Pricing Capability: A Framework for Valuing Proprietary Data Assets in the Age of AI

Capability Alpha, Price-Implied Uplift, and the Spirit Aviation Data Auction

Adam Rida

Tracer AI, Inc.
adam@tracerml.ai

August 23, 2026

Abstract

Companies increasingly use proprietary enterprise data to train, adapt, and evaluate AI models, yet there is no established way to price workflow-linked data whose value depends on the capability it adds to a model or system. Google LLC's selected $10 million bid for Spirit Aviation Holdings' deidentified operational data and software provides a rare public case in which this problem can be examined. Existing research already links information and predictive improvement to economic value, while adjacent work studies intangible assets, AI training-data markets, process data, and the contribution of data to model performance. This paper focuses on the transaction-level problem of combining these pieces for proprietary enterprise corpora used in current AI training and evaluation: measuring the marginal performance contribution of a dataset under controlled evaluation, mapping that contribution into buyer-specific economics, and interpreting an observed bid under explicit rights, useful-life, and market-structure assumptions.

We use *Capability Alpha* to denote the causal improvement in economically relevant model or system performance attributable to a proprietary dataset under controlled evaluation. The measured gain is translated into buyer-specific use value through economic exposure, the mapping from performance to commercial outcomes, the expected duration of the gain, usable rights, and acquisition and integration costs. We also define *Price-Implied Uplift* (PIU), a specialization of reverse discounted-cash-flow analysis that recovers the persistent revenue-equivalent uplift required to justify an observed price under a stated revenue-base scenario.

Applied to Spirit, an illustrative scenario in which products or workflows representing 5% of annual Google Cloud revenue form the relevant revenue base implies a three-year PIU of about 0.24% at the selected $10 million bid. Across 1%–10% revenue-scope assumptions, PIU ranges from about 1.18% to 0.12%. A separate forward sensitivity grid shows that assumptions of a 1% improvement in airline economics and 0.5%–2.0% Google capture of the resulting value produce a present-value range of approximately $50 million–$200 million. The $10 million bid covers a combined package of data and software, and the public record does not allocate the price between them. Spirit's Capability Alpha has not been measured, so these are conditional breakeven and scale calculations, not an appraisal. The framework also considers buyer concentration, restrictive rights, and preemption as potential sources of divergence between observed price and buyer-specific value.

Keywords: data valuation; AI training data; model evaluation; intangible assets; reverse DCF; industrial organization.

1 Introduction

Proprietary enterprise records are becoming inputs to model training, post-training, and evaluation. These assets include expert demonstrations, operational workflows, internal software repositories, communications, transaction records, and linked records of decisions and outcomes. Enterprise information systems have produced such records for decades as event logs and audit trails that record how business processes unfold; Aalst [1] surveys this process-mining tradition. What is changing is their role in AI production. Modern agent training can use multi-step trajectories as demonstrations, preference data, evaluation episodes, or environment-grounded supervision, making the sequence and linkage between context, action, tool use, and outcome economically relevant in a way that a static document or feature table may not capture. Chen and Yuille [2] and Yang et al. [3] provide recent examples of agent post-training from expert or environment-grounded trajectories. Oderinwale and Kazlauskas [4] document the broader shift toward proprietary AI training data and identify context-dependent value and contribution to AI production as open measurement problems. Company-reported studies by Mercor [5] and Harvey [6] also show that expert data can materially change performance on professional tasks while illustrating how difficult it is to isolate the contribution of data from the rest of the training stack. Company market activity points in the same direction. Mercor [7] solicits anonymized enterprise workflow data across messaging, documents, CRM, code, finance, and related systems for contribution to frontier AI labs, while Scale AI [8] describes de-identified operational workflows as training data supplied through enterprise data partnerships. Together these offerings provide evidence that the transaction category is emerging. The training value of any particular operational archive remains an empirical question.

The valuation problem draws on several established literatures. Coyle and Manley [9] and OECD [10] review economic approaches to data valuation, while the International Valuation Standards Council [11, 12] provides established machinery for valuing intangible assets. Howard [13] links information to the economic consequences of decisions. Dalessandro, Perlich, and Raeder [14] directly estimate the economic value of incremental predictive data by translating changes in model-performance metrics into expected economic gain. On the machine-learning side, Ghorbani and Zou [15], Jia et al. [16], Xu, Luo, and Li [17], and Tian et al. [18] provide contribution-based or utility-based approaches to data valuation. OpenAI [19, 20] reports evaluations of model performance on tasks with observable economic context.

The contribution of this paper is an AI-specific integration and transaction framework that combines controlled evaluation of proprietary enterprise corpora with buyer-specific intangible-asset valuation, economic life, rights, market structure, and an observed-price inversion. The performance-to-economic translation follows established prior work, including Dalessandro, Perlich, and Raeder [14]. Two quantities organize the analysis. *Capability Alpha* is the task-level performance gain attributable to a proprietary dataset under a specified evaluation design. *Price-Implied Uplift* reverses a revenue-based discounted-cash-flow model to recover the persistent uplift required to justify an observed data price. Empirical estimates of Capability Alpha should be accompanied by confidence intervals or other uncertainty measures appropriate to the evaluation design.

The framework is applied to Spirit Aviation Holdings' proposed transfer of deidentified data and software assets to Google LLC. Spirit Aviation Holdings, Inc. [21] identifies Google as the successful bidder at $10 million, subject to court approval. The public filing describes a large operational corpus spanning communications, IT records, source code, pricing, revenue, airline operations, maintenance, finance, legal, and employee records. The selected bid covers both data and software, so it is an observable package price. It does not identify a standalone price for data.

Two quantitative exercises are used. The first is a forward sensitivity analysis: given assumptions

about economic impact and value capture, what present value follows for a particular buyer? The second is PIU: given the selected bid and a stated revenue base, how much persistent revenue-equivalent uplift is required to break even? Under an illustrative 5% Google Cloud revenue-scope assumption, the $10 million bid implies a PIU of about 0.24% over three years. A separate forward sensitivity calculation produces a $50 million–$200 million range under the assumptions stated in Section 6.2.

Market structure is treated separately from direct use value. Gans [22] and Korinek and Vipra [23] describe concentration and complementary inputs in AI markets, including compute, model development, and distribution. Restrictive rights can also create strategic value when rival access has a measurable competitive effect. These mechanisms can cause observed price to diverge from buyer-specific economic value.

2 Related literature and positioning

2.1 Intangible assets and data valuation

The valuation literature already provides the financial machinery required to value intangible assets. Cost approaches estimate the resources required to create or replace an asset. Market approaches use comparable transactions. Income approaches discount incremental cash flows attributable to an asset. The International Valuation Standards Council [11, 12] distinguishes Market Value from Investment Value or Worth, which is specific to an owner or prospective owner, and Synergistic Value, which can arise from combinations of assets or interests.

Coyle and Manley [9] find no single empirical valuation method that generalizes across data types and uses. OECD [10] likewise emphasizes that value depends on access, governance, rights, and context of use. Jones and Tonetti [24] formalize the nonrival nature of data and show why ownership, reuse, and market structure can affect whether data is shared or retained. Mohan, Bharathy, and Jalan [25] review enterprise-data valuation methods and document continuing implementation and reporting challenges. Bhutani, Ordoñez, and Veldkamp [26] model data as a productive asset whose value can depreciate, reinforcing the need to represent economic life explicitly.

The AI-training-data literature makes the production problem explicit. Oderinwale and Kazlauskas [4] characterize data as a distinct input to AI production and identify context-dependent value and estimation of data's contribution to production as core open questions. Income-based valuation can translate an attributable economic benefit into value, but it requires a defensible estimate of the benefit first. Rappaport and Mauboussin [27] provide the reverse discounted-cash-flow logic used here to start from price and infer the operating performance required to justify it.

2.2 Economic value of information and incremental predictive data

The closest antecedent to the technical-to-economic bridge in this paper predates modern foundation models. Howard [13] assigns economic value to information through the decisions and consequences it changes. Dalessandro, Perlich, and Raeder [14] estimate the economic value of incremental commercial data by first measuring the change in predictive-model performance and then translating that change into expected economic gain. This is direct prior art for the performance-to-value step and establishes that data value is use-dependent.

The present framework applies that logic to proprietary enterprise corpora used in model adaptation, post-training, retrieval, evaluation, and AI systems; heterogeneous task families and model

stacks; legal and contractual usability; technological obsolescence; buyer-specific complements; and thin markets in which restrictive rights may create strategic effects. Bostoen and Monti [28] similarly emphasize transaction design and bargaining institutions when direct comparables for AI training data are scarce. Capability Alpha is therefore a named empirical input to the framework. Causal inference and the general performance-to-economic translation step are established prior art; the contribution here is the transaction-level integration described above.

2.3 From predictive features to workflow-linked trajectories

Enterprise workflow data has a long history as an internal analytical asset. Aalst [1] describes event logs as records of business-process execution used for process discovery, conformance checking, and operational analysis. The recent change is in the production role assigned to those records. In agentic systems, a linked workflow can become training supervision or an evaluation environment when it preserves enough of the sequence from state and context through action, tool use, intermediate results, and outcome.

Recent technical work makes this distinction concrete. Chen and Yuille [2] treat expert trajectories as supervision for agent policy optimization and convert multi-turn demonstrations into state-conditioned preferences. Yang et al. [3] show that interaction trajectories exposing inspect-act-verify structure can produce strong generalization and data efficiency in terminal-agent post-training. These results do not establish the monetary value of enterprise records. They show why linkage and trajectory structure can matter technically for modern agent training.

Operational records themselves are longstanding. The shift lies in the data object that is being priced and in its role inside the AI production process. Dalessandro, Perlich, and Raeder [14] study incremental third-party feature data for defined predictive applications in advertising and direct mail. Modern enterprise AI can assign value to workflow-linked records that were originally generated as operational exhaust, process evidence, communications, or audit trails, because the same corpus may support demonstrations, evaluation tasks, verifiers, retrieval, and post-training across multiple task families. Current market offerings from Mercor [7] and Scale AI [8] provide direct evidence that such workflow data is now being packaged for external AI buyers. The valuation problem addressed here is the external pricing of that emerging AI use while preserving the established insight that data value is conditional on the buyer, task, and decision environment.

2.4 Machine-learning data valuation

Machine-learning research approaches data value from the technical side. Ghorbani and Zou [15] assign value to training observations according to their marginal contribution to a model-performance payoff, and Jia et al. [16] develop more efficient Shapley-based estimators. Related work studies the effect of changing the training set itself. Ilyas et al. [29] fit *datamodels* that predict how model outputs change across training-set subsets, while Park et al. [30] develop TRAK for scalable attribution of model behavior to training examples. In the LLM setting, Xia et al. [31] use influence estimates to select instruction-tuning data targeted at a desired capability. These methods are useful for locating influential subsets and approximating data counterfactuals.

They do not remove the need for a direct asset-level experiment. Capability Alpha concerns the effect of giving a specified development or deployment procedure access to dataset D as a treatment. Training-data attribution can help decide which parts of D to test, diagnose where gains come from, or reduce the cost of subset search; the final estimate used for valuation should come from a matched treatment-control evaluation whenever that experiment is feasible.

Recent work also addresses buyer-specific utility and information asymmetry in AI data transactions. Yang, Gao, and Zheng [32] propose a privacy-preserving protocol that scores external data against a buyer's model without revealing the raw data, directly addressing the inspection problem that arises when utility must be assessed before exchange. Xu, Luo, and Li [17] combine information-density measures, proxy-model training gain, influence methods, and Data Shapley for utility-aware pricing of LLM training data. These approaches strengthen the technical measurement layer. A contribution measured in benchmark or influence units still requires a separate mapping into buyer-specific economics.

Capability Alpha is used here as that intermediate empirical quantity. It is task- and evaluation-specific, and the valuation layer translates it into buyer-specific economic outcomes; benchmark points have no universal dollar price.

2.5 Evaluation on economically relevant tasks and expert data

Economically grounded evaluations make the measurement step more plausible. OpenAI [19] reports more than 1,400 SWE-Lancer freelance software-engineering tasks associated with roughly $1 million in actual payouts. OpenAI [20] evaluates 1,320 GDPval tasks across 44 occupations selected for economic significance using work products created by experienced professionals. These benchmarks move evaluation closer to tasks whose economic stakes can be observed or approximated.

Company-reported post-training experiments provide examples of the causal-attribution problem. Mercor [5] reports large gains after training GLM 4.6 on 874 expert-created tasks in professional environments. Harvey [6] reports gains from a stack combining public, synthetic, and human-expert data with post-training and harness changes, and separately shows that harness changes alone can move benchmark performance. These reports are useful as examples of measurement design; they are not independent estimates of the value of enterprise data.

A before-and-after comparison is insufficient when data, model checkpoint, tools, compute, rewards, and harness design move together. The empirical object needed for valuation is the marginal contribution of the proprietary corpus under a controlled stack.

2.6 Economics of AI and buyer concentration

The same measured capability gain can have different economic value across buyers. Demirer et al. [33] document substantial variation in demand for model capability across applications using OpenRouter and Microsoft Azure usage data. Gans [22] emphasizes that data markets interact with bottlenecks such as compute, model openness, and platform design in determining AI market power. Korinek and Vipra [23] analyzes scale economies and concentration in foundation-model development, including the role of compute, data, and talent.

For data transactions, this means the set of firms able to receive a dataset can be much larger than the set able to monetize it at scale. Buyer-specific values can differ substantially, and thin buyer markets can make observed prices sensitive to bargaining power, process design, and the value of the next credible bidder.

2.7 Exclusivity, foreclosure, and preemption

Exclusive or restrictive rights add a strategic channel. OECD [34] treats access to data and input foreclosure as potentially relevant competitive mechanisms in digital mergers. OECD [35] similarly

discusses data access, vertical integration, model restrictiveness, and control of enabling inputs in AI downstream competition.

Preemption also appears in models of strategic investment. Weeds [36] analyzes R&D investment under uncertainty when competition for a winner-takes-all asset changes investment timing through preemption incentives. In a data transaction, the narrower analogue is that restrictive rights can create value for a buyer if denying a close rival access to the same capability-relevant input changes expected profits.

The empirical framework separates direct use value from any strategic component. The Spirit calculations below include direct use value only because the public record does not provide enough evidence to estimate a competitive-denial effect.

2.8 Research gap

The component literatures already cover much of the underlying machinery. The gap addressed here is narrower. Earlier work establishes the economic value of information and the translation of incremental predictive performance into money; process mining establishes the analytical value of enterprise event logs; recent AI-data work develops contribution scoring and privacy-preserving buyer-side utility measurement; and agent-training research establishes the technical value of interaction trajectories. What remains underdeveloped is a transaction-level framework for workflow-linked enterprise corpora used in modern AI production: a controlled estimate of incremental capability across economically relevant tasks, an explicit mapping into buyer-specific economics, treatment of useful life and rights, and an inversion of an observed price into a testable break-even threshold. The Spirit case is used to show how these pieces can be stated together without treating any sensitivity scenario as an appraisal.

3 Framework

3.1 Measurement concepts

Three concepts organize the measurement layer.

Capability gap. For task family t, let H_t denote a reference level of performance and F_t the performance of a defined practical baseline under a standardized harness. The gap $G_t = H_t - F_t$ is a diagnostic of how much task-level performance remains available to improve. A large gap creates room for improvement; it says nothing by itself about whether dataset D contains useful signal.

Capability Alpha. Capability Alpha is a task-level treatment effect measured under a declared reference adaptation stack. For dataset D, task t, and a pre-specified data dose k, let M_0 and M_1 denote the matched control and treatment models defined in Section 4. We define

$$CA_{D,t}(k) = \mathbb{E}[Y_t(M_1)] - \mathbb{E}[Y_t(M_0)], \tag{1}$$

where Y_t is a pre-specified held-out task metric and the expectation is over training and evaluation randomness. The reference model, adaptation method, token and compute budget, and evaluation harness are fixed before the comparison. Capability Alpha is therefore conditional on a disclosed experimental stack, in the same sense that contribution-based data valuation depends on a learning algorithm and utility metric [15, 16]. It can be positive or negative. Different task metrics are kept separate and translated into economic units task by task through $\lambda_{b,t}$; no cross-task technical average is used.

Verifiability and outcome linkage. These are descriptive corpus properties and are not separate valuation variables. Outcome linkage is related to the event-log and case-linkage concepts described by Aalst [1], where enterprise records are connected into executions of a business process. Here the emphasis is on whether the recovered sequence is usable for AI evaluation or post-training. Verifiability concerns the share and quality of cases for which success can be judged by an objective, expert, or reproducible criterion. Outcome linkage concerns the ability to connect context, action, intermediate state, and eventual outcome across records. These properties can matter because linked trajectories can supply demonstrations, environments, preference signals, and reward or verifier targets, as illustrated by Chen and Yuille [2] and Yang et al. [3].

Figure 1 summarizes the measurement-to-valuation pipeline.

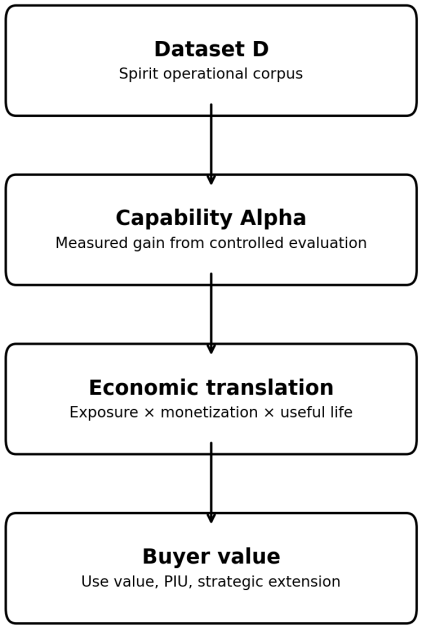

Figure 1: From proprietary data to buyer-specific value.

3.2 Direct use value

For buyer b, task family t, and future year h, let $E_{b,t,h}$ denote annual economic exposure measured in dollars of revenue, cost, or another defined economic base. Let $\lambda_{b,t}$ map one unit of the task-performance metric into a fractional change in that economic base. Let $d_{D,t,h}$ denote the fraction of the initial data advantage expected to persist at horizon h. A reduced-form direct use value is

$$V_b^{use}(D) = \rho_D \sum_{h=1}^{T} \frac{1}{(1+r)^h} \sum_t \left(E_{b,t,h}\lambda_{b,t}CA_{D,t}d_{D,t,h}\right) - C_b(D), \tag{2}$$

where $\rho_D \in [0,1]$ is a reduced-form rights-and-usability adjustment, r is the annual discount rate, and $C_b(D)$ is the buyer-specific cost of acquisition, legal work, deidentification, cleaning, training, integration, and deployment. The factor ρ_D should represent restrictions on usable rights and

access; technical uncertainty belongs in the Capability Alpha estimate, and strategic exclusivity is handled separately below. When a legal restriction removes an entire use case, the corresponding task exposure should be set to zero instead of relying on a single global haircut. Under these definitions, the inner product has units of dollars per year. The $\lambda_{b,t} CA_{D,t}$ term is a local linear approximation. If commercial response is thresholded or strongly nonlinear, it should be replaced by an estimated response function. Task-level economic bases must also be defined to avoid double counting overlapping revenue or cost exposure.

The distinction between r and $d_{D,t,h}$ is important. The discount rate converts future cash flows into present value and reflects time value, risk, and opportunity cost. The persistence term describes whether the data advantage itself erodes as models improve, substitute data appears, or business processes change, consistent with economic-life treatment in the International Valuation Standards Council [12] and data-depreciation analysis by Bhutani, Ordoñez, and Veldkamp [26].

The same equation can be applied to the current owner by setting b equal to the owner. Capability Alpha remains the measured technical contribution under the reference evaluation, while owner or buyer heterogeneity enters through economic exposure, monetization, costs, and the rights package. This allows internal use and transfer to be compared within the same valuation structure.

3.3 Price-Implied Uplift

For the Spirit case, we also use a simplified revenue-base specialization. Let R_{scope} be a defined annual revenue base, m the operating margin applied to incremental revenue, T the useful life, r the discount rate, ρ the rights adjustment, and C_b buyer-specific integration and transaction costs. Define the annuity factor

$$A(T, r) = \sum_{h=1}^{T} \frac{1}{(1+r)^h}.$$

The revenue-equivalent uplift u required to support transaction price P is

$$\text{PIU} = u^* = \frac{P + C_b}{R_{\text{scope}} m A(T, r) \rho}. \tag{3}$$

This is a reverse discounted-cash-flow calculation applied to a data transaction, following the price-implied expectations logic of Rappaport and Mauboussin [27]. In this specialization, PIU is a percentage uplift to the scoped revenue base, with operating margin converting incremental revenue into operating profit. A cost-savings application would use a corresponding cost base and conversion rule instead of mechanically applying an operating margin.

PIU reports the break-even threshold implied by the stated revenue-base scenario. If PIU equals 0.236%, the purchase requires a persistent 0.236% revenue-equivalent uplift over the modeled horizon. Realized performance remains a separate empirical quantity.

3.4 Strategic extension

Direct use value can be extended when a separately identifiable strategic effect is supported by evidence:

$$V_b^{total}(D) = V_b^{use}(D) + S_b(D). \tag{4}$$

Here $S_b(D)$ is the present value, to buyer b, of named strategic effects that are outside the buyer's own task-level cash flows in Equation 2. Each component included in $S_b(D)$ must correspond to a stated strategic mechanism and be estimated separately.

One possible component is preemption value. For buyer b,

$$V_b^{preemption}(D) = \Pi_b(\text{rivals without access to } D) - \Pi_b(\text{rivals with access to } D). \tag{5}$$

A separately estimated option value can also enter $S_b(D)$ when ownership creates a credible future use that is not represented in the enumerated task cash flows. The Spirit sensitivity range sets $S_b(D) = 0$ because the public record does not support an estimate of either component.

4 Estimating Capability Alpha

The empirical object is a dataset-level counterfactual: under a fixed model-adaptation procedure and fixed resource budget, how much does replacing available reference data with proprietary data D change out-of-sample performance on task t? This follows the use-dependent utility logic of Data Shapley [15, 16], the training-set counterfactual view of Datamodels [29], and recent LLM work that prices data from empirical training gain [17]. Capability Alpha is estimated by direct matched retraining; influence and attribution methods are used only as secondary diagnostics.

4.1 Estimand and reference design

The measurement fixes a base model M, an adaptation operator $\mathcal{A}$, a total adaptation budget B, a reference corpus R, a proprietary-data dose $D_k \subseteq D$ of size k, and a held-out task evaluation $\mathcal{E}_t$ with metric Y_t. The reference corpus represents the data that would occupy the same adaptation budget if the proprietary asset were unavailable.

The adaptation operator is the mechanism through which training data can change model capability. It may be full fine-tuning or a parameter-efficient method such as LoRA [37]; whichever procedure is chosen is held fixed across treatment and control. The resulting Capability Alpha is therefore conditional on a disclosed reference stack; an unspecified notion of "data quality" is insufficient.

For training seed s, define

$$M_0^{(s)} = \mathcal{A}(M, R_B; s), \tag{6}$$

$$M_1^{(s)} = \mathcal{A}(M, R_{B-k} \cup D_k; s), \tag{7}$$

where R_B contains B units of reference data and R_{B-k} contains $B - k$ units drawn under the same sampling rule. Both arms use the same checkpoint, optimizer, adaptation method, training schedule, token or example budget, compute budget, and evaluation harness. The intervention changes only the provenance of k units of training data. Comparing an unadapted model with a model trained on D would mix the effect of the data with the effect of additional optimization.

4.2 Held-out task construction

The task family t, target distribution, metric Y_t, and verifier are fixed before treatment results are observed. Economically relevant tasks are preferred because the subsequent valuation requires a mapping from technical performance into business outcomes. SWE-Lancer and GDPval provide examples of evaluations built from realistic work with explicit success criteria [19, 20].

The evaluation set $\mathcal{E}_t$ is disjoint from the adaptation data. For workflow-linked enterprise corpora, splitting occurs at the case or entity level needed to prevent leakage: an incident, customer,

project, route, contract, or later time period is held out as a unit. Objective outcomes are used as verifiers when available; otherwise a calibrated expert rubric is fixed in advance. Near duplicates, copied resolutions, shared identifiers, and temporal overlap are screened because they can turn memorization into apparent capability. Datamodels shows how training-set counterfactuals can expose this form of leakage [29].

4.3 Treatment-effect estimator

Treatment and control are evaluated on the same held-out cases. With S matched training seeds,

$$\widehat{CA}_{D,t}(k) = \frac{1}{S}\sum_{s=1}^{S}\left[Y_t(M_1^{(s)}, \mathcal{E}_t) - Y_t(M_0^{(s)}, \mathcal{E}_t)\right]. \tag{8}$$

A confidence interval is estimated from the paired evaluation outcomes and run-to-run variation across seeds. Negative estimates are retained. A placebo arm using an equally sized irrelevant or shuffled corpus can be added when feasible to detect procedural gains unrelated to information in D.

As a stylized example, let D contain IT-support trajectories and let Y_t be verified resolution rate on 1,000 held-out incidents. Suppose both arms use the same LoRA configuration and a 20-million-token adaptation budget. The control uses 20 million reference tokens; the treatment replaces 5 million of those tokens with D. If mean resolution is 55.0% for control and 61.2% for treatment across matched runs, then $\widehat{CA}_{D,t}(5\text{M}) = +6.2$ percentage points. That number is the measured technical contribution of the proprietary data under the declared stack. It contains no monetary assumption.

4.4 Scaling, transfer, and valuation handoff

Repeating the experiment at pre-specified values of k produces a capability curve $CA_{D,t}(k)$ and reveals saturation. Repeating the matched intervention on a second model family or checkpoint tests whether the sign and approximate magnitude transfer beyond one reference model. Regressions on adjacent economically relevant tasks are measured alongside the target-task gain.

Data Shapley [15], Datamodels [29], TRAK [30], and LESS [31] can then rank records or trajectories for follow-up ablations and data selection. They do not replace the direct treatment-control estimate in Equation 8. For a corpus spanning several task families, the measurement output is the set of task-specific estimates and uncertainty intervals $\{\widehat{CA}_{D,t}(k), CI_{D,t}(k)\}_{t\in\mathcal{T}}$. Each task is carried separately into Equation 2, where $E_{b,t,h}$ and $\lambda_{b,t}$ translate the measured capability contribution into buyer-specific economic value.

5 Spirit Aviation case study

5.1 Transaction and asset

Spirit Aviation Holdings, Inc. [21] designated Google LLC as the successful bidder for a set of data and software assets at $10 million, subject to court approval and the required deidentification process. The same filing identifies Mercor as the $7.5 million alternate bidder. Business Insider [38] later reported micro1's $12.5 million offer outside the initial auction process. As of August 23, 2026, approval remained pending after employee-privacy objections delayed the sale hearing,

as reported by Bloomberg Law [39]. Bloomberg Law [40] reported Google's statement that the acquired portion of Spirit's enterprise dataset could help improve its products and AI models.

Spirit Aviation Holdings, Inc. [21] describes the selected $10 million bid as covering a combined package of data and internally developed software. The public record does not allocate the package price between those components. The PIU exercise below uses the full $10 million as P. If the software has positive standalone value, that convention places an upper bound on the price attributed to the data component; if the components are complementary, a clean allocation may not be economically meaningful. Spirit Aviation Holdings, Inc. [21] also requires Google to bear third-party deidentification costs in addition to the headline bid, so the zero-cost PIU case is explicitly a base sensitivity and does not represent total acquisition cost.

The asset schedule in Spirit Aviation Holdings, Inc. [21] describes a broad operating corpus. It includes approximately 100 million emails, 500 million Microsoft Teams items, 17.1 million OneDrive items, 20.6 million SharePoint items, 667,563 IT tickets, 516 code repositories containing roughly 30 million lines of code, 372,585 commits, 3.53 billion booking-curve observations, 7.25 billion competitor-flight observations, 7.51 billion revenue transactions, and extensive airline operations, maintenance, finance, legal, and employee records.

The agreement in Spirit Aviation Holdings, Inc. [21] requires deidentification while preserving referential integrity across the data structure. Stable identifiers and timestamps could preserve useful workflow trajectories, for example linking a pricing state to a booking outcome or an IT ticket to a code change and deployment result. Whether those linkages are complete enough to support training or evaluation is an empirical question that cannot be answered from the asset schedule alone.

The transferred assets are a bounded subset of Spirit's data estate. Spirit Aviation Holdings, Inc. [21] excludes major consumer datasets, including customer profiles and Free Spirit member records.

5.2 Why the asset could matter to Google

Google LLC is the buyer. Google Cloud is used below as one observable monetization base because Alphabet Inc. [41] publicly reports its revenue and operating margin. Cloud serves solely as a transparent economic base for the sensitivity analysis.

Google Cloud already supports airline optimization and decision-support use cases. Google Cloud [42] describes operational decision support using crew, passenger, rotation, and technical data and reports savings from rotation optimization. Google Cloud [43] describes machine-learning revenue management and reports material revenue effects in customer deployments. These examples establish that airline operations and revenue-management tasks can have measurable commercial consequences. They provide no estimate of Spirit's contribution.

6 Spirit valuation scenarios

6.1 Observed bids

The observed process gives three price signals. Spirit Aviation Holdings, Inc. [21] records Mercor at $7.5 million and Google selected at $10 million; Business Insider [38] reports micro1's later $12.5 million indication. The micro1 figure is stated willingness to pay because it arrived outside the initial auction process and had not become a completed transaction in the record used here.

Figure 2 places these market observations next to the forward sensitivity range developed below. The observed bids and the modeled interval answer different questions.

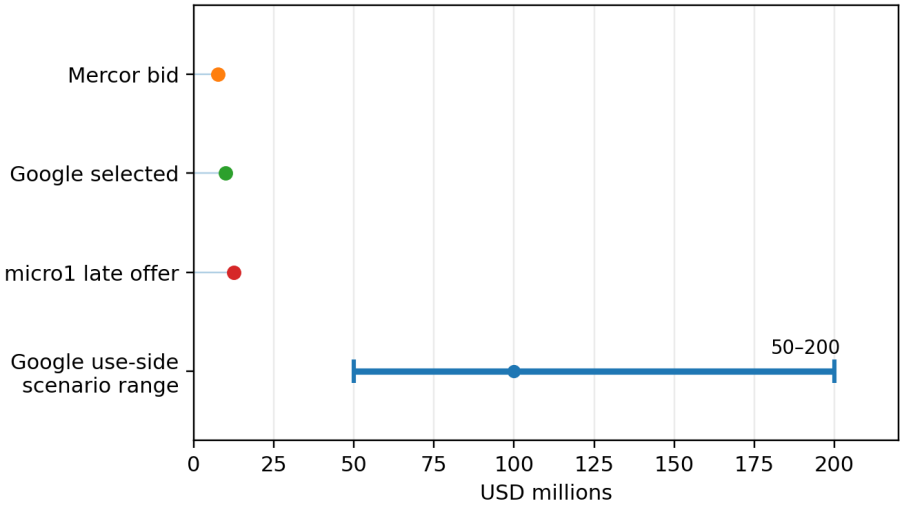

Figure 2: Observed bids and one conditional buyer-specific use-value range.

6.2 Illustrative forward sensitivity, not an appraisal

The $50 million–$200 million range is generated by the assumptions in Table 1. Observed inputs and scenario inputs are shown separately.

Input	Value	Status
2026 global airline revenue	$1.165T	IATA [44]
Modeled improvement in airline economics	1.0%	Scenario assumption
Google capture of created value as incremental revenue	0.5%–2.0%	Scenario assumption
Google Cloud operating margin	35.6%	Alphabet [41]
Economic life	3 years	Scenario assumption
Discount rate	12%	Scenario assumption; not a Google WACC estimate

Table 1: Inputs to the forward Spirit use-value sensitivity calculation.

A 1% improvement in airline economics corresponds to $11.65 billion of annual value across the industry. If Google captures 0.5%–2.0% of that value as incremental revenue, the implied annual Google revenue is approximately $58.25 million–$233 million. Applying the 35.6% Google Cloud operating margin gives approximately $20.7 million–$82.9 million of annual operating profit. A three-year life discounted at 12% has an annuity factor of approximately 2.402, producing a present value of approximately $49.8 million–$199.2 million.

The calculation is a scale sensitivity. The 1% industry effect and 0.5%–2.0% capture rates are assumptions. They are not estimated from Spirit data and should not be interpreted as a central case. Table 2 shows how strongly the result moves when the airline-impact assumption changes.

Modeled airline economic improvement	0.5% capture	1.0% capture	2.0% capture
0.25%	\$12.5M	\$24.9M	\$49.8M
0.50%	\$24.9M	\$49.8M	\$99.6M
1.00%	\$49.8M	\$99.6M	\$199.2M
2.00%	\$99.6M	\$199.2M	\$398.4M

Table 2: Present-value sensitivity to the two unobserved forward assumptions. Values use the observed Google Cloud margin, a three-year life, and a 12% discount rate.

The \$50 million–\$200 million interval is therefore one row of a wider sensitivity surface, not a valuation estimate. A measured Capability Alpha and a production estimate of λ would be required to replace these assumptions with evidence.

6.3 Price-Implied Uplift for Spirit

PIU starts from the selected \$10 million package bid. Alphabet Inc. [41] reports Q2 2026 Google Cloud revenue that annualizes to approximately \$99.072 billion. Define a *relevant revenue scope* as the share of annual Google Cloud revenue associated with products or workflows that could plausibly experience an economic effect from capabilities affected by the dataset.

For the illustrative 5% scenario,

$$R_{\text{scope}} = 0.05 \times \$99.072\text{B} = \$4.954\text{B}.$$

The 5% figure refers only to modeled scope; it carries no claim about aviation's share of Cloud revenue or a 5% Spirit-driven revenue increase.

Applying the reported 35.6% Google Cloud segment operating margin as a proxy gives approximately \$1.763 billion of annual operating profit associated with the scoped revenue base. The segment margin is not an observed incremental margin for a Spirit-derived product, so it is another sensitivity input. Discounting three years at 12% gives a present-value profit base of about \$4.23 billion. With zero integration cost and $\rho = 1$ for this base sensitivity,

$$\text{PIU} = \frac{\$10\text{M}}{\$4.23\text{B}} \approx 0.236\%.$$

Table 3 and Figure 3 show the sensitivity to the scope assumption.

Relevant Cloud revenue scope	Annual revenue base	PIU for \$10M price
1%	\$0.991B	1.181%
2.5%	\$2.477B	0.472%
5%	\$4.954B	0.236%
10%	\$9.907B	0.118%

Table 3: Price-Implied Uplift under alternative revenue-scope assumptions.

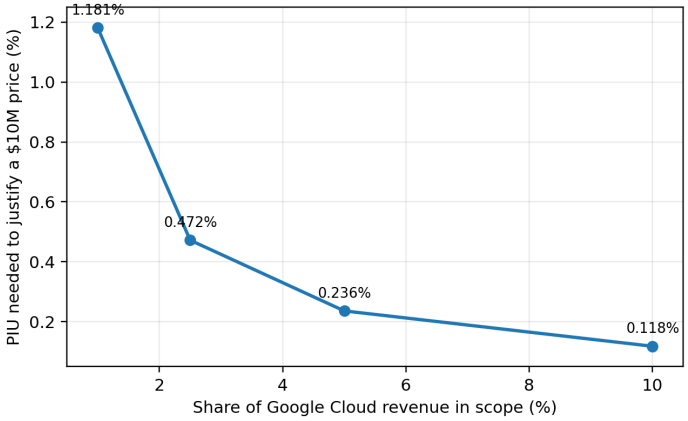

Figure 3: PIU declines as the modeled revenue scope grows.

Under the 5% scope assumption, a persistent 0.236% revenue-equivalent uplift in the scoped Google Cloud base is sufficient for a $10 million purchase to break even over three years under the stated margin, life, rights, and cost assumptions. Holding the 5% scope and 35.6% margin fixed, changing the discount rate from 8% to 20% moves PIU only from about 0.220% to 0.269%; the revenue-scope and economic-translation assumptions are materially more consequential. Whether the dataset can generate the required effect is the empirical question.

7 Market structure, retention, and strategic value

The Spirit process illustrates why transaction price and buyer-specific value can diverge. A proprietary corpus can be technically transferable to many firms while only a small number possess the models, compute, post-training infrastructure, distribution, and commercial exposure needed to monetize it at scale, consistent with the market-structure analyses of Gans [22] and Korinek and Vipra [23]. In a thin buyer market, a high-value buyer can obtain substantial surplus even when the observed price is much lower than its own Investment Value or Worth.

The current owner can be evaluated with the same framework. Setting b equal to the owner produces an internal retention value. A firm with limited economic exposure, weak commercialization channels, or high integration costs may rationally assign low internal value to a dataset even when the dataset has measurable Capability Alpha. A buyer with larger complementary assets can value the same measured gain much more highly. This difference creates a potential gain from trade without requiring a buyer-specific definition of Capability Alpha.

Transfer can also impose a competitive cost on the seller if the buyer, or customers of the buyer, use the resulting capability against the seller. That cost affects the seller's reservation value and the contract terms it should accept. Field-of-use restrictions, nonexclusive licensing, time limits, or higher consideration can change that trade-off. The relevant strategic effect belongs in the rights package or in a separately estimated strategic component, not inside Capability Alpha.

Exclusivity can create preemption value when a close rival would gain meaningful capability from the same data and cannot obtain an equivalent substitute. Related mechanisms appear in OECD [34], OECD [35], and Weeds [36]. Equation 5 provides a way to state that effect. The Spirit sensitivity range excludes preemption because the public record does not provide enough evidence to estimate it.

Technology can change both sides of the market over time. Improvements in models, hardware,

and post-training tooling can lower the complementary investment required to exploit a corpus and expand the set of economically capable buyers. The same progress can reduce the use value of a fixed corpus if baseline performance closes the relevant task gap or substitute data and synthetic environments improve. Asset value and seller bargaining power can therefore move in different directions.

8 Limitations and falsification

Spirit's Capability Alpha has not been measured. The available evidence therefore cannot establish whether the assets were underpriced. The framework identifies the technical and economic quantities required to evaluate that claim.

Capability Alpha is conditional on the evaluation stack. Model family, compute, prompting, tools, retrieval, reward design, and harness engineering can all affect the observed treatment effect. Robust estimation requires controlled ablations, uncertainty intervals, and transfer tests across model families. Company-reported post-training results cited in Section 2 illustrate these design issues but are not substitutes for an independent Spirit evaluation.

Economic translation is a second major uncertainty. The coefficients $E_{b,t,h}$ and $\lambda_{b,t}$ determine how a technical gain becomes money, and neither is observed for Spirit. Equation 2 also uses a local linear and additive approximation; nonlinear product response, interactions across tasks, and overlapping economic bases require a richer production model. The forward \$50 million–\$200 million range is especially sensitive to the assumed 1% airline-industry effect and 0.5%–2.0% Google capture rate. The PIU calculation is algebraically tighter because it starts from an observed price, but its denominator still depends on the chosen revenue scope, margin, life, rights, and costs.

Rights and usability add further uncertainty. A technically useful corpus can lose economic value through privacy restrictions, deidentification burdens, broken linkage, contractual limits, or high integration costs. Economic life is uncertain because frontier progress and substitute data can erode scarcity, consistent with the International Valuation Standards Council [12] and Bhutani, Ordoñez, and Veldkamp [26].

The Spirit buyer-specific value thesis would weaken if representative samples show low practical usability, weak Outcome Linkage, low Verifier Density, little transfer to economically relevant tasks, or a measured Capability Alpha that translates into commercial uplift below the PIU threshold under defensible scope assumptions. The Spirit package also combines data and software, and Google bears deidentification costs beyond the headline bid. A standalone data valuation would therefore require an allocation of package value and a measured cost term. PIU remains usable as a package-level sensitivity. Those issues limit any claim about a standalone Spirit data price.

9 Conclusion

This paper adapts and combines established ideas from value-of-information analysis, predictive-data valuation, intangible-asset finance, and machine-learning data valuation for proprietary enterprise corpora used in contemporary AI systems. Capability Alpha is the paper's shorthand for the controlled task-level performance contribution used as the technical input. Direct use value then depends on economic exposure, the mapping from performance to commercial outcomes, persistence, usable rights, and buyer-specific costs. PIU provides the inverse calculation for a revenue-base scenario by recovering the persistent revenue-equivalent uplift required to justify an observed trans-

action price.

The Spirit Aviation transaction provides a public case with observable bids and a disclosed asset schedule. Under an illustrative 5% Google Cloud revenue-scope assumption, the selected $10 million package bid implies a three-year PIU of about 0.24%. A separate forward sensitivity calculation produces a present-value range of approximately $50 million–$200 million under explicit 1% airline-impact and 0.5%–2.0% Google-capture assumptions. These calculations are hypotheses about the scale of the economics, not measurements of Spirit's realized contribution or a standalone price for the data component.

The same framework also clarifies the owner's decision to retain, use, license, or sell data. The current owner and an external buyer can assign different values to the same measured Capability Alpha because their economic exposure, commercialization channels, costs, and strategic incentives differ. An important emerging category is workflow-linked operational data: records that predate modern AI as business-process exhaust, communications, audit trails, or process evidence, but can now be reused as demonstrations, evaluation episodes, verifiers, or post-training trajectories for agentic systems. Empirical progress depends on controlled data evaluations, credible production mappings, transparent rights packages, and repeated observations of actual transactions and realized outcomes.

References

[1] W. M. P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011. DOI: 10.1007/978-3-642-19345-3. URL: https://doi.org/10.1007/978-3-642-19345-3.

[2] Y. Chen and A. L. Yuille. "Agentic-DPO: From Imitation to Agentic Policy Optimization on Expert Trajectories". In: *arXiv preprint arXiv:2607.10601* (2026). DOI: 10.48550/arXiv.2607.10601. URL: https://arxiv.org/abs/2607.10601.

[3] S. Yang et al. "What Makes Interaction Trajectories Effective for Training Terminal Agents?" In: *arXiv preprint arXiv:2606.03461* (2026). DOI: 10.48550/arXiv.2606.03461. URL: https://arxiv.org/abs/2606.03461.

[4] H. Oderinwale and A. Kazlauskas. "The Economics of AI Training Data: A Research Agenda". In: *arXiv preprint arXiv:2510.24990* (2025). URL: https://arxiv.org/abs/2510.24990.

[5] Mercor. *Expert data drives model performance*. Jan. 2026. URL: https://www.mercor.com/blog/expert-data-drives-model-performance/.

[6] Harvey. *Update on Our Post-Training Effort*. Aug. 2026. URL: https://www.harvey.ai/blog/post-training-update-harvey-tenet.

[7] Mercor. *Enterprise Workflow Data for Frontier AI*. 2026. URL: https://www.mercor.com/data/ (visited on 08/23/2026).

[8] Scale AI. *Data Partnerships*. 2026. URL: https://scale.com/data-partnership (visited on 08/23/2026).

[9] D. Coyle and A. Manley. "What is the value of data? A review of empirical methods". In: *Journal of Economic Surveys* 38.4 (2024), pp. 1317–1337. DOI: 10.1111/joes.12585.

[10] OECD. *Measuring the value of data and data flows*. Tech. rep. Digital Economy Papers No. 345. OECD, 2022. URL: https://www.oecd.org/en/publications/measuring-the-value-of-data-and-data-flows_923230a6-en.html.

[11] International Valuation Standards Council. *International Valuation Standards: Bases of Value*. 2025. URL: https://ivsc.org/standards-glossary/.

[12] International Valuation Standards Council. *IVS 210: Intangible Assets*. 2025. URL: https://ivsc.org/wp-content/uploads/2021/10/IVS210IntangibleAssets.pdf.

[13] R. A. Howard. "Information Value Theory". In: *IEEE Transactions on Systems Science and Cybernetics* 2.1 (1966), pp. 22–26. DOI: 10.1109/TSSC.1966.300074. URL: https://doi.org/10.1109/TSSC.1966.300074.

[14] B. Dalessandro, C. Perlich, and T. Raeder. "Bigger is Better, but at What Cost? Estimating the Economic Value of Incremental Data Assets". In: *Big Data* 2.2 (2014), pp. 87–96. DOI: 10.1089/big.2014.0010. URL: https://doi.org/10.1089/big.2014.0010.

[15] A. Ghorbani and J. Zou. "Data Shapley: Equitable Valuation of Data for Machine Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019. URL: https://proceedings.mlr.press/v97/ghorbani19c.html.

[16] R. Jia et al. "Towards Efficient Data Valuation Based on the Shapley Value". In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. 2019. URL: https://proceedings.mlr.press/v89/jia19a.html.

[17] M. Xu, Q. Luo, and K. Li. "Utility-Aware Data Pricing: Token-Level Quality and Empirical Training Gain for LLMs". In: *arXiv preprint arXiv:2604.22893* (2026). URL: https://arxiv.org/abs/2604.22893.

[18] Z. Tian et al. "Private Data Valuation and Fair Payment in Data Marketplaces". In: *arXiv preprint arXiv:2210.08723* (2022). URL: https://arxiv.org/abs/2210.08723.

[19] OpenAI. *Introducing the SWE-Lancer benchmark*. Feb. 2025. URL: https://openai.com/index/swe-lancer/.

[20] OpenAI. *Measuring the performance of our models on real-world tasks: GDPval*. Sept. 2025. URL: https://openai.com/index/gdpval/.

[21] Spirit Aviation Holdings, Inc. *Notice of Auction Results and Scheduled Hearing for the Deidentified Data, including Asset Purchase Agreement and Spirit Data Categories schedule*. U.S. Bankruptcy Court, S.D.N.Y., Case No. 25-11897 (SHL), Doc. 1463, filed Aug. 14, 2026. Aug. 2026. URL: https://document.epiq11.com/document/getdocumentbycode?docId=4606206&projectCode=SPJ&source=DM.

[22] J. S. Gans. "Market Power in Artificial Intelligence". In: *Annual Review of Economics* 18 (2026), pp. 417–445. DOI: 10.1146/annurev-economics-051624-061832. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-economics-051624-061832.

[23] A. Korinek and J. Vipra. "Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence". In: *Economic Policy* 40.121 (2025), pp. 225–256. DOI: 10.1093/epolic/eiae057. URL: https://academic.oup.com/economicpolicy/article/40/121/225/7905140.

[24] C. I. Jones and C. Tonetti. "Nonrivalry and the Economics of Data". In: *American Economic Review* 110.9 (2020), pp. 2819–2858. DOI: 10.1257/aer.20191330. URL: https://doi.org/10.1257/aer.20191330.

[25] S. K. Mohan, G. Bharathy, and A. Jalan. "Enterprise Data Valuation—A Targeted Literature Review". In: *Journal of Economic Surveys* 40.1 (2026), pp. 73–92. DOI: 10.1111/joes.12705. URL: https://onlinelibrary.wiley.com/doi/10.1111/joes.12705.

[26] A. Bhutani, G. Ordoñez, and L. Veldkamp. *The Missing Value of Data*. Tech. rep. 35266. National Bureau of Economic Research, 2026. DOI: 10.3386/w35266. URL: https://www.nber.org/papers/w35266.

[27] A. Rappaport and M. J. Mauboussin. *Expectations Investing: Reading Stock Prices for Better Returns*. Revised and Updated. Columbia University Press, 2021.

[28] F. Bostoen and G. Monti. *Putting a Price on AI Training Data*. TILEC Discussion Paper 2025-10. Tilburg Law and Economics Center, 2025. DOI: 10.2139/ssrn.5840062. URL: https://ssrn.com/abstract=5840062.

[29] A. Ilyas, S. M. Park, L. Engstrom, G. Leclerc, and A. Madry. "Datamodels: Understanding Predictions with Data and Data with Predictions". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. 2022, pp. 9525–9587. URL: https://proceedings.mlr.press/v162/ilyas22a.html.

[30] S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Madry. "TRAK: Attributing Model Behavior at Scale". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. 2023, pp. 27074–27113. URL: https://proceedings.mlr.press/v202/park23c.html.

[31] M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen. "LESS: Selecting Influential Data for Targeted Instruction Tuning". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. 2024, pp. 54104–54132. URL: https://proceedings.mlr.press/v235/xia24c.html.

[32] M. Yang, R. Gao, and Z. Zheng. "Sell Data to AI Algorithms Without Revealing It: Secure Data Valuation and Sharing via Homomorphic Encryption". In: *arXiv preprint arXiv:2512.06033* (2025). DOI: 10.48550/arXiv.2512.06033. URL: https://arxiv.org/abs/2512.06033.

[33] M. Demirer, A. Fradkin, N. Tadelis, and S. Peng. *The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs*. Tech. rep. 34608. National Bureau of Economic Research, 2025. DOI: 10.3386/w34608. URL: https://www.nber.org/papers/w34608.

[34] OECD. *Theories of Harm for Digital Mergers*. Tech. rep. OECD, 2023. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/05/theories-of-harm-for-digital-mergers_7bae0553/0099737e-en.pdf.

[35] OECD. *Artificial intelligence and competitive dynamics in downstream markets*. Tech. rep. 331. OECD, 2025. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/11/artificial-intelligence-and-competitive-dynamics-in-downstream-markets_c6e81d0e/ccf0624a-en.pdf.

[36] H. Weeds. "Strategic Delay in a Real Options Model of R&D Competition". In: *The Review of Economic Studies* 69.3 (2002), pp. 729–747. URL: https://realoptions.org/papers1999/WeedsStrategic.pdf.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[38] Business Insider. *AI startup Micro1 wants to challenge Google's winning bid for Spirit Airlines data with a higher offer*. Aug. 20, 2026. URL: https://www.businessinsider.com/micro1-challenges-google-bid-spirit-airlines-data-2026-8.

[39] Bloomberg Law. *Spirit Data Sale to Google Shows Gaps in Employee Protections*. Aug. 2026. URL: https://news.bloomberglaw.com/tech-and-telecom-law/spirit-data-sale-to-google-shows-gaps-in-employee-protections.

[40] Bloomberg Law. *Google Aims to Boost AI With Purchase of Spirit Airlines Data*. Aug. 17, 2026. URL: https://news.bloomberglaw.com/mergers-and-acquisitions/google-aims-to-boost-ai-with-purchase-of-spirit-airlines-data.

[41] Alphabet Inc. *Q2 2026 Results and Form 10-Q*. 2026. URL: https://www.sec.gov/Archives/edgar/data/1652044/000165204426000066/googexhibit991q22026.htm.

[42] Google Cloud. *SWISS: Making air travel more sustainable and improving operational quality with Google Cloud*. Accessed Aug. 2026. 2026. URL: https://cloud.google.com/customers/swiss.

[43] Google Cloud. *FLYR Labs applies AI and machine learning to help airlines reach new heights*. Oct. 2021. URL: https://cloud.google.com/blog/topics/startups/improving-airline-revenue-management-with-google-cloud.

[44] International Air Transport Association. *Middle East Disruptions and High Fuel Prices Halve Airline Industry Profitability*. June 2026. URL: https://www.iata.org/en/pressroom/2026-releases/06-07-middle-east-disruptions-high-fuel-prices-halve-airline-industry-profitability/.