message contains exactly one artist name and an instruction not to mention any other artist; previous names and scores are never included. Calls are sequential (or parallel with disjoint payloads), and no context is shared between artists.

Temperature. Temperature scales the randomness of next-token sampling. At zero the model is as deterministic as the API allows, which is what a fixed rubric requires. All ratings in this paper use `temperature = 0`. Higher values would add noise unrelated to the artist.

Batch-size sensitivity. Placing many names in one prompt is cheaper, but it reintroduces the contamination that isolated calls are designed to avoid. We re-rated the same 100 artists at temperature 0 under six regimes: one name per call; ten names per call (ten calls); and 100, 200, 500 and 1,000 names in a single call (200 and above padded with unrelated names, after which the original 100 were extracted). Table 2 compares each regime to the 1-by-1 reference on the overall score.

Names in call	r vs 1-by-1	MAE (pp)	Mean overall score (%)	Shift vs 1-by-1
1	1.000	0.00	63.16	0.00
10	0.918	3.38	60.58	−2.58
100	0.878	3.63	60.33	−2.84
200	0.880	5.30	58.16	−5.00
500	0.856	8.93	54.44	−8.72
1,000	0.843	7.34	56.03	−7.13

Table 2: Batch-size sensitivity. Same 100 artists, same model and rubric. Overall score relative to isolated 1-by-1 ratings.

Correlation remains high, so coarse rankings are persistent. The *level* of the scores is less stable: larger contexts pull scores down by several percentage points, from a mean of 63.2% in isolation to 60.6% with ten names and 54.4% with 500, where the shift peaks (MAE 8.93 pp, shift −8.72 pp) before easing slightly at 1,000. A repeat of the 100-names-in-one-prompt regime reproduced the same pattern ($r = 0.896$, shift −3.41 pp). Production scores for the 2,460-artist regression sample therefore use only isolated 1-by-1 calls.

5.2 Estimation

The regression sample is 2,460 artists. The outcome is the natural log of average realised USD per m^2. The thirteen LLM scores are standardised (mean 0, SD 1) so that a coefficient is the change in log price associated with a one-standard-deviation move in that score. The reported specification is a single OLS fit on the full sample. We also

draw 10,000 random 10% holdouts (246 test artists, seed 42), refit on the complement, and record out-of-sample R^2, as a check that the fit is not an in-sample artefact.

5.3 Univariate Associations

Table 3 reports, for each score, a separate regression of log price on that score alone.

Table 3: Independent OLS of log average USD per m^2 on each standardised LLM score (n = 2,460), ordered by univariate R^2. Unique ΔR^2 is the drop in in-sample R^2 of the joint 13-score model when that score is removed. Bracketed numbers give the rubric order of Appendix **??**, as in Figure 5.

No.	**Factor**	**Description**	β	R^2	**Unique** ΔR^2
1	Probability of Becoming Canonical in 10 Years [13]	Likelihood of future canonical status based on documented indicators.	0.842	0.275	0.045
2	Market Position & Institutional Validation [7]	Institutional recognition, gallery representation and market positioning.	0.825	0.264	0.002
3	Critical Reception & Scholarly Attention [11]	Quality and persistence of critical and academic engagement.	0.770	0.230	0.005
4	Collection Building & Context [8]	Value for filling collection gaps and supporting curatorial narratives.	0.763	0.226	0.001
5	Educational & Public Engagement Value [10]	Usefulness for teaching, research and public interpretation.	0.752	0.220	0.001
6	Historical Significance & Art Historical Contribution [1]	Contribution to art history, artistic influence and canonical status.	0.738	0.212	0.008
7	Alignment with Institutional Mission & Values [12]	Fit with institutional objectives, education and long-term stewardship.	0.718	0.200	0.003
8	Aesthetic Quality & Visual Impact [5]	Formal coherence, visual strength and exhibition presence.	0.705	0.193	0.003
9	Contemporary Relevance & Urgency [3]	Engagement with current cultural, political, social or technological conditions.	0.688	0.184	0.002
10	Conceptual Depth & Intellectual Rigor [6]	Coherence, intellectual grounding and lasting interpretive depth.	0.662	0.170	0.000
11	Medium Innovation & Technical Excellence [2]	Technical mastery and innovation within or across media.	0.645	0.161	0.015

Continued on next page

Table 3 – continued from previous page

No.	Factor	Description	β	R^2	Unique ΔR^2
12	Sustainability & Conservation [9]	Durability, conservation risk and long-term display viability.	0.613	0.146	0.000
13	Representing Underrepresented Narratives [4]	Contribution to overlooked perspectives and institutional collection gaps.	0.467	0.085	0.001

Ten-year canonical probability is the strongest single associate ($R^2 = 0.275$, $\beta = 0.842$), followed by market and institutional validation (0.264), critical reception (0.230), collection building (0.226) and educational engagement (0.220). Historical significance lies in the middle of the distribution (0.212). Underrepresented narratives is last (0.085): the rubric explicitly states that this category alone must not drive a high overall score, and the price data confirm that it is the category least aligned with the auction cross-section.

A one-standard-deviation higher canonical-probability score is associated with an $e^{0.842} - 1 \approx 132\%$ higher average price per area in a univariate regression, which describes a ranking and should not be read as a structural effect. Every univariate p-value is vanishingly small because $n = 2{,}460$; the values establish a detectable association, and do not establish thirteen independent channels. Section 5.5 shows that the channels largely coincide.

5.4 Joint Fit

Taken together, the thirteen scores explain 39.7% of the cross-artist variation in log average price (adjusted $R^2 = 39.4\%$). Across the 10,000 holdouts the mean out-of-sample R^2 is 38.4% (5th to 95th percentiles 28.1% to 47.3%), so the in-sample figure owes little to the fitting of 13 coefficients on 2,460 observations.

Figure 4 shows the full-sample fit. The cloud follows the 45-degree line through the bulk of the sample, where actual log prices lie between roughly 8 and 12, but the predictions are compressed: actual log prices run from about 5 to 18, while fitted values remain within a band of roughly 8 to 12. Artists at the top of the price distribution are recognised as high-prestige and are nonetheless under-predicted, and inexpensive artists are drawn up toward the mean.

The compression is informative. An artist-level institutional rubric orders the middle of the market well, while work-level idiosyncrasy, the thinness of particular sales and right-tail events that lie outside prestige remain in the residual. The LLM scores constitute a ranking of institutional standing.

5.5 One Prestige Axis

The thirteen scores do not describe thirteen independent traits. Principal-component analysis on their correlation matrix places 89.4% of score variance on the first component, the same dependence that Figure 5 displays as a near-uniform block of high pairwise correlations (typically $r > 0.8$). The exception is underrepresented narratives (factor 4), which the rubric already treats as a corrective category. Conservation (9) is only slightly less aligned. Once the other twelve scores are in the joint model, dropping canonical probability costs 4.5 percentage points of R^2, and every other unique ΔR^2 is at most

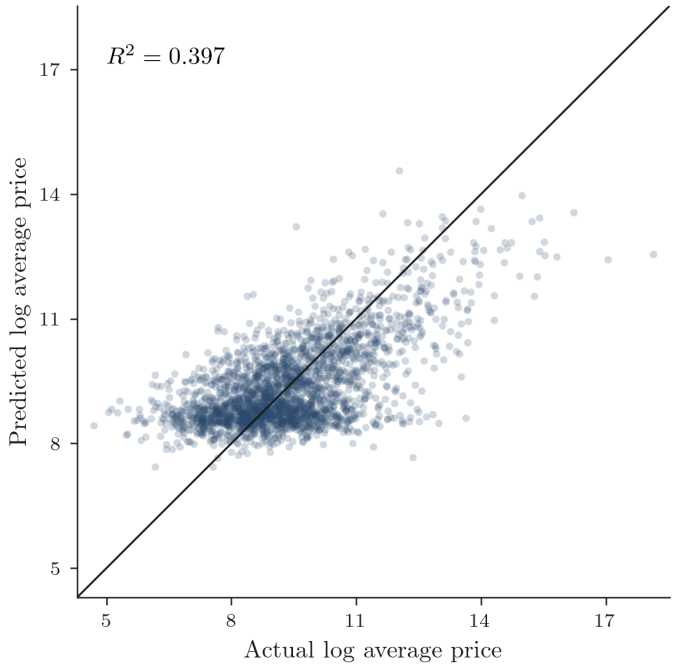


Figure 4: Full-sample OLS of log average USD per m^2 on the 13 standardised LLM scores ($n = 2{,}460$). The line is $y = \hat{y}$. Predictions occupy a much narrower band than realised values: the scores rank artists but do not reproduce the tails.

1.5 percentage points. Multivariate coefficients therefore provide no ranking of what the market pays for. Six of the thirteen scores, among them historical significance and critical reception, are positive in univariate regressions and turn negative in the joint fit, while the coefficient on canonical probability rises several-fold, because all are collinear with the same prestige factor. For this reason Table 3 is univariate. The axis reflects genuine institutional standing, the same convergence of museums, galleries and curators described in the introduction, and Section 5.6 argues that the way the model represents an artist compresses that standing into a near-single dimension. We refer to this first component as the *prestige axis*.

5.6 Why the Scores Share One Axis

The dominance of the first component is a property of the instrument as much as of the artists, and three strands of research on language models bear on it. First, word and contextual representations in neural language models are strongly anisotropic: they occupy a narrow region of the embedding space and are dominated by a small number of common directions (Mu and Viswanath, 2018; Ethayarajh, 2019). Second, factual knowledge about a named entity is concentrated in the representation of the entity's own tokens. Feed-forward layers act as key-value memories that are queried by the subject (Geva et al., 2021), and interventions on the middle-layer representation of the subject alter what the model asserts about it across a wide range of prompts (Meng et al., 2022). The model's knowledge of an artist is therefore retrieved through a single representation keyed to the name, in the sense in which language models have been described as knowledge bases (Petroni et al., 2019). Third, the reliability of what is recalled scales with the frequency of the entity in the pre-training data, and accuracy degrades sharply for less

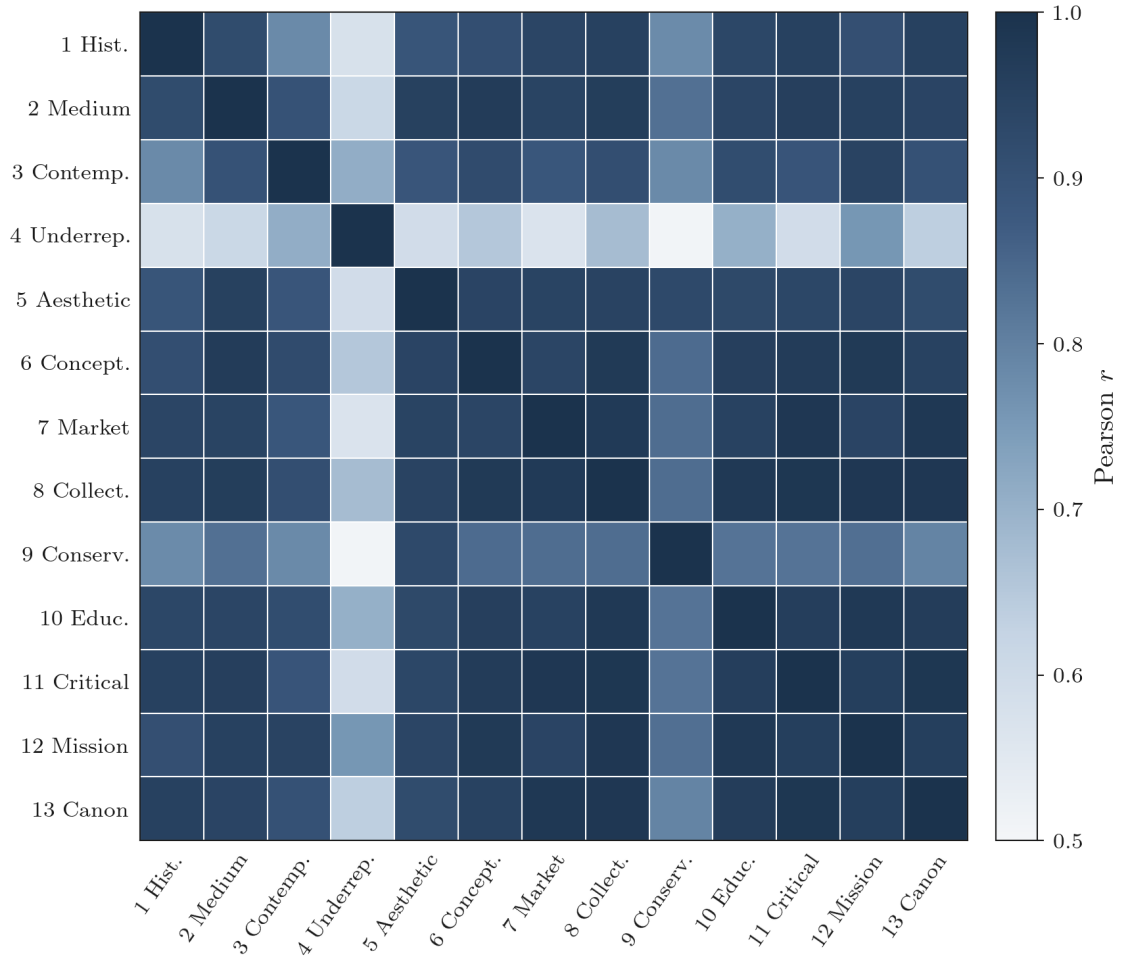


Figure 5: Pearson correlations among the 13 standardised LLM scores ($n = 2{,}460$). The diagonal is 1 and the colour scale starts at 0.5. Factor numbers follow the rubric order in Appendix **??**. Underrepresented narratives (4) is the light cross; the remaining twelve scores form a single block.

frequently discussed entities (Mallen et al., 2023; Kandpal et al., 2023).

Read together, these results suggest a mechanism for the structure in Figure 5. Each of the thirteen categories conditions on the same name, and the category enters as a comparatively small perturbation of a representation dominated by the artist. Where the model holds rich and differentiated knowledge, the category can move the score. For most artists in the sample the representation carries little beyond a general level of standing, a compressed summary of how widely and how favourably the name is discussed, and the model expresses that summary thirteen times. Asked about aesthetic quality and about cultural impact, it returns essentially the same judgement, because both questions retrieve the same representation of the name. The behaviour is a computational analogue of the halo effect documented in human raters, in whom a global impression of a person colours judgements of conceptually distinct attributes (Thorndike, 1920; Nisbett and Wilson, 1977), and it is consistent with the systematic biases reported for language models used as evaluators (Zheng et al., 2023). Underrepresented narratives, the one category that departs from the axis, is also the one whose rubric criterion (geography, origin, and representation in collections) can be answered in part from biographical fact, independently of reputation.

The mechanism carries a testable implication: where the model knows more about an artist, its thirteen answers should diverge more. The data bear this out. Dividing the 2,278 artists with a quantitative record into thirds by the number of recorded exhibitions, the first component accounts for 88.6% and 89.4% of score variance among the least and moderately documented artists, and falls to 79.9% among the most documented. The result carries a practical lesson for the use of language models as raters. When a model is asked a long series of questions about a single named entity, the number of questions overstates the number of independent measurements obtained: in our case, thirteen scores carry approximately the information of one. Such a rubric measures prestige consistently and at low cost, and its value lies there. Separating attributes such as aesthetic quality from cultural impact requires that the distinguishing evidence be supplied in the prompt, or a design in which artists are compared with one another on a single attribute at a time.

6 Comparing the Subjective and Quantitative Models

6.1 Explanatory Power (H1)

To compare like with like, we re-estimate the subjective model on the 2,278 artists for which the quantitative record is available, with the LLM scores re-standardised on that sample. On this common sample the thirteen LLM scores explain 40.3% of the variation in log price, against 49.3% for the thirteen quantitative factors. Out of sample, on identical holdout splits, the quantitative model retains 48.2% and the subjective model 38.8%, and the quantitative model is the stronger of the two in 95.5% of the 10,000 splits. The public record explains roughly nine percentage points more of the cross-section than the curatorial judgement of the language model, with an equal number of factors and on an identical outcome. H1 is therefore rejected. A model with access to a large share of the written record on contemporary art, prompted to reason as a curator, is less informative about relative prices than a set of measures of exhibitions, galleries and auction activity. The comparison has a striking practical counterpart. The language model recovers about 82% of the explanatory power of the quantitative model (40.3% against 49.3%) from the artist's name alone, at a total cost of $11.77 for the entire panel, whereas the public record requires the assembly of exhibition, gallery and auction histories artist by artist.

6.2 Linear Independence (H2)

The two information sets are related, and the question is by how much. We quantify their overlap with a commonality analysis of the explained variance (Kraha et al., 2012). Let R^2_L and R^2_Q denote the in-sample explained variance of the subjective and quantitative models, and R^2_{LQ} that of the union of the two sets on the same sample. The variance explained uniquely by each set, and the variance they explain in common, are then

$$U_L = R^2_{LQ} - R^2_Q, \qquad U_Q = R^2_{LQ} - R^2_L, \qquad C = R^2_L + R^2_Q - R^2_{LQ}.$$

Table 4 reports the partition.

Component	Share of log-price variance (%)	Share of explained variance (%)
Unique to the subjective (LLM) set, U_L	9.26	15.8
Unique to the quantitative set, U_Q	18.25	31.2
Common to both sets, C	31.06	53.0
Total explained by the two sets, R^2_{LQ}	58.57	100.0

Table 4: Commonality analysis of the subjective and quantitative information sets on the common sample ($n = 2{,}278$). $R^2_L = 40.3\%$ and $R^2_Q = 49.3\%$.

Of the 40.3% of price variation explained by the language model, 31.1 percentage points, or 77%, are also explained by the public record; the remaining 23% is linearly independent of it. Of the 49.3% explained by the quantitative factors, 63% is shared with the language model and 37% is unique. The overlap is visible directly in the factors. The first canonical correlation between the two sets of thirteen variables is 0.86 (Hotelling, 1936), and a regression of the LLM prestige axis on the thirteen quantitative factors explains 57.6% of its variance. Across individual scores, the share explained by the public record ranges from 19.1% for underrepresented narratives to 66.3% for critical reception, with a mean of 53.4%. The two fitted price indices correlate at 0.60.

H2 is therefore supported, with a precise answer to the question of degree. The two sets are strongly but imperfectly collinear: roughly three quarters of what the language model knows about prices is already contained in the documentary record, and roughly a quarter is not. The independent component is economically meaningful, at 9.3 percentage points of log-price variance, and it suggests that the model encodes part of the critical and art-historical reception of an artist that counts of exhibitions and lots do not reach. The residuals of the two models, however, correlate at 0.68. Both models miss largely the same artists in the same direction, which indicates that a substantial part of the unexplained variation lies outside both information sets. The residual is, in this sense, a shared alpha: a component of price that neither the curator nor the documentary record sees. It motivates the composite model of the next section, which adds a third and largely orthogonal source of information.

7 Composite Model

7.1 Construction

The composite model is a single linear specification of 68 characteristics on the 2,278 artists of the common sample: the thirteen LLM scores, the thirteen quantitative factors, and 42 characteristics of each artist's auction record. The specification, preprocessing and evaluation follow Sections 3.3 and 4 without modification, and the estimator remains

ordinary least squares with an intercept. The composite thus combines the information sets at the level of the regressors.

The 42 auction-record characteristics are generated from the same lot-level sale file as the outcome, from which every price field (hammer price, estimates and premium) is removed before aggregation. The remaining lot attributes are rolled up to the artist and summarise what has been offered, when, and where. They comprise biography (birth year and living status); the volume and breadth of the record (the number of lots and the number of distinct houses, sale locations and supports); format (mean, median and log mean area, mean aspect ratio, and the share of lots with parsed dimensions); the chronology of production (mean, earliest and latest year of creation, the span of creation dates, and the share of lots without a date); the chronology of sale (mean, earliest and latest sale year, the span of sale years, and the years since the most recent sale); medium and support (the shares of untitled, oil, acrylic, spray, mixed-media, canvas, linen and panel works); venue (the shares of lots at Christie's, Sotheby's, Phillips, Bonhams and the principal Asian houses); and region of sale (the shares in the United States, the United Kingdom, Hong Kong, Korea, Japan, South Africa, continental Europe and elsewhere). A log transform of the lot count and three aggregates that duplicate quantitative factors (lots offered, sell-through and big-3 share) are excluded.

With 68 parameters and 2,278 observations, the model retains about 33 observations per parameter. The composite is a deliberately broad description of the artist and gives up the interpretability of the smaller models; its purpose is to measure how much of the price cross-section the combined information can account for, and to reveal the structure that organises it.

7.2 Explanatory Power (H3)

The composite explains 72.5% of the cross-sectional variation in log price (adjusted $R^2 = 71.7\%$). Across 10,000 random holdouts of 10% of artists, the mean out-of-sample R^2 is 70.1%, with 5th and 95th percentiles of 63.9% and 75.5%, and the mean out-of-sample MAE on log price falls to 0.68, against 0.90 for the quantitative model. The gap between in-sample and out-of-sample fit widens from 1.1 to 2.4 percentage points, a modest price for a fivefold increase in the number of regressors. Figure 6 shows the full-sample fit. The compression that characterises the two smaller models is markedly reduced: the cloud now extends along the 45-degree line into both tails, and the model reproduces a substantially wider range of prices.

Figure 7 places the three models side by side. Relative to the stronger of the two smaller models, the composite adds 23.2 percentage points in sample and 21.9 percentage points out of sample, and the improvement is far larger than the difference between the two smaller models themselves. H3 is therefore supported. On identical holdout splits the composite outperforms the quantitative model in every one of the 10,000 draws, by 21.9 percentage points on average and by no less than 6.0.

7.3 Sources of Incremental Information

Each information set retains a distinct contribution inside the composite. Removing the LLM scores reduces in-sample R^2 by 5.7 percentage points, removing the quantitative factors by 6.6, and removing the auction-record characteristics by 13.9. The unique contributions of the three blocks sum to 26.3 percentage points, and the remaining 46.2 points are shared between two or more of them. The auction-record block is the largest single source of new information, and on its own it explains 55.0% of log-price variation,

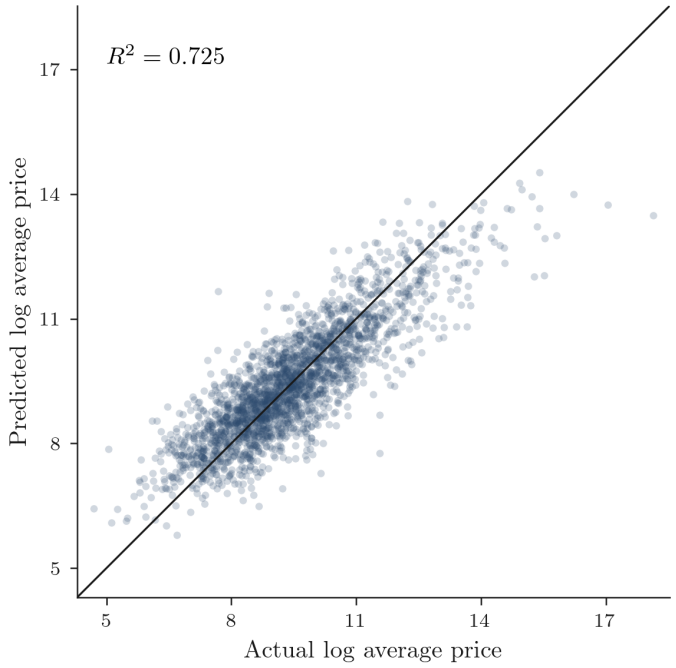


Figure 6: Full-sample OLS of log average USD per m^2 on the 68 composite characteristics ($n = 2{,}278$). The reference line is $y = \hat{y}$.

more than either the subjective or the quantitative model. Part of this strength reflects the composition of what an artist brings to auction: format, medium, the period of the work and the venue are themselves standard hedonic determinants of price (Higgs and Worthington, 2005; Renneboog and Spaenjers, 2013), and an artist-level average inherits the mix.

Among individual characteristics, hammer performance relative to the high estimate leads (4.4 percentage points), followed by ten-year canonical probability (1.4), the number of lots offered (0.8) and critical reception (0.8). Two of the four leading characteristics are LLM scores, and the subjective set retains a distinct role once the documentary record is controlled for. Format follows, with the log mean and median area of works offered. Because the outcome is price per unit area, these terms capture the well-documented concavity of price in size, by which larger works command lower prices per square metre. Living status, the latest year of creation, museum recognition and the reciprocal gallery score complete the leading group. The big-3 share, the strongest univariate correlate in the quantitative model, retains a unique contribution of only 0.01 percentage points, because the individual shares of Christie's, Sotheby's and Phillips enter the auction-record block and absorb its information.

7.4 Latent Structure

Figure 8 reports the principal-component structure of the 68 characteristics. The first component accounts for 26.9% of their joint variance and loads almost uniformly on the LLM scores (critical reception, market and institutional validation, collection building, canonical probability, conceptual depth and medium innovation, each with a loading of about 0.22), together with the quantitative measures of institutional standing that move with them. It is the prestige axis of Section 5.5, recovered from a far broader set of

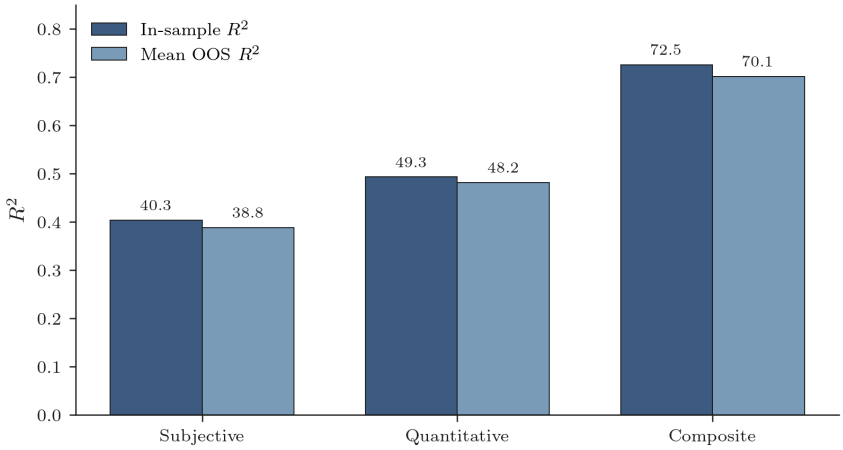


Figure 7: In-sample and mean out-of-sample R^2 of the subjective (13 LLM scores), quantitative (13 public-record factors) and composite (68 characteristics) models. All values are estimated on the common sample of 2,278 artists; out-of-sample values are means over the same 10,000 random 10% holdouts.

characteristics, and on its own it is associated with 24.7% of the cross-sectional variation in log price. The second component accounts for a further 9.0% and orders artists by generation and working life. Artists who score highly on it were born earlier and produced the works now at auction earlier, as measured by birth year and the earliest, mean and latest years of creation. They also have longer exhibition careers and a wider span of production that has reached auction. The third to fifth components describe auction recency and volume, the Asian sale circuit, and format and region of sale. The structure of the composite is considerably less concentrated than that of either smaller set. The first component carries 89.4% of the variance of the LLM scores and the first five carry 85.4% of the variance of the quantitative factors, whereas the first five components of the composite account for 51.2%, and 27 are required to reach 90%. The composite is therefore a genuinely richer description of the artist, and its additional columns contribute dimensions of their own.

The first two components admit a joint reading. The first measures how highly an artist is regarded; the second, how long and across how many bodies of work that regard has been sustained. Together they account for 35.9% of the variation in the 68 characteristics, and we interpret them jointly as *legacy*. Section 8 returns to this interpretation.

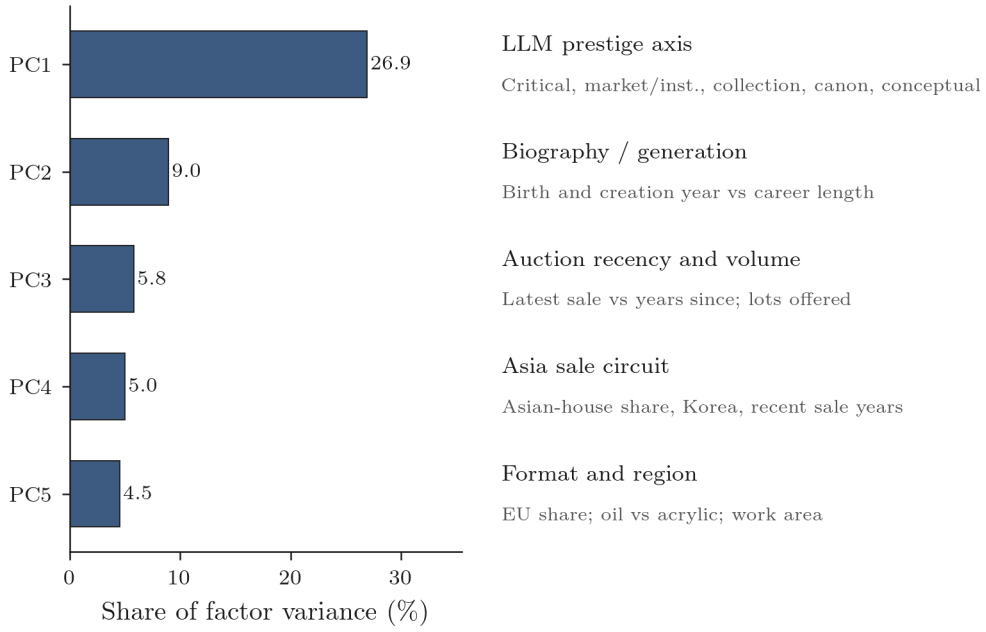


Figure 8: Correlation-based principal-component analysis of the 68 composite characteristics (n = 2,278), with the interpretation of each of the first five components. PC1 (26.9%) is the LLM prestige axis; PC2 (9.0%) orders artists by generation, career length and the span of their body of work.

8 Conclusion

This paper set out to measure how much of the cross-section of artist-level auction prices can be explained by characteristics recorded without reference to price. The first hypothesis is rejected: thirteen measures of the public record explain 49.3% of the variation in log price, against 40.3% for thirteen curatorial scores of a language model on the same artists. The second is supported, since 23% of what the language model explains is linearly independent of the public record. The third is strongly supported: a composite of 68 characteristics explains 72.5% of log-price variation in sample and 70.1% on held-out artists.

The structure of the composite also offers a compact account of what the market prices. Its first principal component is the prestige axis of the language model, joined by the institutional measures that move with it, and its second orders artists by generation, by the length of their careers and by the span of the work they have produced. The first describes the standing an artist has attained; the second, the time and the breadth of production over which that standing has been sustained. Together they describe legacy: prestige that has persisted across a career, outlived the generation that produced it, and extended across many bodies of work. The pairing recovers the distinction drawn by Lang and Lang (1988) between recognition and the more durable phenomenon of renown, with the first component measuring the recognition an artist has attained and the second its persistence over time, so that legacy in our sense is renown, measured. Collectors, dealers and critics routinely invoke legacy as the ultimate source of value in art. The composite gives that intuition a measurable form, assembled from a qualitative judgement and a documentary record that were constructed independently of one another, and it is notable that the two converge on the same organising structure. The two components account for

35.9% of the joint variation in the characteristics, and the remaining dimensions, among them auction recency and volume, the sale circuit and the format of the work, describe the conditions under which legacy is sold.

A methodological finding accompanies these results. The thirteen curatorial scores of the language model share 89.4% of their variance on a single component. We attribute this to the way such models retrieve knowledge of a named entity, through a representation keyed to the name and dominated by a general level of standing, so that questions about distinct attributes of an artist return essentially one judgement. Consistent with this account, the concentration falls to 79.9% for the best-documented artists, against about 89% for the rest. Researchers who use language models as raters should accordingly treat a battery of entity-level questions as a single measurement of reputation, and should supply the distinguishing evidence directly where finer distinctions are required.

Several limitations qualify these conclusions. The analysis is cross-sectional and associational, and cannot resolve whether recognition precedes price or anticipated demand precedes recognition. Valuing an artist differs from valuing an artwork: the premium on early work, the colour, subject and emotional force of an individual piece, and the size of an artist's total output are absent from an artist-level average. Concentrated ownership is also unobserved. A dealer who acquires a sufficient stake in an artist's production can act as a market maker and has a direct incentive to promote the artist through gallery placement, museum exhibitions and press coverage. The Los Angeles dealer Stefan Simchowitz, for example, reported in 2014 that he owned 34 paintings by Oscar Murillo, some acquired early in the artist's career for as little as $1,500 each, while Murillo's works were fetching several hundred thousand dollars at auction (Kazakina, 2014). Further research using panels of factors and prices could establish the sequence in which recognition and price move, and work-level data could separate the legacy of the artist from the qualities of the individual work.

Roughly 28% of the variation in log price remains unexplained. It is the artist's alpha: the part of value that legacy, the record and the conditions of sale do not account for. A positive alpha means price has run ahead of what the artist's legacy supports; a negative alpha means price lags behind it. Whether such gaps reflect noise, characteristics not yet measured, or mispricing that price discovery will eventually correct over time cannot be settled in a single cross-section, which makes alpha the natural object for future work on which artists the market over- and under-values.

References

McAndrew, C. (2025). *The Art Basel and UBS Global Art Market Report 2025*. Basel: Art Basel and UBS. https://theartmarket.artbasel.com.

Ashenfelter, O. and Graddy, K. (2003). "Auctions and the Price of Art." *Journal of Economic Literature*, 41(3), 763–787. https://doi.org/10.1257/002205103322436188.

Beckert, J. and Rössel, J. (2013). "The Price of Art: Uncertainty and Reputation in the Art Field." *European Societies*, 15(2), 178–195. https://doi.org/10.1080/14616696.2013.767923.

Braden, L.E.A. (2009). "From the Armory to Academia: Careers and Reputations of Early Modern Artists in the United States." *Poetics*, 37(5–6), 439–455. https://doi.org/10.1016/j.poetic.2009.09.004.

Braden, L.E.A. and Teekens, T. (2019). "Reputation, Status Networks, and the Art Market." *Arts*, 8(3), 81. https://doi.org/10.3390/arts8030081.

Braden, L.E.A. and Teekens, T. (2020). "Historic Networks and Commemoration: Connections Created Through Museum Exhibitions." *Poetics*, 81, 101446. https://doi.org/10.1016/j.poetic.2020.101446.

Cenedella v. Metropolitan Museum of Art, 348 F. Supp. 3d 346 (S.D.N.Y. 2018).

Cochrane, J.H. (2011). "Presidential Address: Discount Rates." *Journal of Finance*, 66(4), 1047–1108. https://doi.org/10.1111/j.1540-6261.2011.01671.x.

Deloitte Private and ArtTactic (2025). *Art & Finance Report 2025*, 9th edition. Luxembourg: Deloitte. https://arttactic.com/reports/deloitte-and-arttactic-or-art-and-finance-report-2025.

Efthymiou, A., Kackovic, M., Rudinac, S., Worring, M. and Wijnberg, N. (2026). "Being Ranked in a Material World: The Visual Originality of an Artwork and Its Effects on the Artist's Canonization." *Organization Studies*, 47(1), 93–125. https://doi.org/10.1177/01708406251397720.

Ethayarajh, K. (2019). "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 55–65. https://doi.org/10.18653/v1/D19-1006.

Fama, E.F. and French, K.R. (1993). "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*, 33(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5.

Feng, G., Giglio, S. and Xiu, D. (2020). "Taming the Factor Zoo: A Test of New Factors." *Journal of Finance*, 75(3), 1327–1370. https://doi.org/10.1111/jofi.12883.

Fraiberger, S.P., Sinatra, R., Resch, M., Riedl, C. and Barabási, A.-L. (2018). "Quantifying Reputation and Success in Art." *Science*, 362(6416), 825–829. https://doi.org/10.1126/science.aau7224.

Geva, M., Schuster, R., Berant, J. and Levy, O. (2021). "Transformer Feed-Forward Layers Are Key-Value Memories." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5484–5495. https://doi.org/10.18653/v1/2021.emnlp-main.446.

Higgs, H. and Worthington, A. (2005). "Financial Returns and Price Determinants in the Australian Art Market, 1973–2003." *Economic Record*, 81(253), 113–123. https://doi.org/10.1111/j.1475-4932.2005.00237.x.

Hotelling, H. (1936). "Relations Between Two Sets of Variates." *Biometrika*, 28(3/4), 321–377. https://doi.org/10.2307/2333955.

Kandpal, N., Deng, H., Roberts, A., Wallace, E. and Raffel, C. (2023). "Large Language Models Struggle to Learn Long-Tail Knowledge." *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 15696–15707.

Kazakina, K. (2014). "Art Flippers Pursue Emerging Stars as Doodles Surge 5,600%." *Bloomberg*, 6 February 2014. https://www.bloomberg.com/news/articles/2014-02-06/art-flippers-chase-fresh-stars-as-murillo-s-doodles-soar.

Kraha, A., Turner, H., Nimon, K., Zientek, L.R. and Henson, R.K. (2012). "Tools to Support Interpreting Multiple Regression in the Face of Multicollinearity." *Frontiers in Psychology*, 3, 44. https://doi.org/10.3389/fpsyg.2012.00044.

Lang, G.E. and Lang, K. (1988). "Recognition and Renown: The Survival of Artistic Reputation." *American Journal of Sociology*, 94(1), 79–109. https://doi.org/10.1086/228952.

Lee, K., Park, J., Goree, S., Crandall, D. and Ahn, Y.-Y. (2024). "Social Signals Predict Contemporary Art Prices Better than Visual Features, Particularly in Emerging Markets." *Scientific Reports*, 14, 11615. https://doi.org/10.1038/s41598-024-60957-z.

Li, Y., Ma, M.X. and Renneboog, L. (2022). "Pricing Art and the Art of Pricing: On Returns and Risk in Art Auction Markets." *European Financial Management*, 28(5), 1139–1198. https://doi.org/10.1111/eufm.12348.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D. and Hajishirzi, H. (2023). "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories." *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546.

Mei, J. and Moses, M. (2002). "Art as an Investment and the Underperformance of Masterpieces." *American Economic Review*, 92(5), 1656–1668. https://doi.org/10.1257/000282802762024719.

Meng, K., Bau, D., Andonian, A. and Belinkov, Y. (2022). "Locating and Editing Factual Associations in GPT." *Advances in Neural Information Processing Systems*, 35, 17359–17372.

Mu, J. and Viswanath, P. (2018). "All-but-the-Top: Simple and Effective Postprocessing for Word Representations." *International Conference on Learning Representations (ICLR)*.

Nisbett, R.E. and Wilson, T.D. (1977). "The Halo Effect: Evidence for Unconscious Alteration of Judgments." *Journal of Personality and Social Psychology*, 35(4), 250–256. https://doi.org/10.1037/0022-3514.35.4.250.

Park, J., Lee, Y., Yang, D., Park, J. and Jung, H. (2024). "Artwork Pricing Model Integrating the Popularity and Ability of Artists." *AStA Advances in Statistical Analysis*, 108, 889–913. https://doi.org/10.1007/s10182-024-00504-3.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y. and Miller, A. (2019). "Language Models as Knowledge Bases?" *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. https://doi.org/10.18653/v1/D19-1250.

Renneboog, L. and Spaenjers, C. (2013). "Buying Beauty: On Prices and Returns in the Art Market." *Management Science*, 59(1), 36–53. https://doi.org/10.1287/mnsc.1120.1580.

Sharpe, W.F. (1964). "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance*, 19(3), 425–442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x.

Thorndike, E.L. (1920). "A Constant Error in Psychological Ratings." *Journal of Applied Psychology*, 4(1), 25–29. https://doi.org/10.1037/h0071663.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E. and Stoica, I. (2023). "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." *Advances in Neural Information Processing Systems*, 36.