# FISD Alternative Data Council Buy-Side Perspective: Understanding Investment Firms' Use of Licensed Data with GenAI

## Executive Summary

The integration of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) into investment workflows has created new opportunities for data utilization while simultaneously introducing novel contractual and intellectual property considerations. This document serves as a resource for data vendors to understand how investment firms are using licensed data with AI technologies, and provides practical approaches to address common concerns in data licensing negotiations.

As a starting point, different types of AI models carry varying risk profiles and implications for data licensing. Public versions (such as free versions of ChatGPT) present higher data security and confidentiality risks as firm inputs may be used to train public models and/or generate outputs for other users. Enterprise-licensed versions hosted on private platforms (such as OpenAI offered through Microsoft Azure) provide enhanced security controls that prevent the sharing of firm inputs and typically include contractual protections against using firm data for model training. Private or proprietary versions (including open-source or firm-developed models hosted on firm infrastructure) offer the highest level of data control but require significant technical resources. Most investment firms employ enterprise or private AI implementations specifically to address data security and IP protection concerns.

As investment firms increasingly adopt GenAI technologies to enhance research capabilities, perform semantic analysis, and generate insights from licensed datasets, data vendors are expressing heightened concerns about intellectual property protection, data substitution risks, and appropriate usage restrictions.

This guide aims to bridge the knowledge gap between data vendors' IP protection objectives and investment firms' operational needs, providing a framework for balanced contractual approaches that protect vendor interests while enabling innovative data utilization in the modern investment landscape. The recommendations herein build upon FISD's longstanding best practices, including the widely adopted Derived Data definition that has guided industry practices since 2015.

## Understanding AI Technologies: Key Definitions and Distinctions

One of the most significant challenges in data licensing negotiations has been the conflation of different AI technologies, particularly the distinction between traditional machine learning (ML) and LLMs. This confusion has led to an inconsistent approach where data vendors readily permit their data to be used with traditional ML models but express heightened concerns about LLM applications. Understanding these nuances is crucial for crafting appropriate contractual terms.

https://fisd.net/alternative-data-council/

Version 1.0 March 2025

**Derived Data** consists of pricing data or other information that is created in whole or in part from the Information and that cannot be: (i) readily reverse-engineered to recreate the Information; or (ii) used to create other data that is a reasonable facsimile for the Information.[1]

**Generative AI (GenAI)** refers broadly to AI systems capable of creating new content, including text, images, or other media. While LLMs are a type of GenAI, the term encompasses various technologies beyond language models.

**Machine Learning (ML)** encompasses a broad category of computational methods that enable systems to learn patterns from data and make predictions or decisions without explicit programming for each specific task. Traditional ML applications in investment contexts include quantitative trading models, risk assessment algorithms, and pattern recognition systems. These models typically process structured data and produce specific, measurable outputs such as price predictions or risk scores.

**Large Language Models (LLMs)** represent a specific subset of AI technology designed to understand and generate human-like text. LLMs are trained on vast corpora of text data and can perform tasks such as language translation, summarization, question answering, and content generation. When investment firms use LLMs with licensed data, they are typically seeking to leverage the model's natural language processing capabilities rather than creating a substitute for the underlying dataset.

**Fine-Tuning** involves providing additional information to a pre-trained, third-party large language model to augment its capability and make it more effective for specific types of tasks or data. This process does not create a new model from scratch but rather adapts an existing model to perform better on particular datasets or use cases. Fine-tuning applications in investment contexts may include training models to better understand financial terminology and concepts, improving analysis of regulatory documents and filings, enhancing processing of earnings call transcripts and investor communications, developing better comprehension of sector-specific language and metrics, or adapting models to recognize patterns in alternative data sources. Fine-tuning enables investment firms to leverage the broad capabilities of pre-trained models while customizing them for specialized financial applications.

**Model Training** involves building an entirely new AI model using specific datasets as the primary training material. This is distinct from fine-tuning and represents a more intensive use of source data.

## Investment Management Use Cases for GenAI with Licensed Data

The following use cases illustrate how discretionary and quantitative investment managers apply LLM processing within their respective analytical frameworks. Discretionary managers typically utilize these technologies for qualitative insight extraction and narrative development, while quantitative managers generally focus on systematic signal generation and automated processing. Each approach reflects different methodologies for incorporating licensed content into investment decision-making processes.

[1] The Financial Information Services Division (FISD) created this longstanding and widely adopted definition of Derived Data. This established definition provides a practical test for determining whether LLM outputs should be considered derived data, focusing on the key questions of whether outputs can be reverse-engineered to expose underlying vendor content or serve as functional substitutes for the original licensed data. The FISD framework has proven robust across various data transformation use cases and applies equally well to AI-generated content.

The examples below represent the most common use cases, but are not intended as a comprehensive list of all possible use cases for GenAI with licensed data.

### Qualitative Research Enhancement

Discretionary managers may use LLMs to process licensed equity research, credit analyses, and industry reports to extract qualitative insights about management quality, competitive positioning, and business model sustainability that inform fundamental investment decisions. These qualitative assessments require ongoing access to detailed analyst commentary and expert research opinions from licensed providers.

### Investment Thesis Development and Testing

Licensed research reports and company analyses may be processed through LLMs to help discretionary managers develop and refine investment theses by identifying supporting evidence, potential counterarguments, and relevant precedent transactions across multiple sources. The AI-generated thesis frameworks require validation against comprehensive licensed research and cannot substitute for the detailed fundamental analysis and expert opinions that support investment decisions.

### Client and Committee Communication

Discretionary managers may utilize LLMs to synthesize licensed research and market commentary into investment committee presentations and client reports that explain investment rationale and market positioning in narrative form. These communications must be supported by authoritative research sources for fiduciary compliance, ensuring continued reliance on licensed research providers for credible investment justification.

### Signal Extraction

Quantitative managers may process licensed alternative datasets through LLMs to extract text-based signals that can be converted into quantitative factors for systematic trading models. The value of these signals depends on the proprietary data collection methodologies and specialized datasets that only licensed providers can supply.

### Research Enhancement and Semantic Search

Investment firms are increasingly using licensed datasets to perform semantic search across large document collections. This enables analysts to quickly identify relevant research across multiple data sources based on conceptual similarity rather than keyword matching alone. GenAI systems process the licensed content to understand context and meaning, allowing for more sophisticated information retrieval that enhances rather than replaces traditional research workflows.

### Content Summarization and Analysis

Licensed data is being used with GenAI systems to generate summaries, extract key insights, and identify trends across multiple sources. This application helps investment professionals process larger volumes of information more efficiently while maintaining the need to access and review the underlying licensed content. The AI-generated summaries serve as a starting point for analysis rather than a replacement for the original research.

## Addressing Vendor Concerns About LLM usage with their data

### Data Redistribution Concerns

Vendors may worry that LLMs will "memorize" their licensed content and reproduce it verbatim in outputs provided to other users. However, when investment firms use LLMs with licensed data through private implementations, the data is not being incorporated into publicly accessible models. The LLM is being used as a processing tool within the firm's controlled environment, similar to how traditional ML models process licensed data to generate insights. Additionally, the concern about vendor data being used with other users via the LLM applies uniformly across all buy-side firms implementing similar enterprise AI solutions with appropriate security controls.

### Seat and Usage Limitations

A related vendor concern involves whether traditional seat-based or named-user licensing models remain viable when licensed data is processed through firm-wide LLM systems that could theoretically be accessed by any employee. Investment firms can address this concern via contractual solutions and auditing capabilities that preserve the integrity of existing licensing models while enabling LLM-enhanced workflows.

### Fine-Tuning Data Persistence

When firms fine-tune models using licensed data, vendors may be concerned that the licensed content becomes permanently embedded in the model parameters, making post-termination data deletion ineffective. However, fine-tuned models typically learn patterns and relationships rather than storing verbatim content, and the original licensed data can still be deleted from firm systems upon contract termination. Additionally, fine-tuning for investment use cases generally focuses on improving the model's ability to understand financial terminology and concepts rather than memorizing specific vendor content, limiting the risk of data reconstruction from model outputs. This concern could also be addressed through contractual prohibitions.

### Data Deletion and Model Rollback Concerns

While vendors have traditionally required customers to delete raw licensed data upon agreement expiration, some now seek broader obligations that extend to "model artifacts" which may embed licensed data. A balanced approach to data deletion recognizes both vendor IP protection needs and the commercial impracticality of comprehensive model rollback. Investment firms can reasonably commit to certifying deletion of raw vendor datasets and removing artifacts that could serve as functional substitutes for the raw dataset, such as high-fidelity embeddings designed for Retrieval-Augmented Generation. However, requiring customers to retrain models or recalculate outputs solely to excise historical use of vendor data is commercially impracticable, can compromise audit trails, and may conflict with regulatory record-keeping requirements. Artifacts that cannot be reverse-engineered to expose underlying vendor data and do not provide a market substitute should remain the customer's property for ongoing use.

### Substitution Risk

Vendors have also expressed a concern that LLM outputs could serve as substitutes for licensed datasets. However, this fails to recognize that investment firms use LLMs to enhance their analysis of licensed content, not to replace it. An LLM-generated summary of research reports, for example, cannot substitute for the detailed analysis, methodologies, and data points contained in the original reports. The outputs are derivative analytical products that maintain (and in some cases enhances) the investment firm's need for ongoing access to the underlying licensed content.

**Premium Data Displacement Concerns**

A more nuanced substitution concern involves whether firms might extract insights from premium datasets through LLM processing and then apply those insights to enhance cheaper, alternative data sources, effectively reducing their reliance on the premium provider. While this scenario is theoretically possible, it faces practical limitations: extracted insights typically represent historical patterns that require ongoing validation against current premium data, competitive advantages in premium datasets often stem from real-time updates and proprietary methodologies that cannot be replicated through LLM processing of historical content, and regulatory requirements for investment decisions generally demand ongoing access to authoritative, up-to-date sources rather than reliance on previously extracted insights. Additionally, investment firms typically do not “turn off” or unsubscribe from datasets once they have begun using them in their investment processes, as doing so would create operational risks and potential compliance issues.

## Functional Equivalence in Risk Profile for ML and LLM Usage

From a data protection and IP perspective, both traditional ML and LLM applications share several key characteristics when used by investment firms.

**No Public Distribution:** Investment firms use both ML models and LLMs internally to enhance their analytical capabilities, not to create competing products for external distribution.

**Controlled Environment:** Both applications occur within the investment firm's controlled technological environment with appropriate security and access restrictions.

**Derived Outputs:** Both technologies generate insights and outputs that are derived from, but not substitutable for, the original licensed data. The outputs generated through GenAI applications generally cannot be reverse-engineered to expose the underlying vendor data and do not provide a functional substitute for licensed content.

**Continued Data Dependency:** Both use cases maintain the investment firm's ongoing need for access to current and updated licensed data.

The primary difference lies in the type of processing capability (statistical pattern recognition vs. natural language processing) rather than the fundamental risk profile regarding data protection and IP concerns.

It is important to note that LLM usage should not be conflated with the derived data ownership rights that have been established and maintained over the past decade+ through established industry practices. The application of LLM technology to licensed data represents a processing methodology rather than a change to fundamental data ownership principles.

## Contractual Approaches

**Derived Data and Output Ownership**

A balanced approach to LLM output restrictions may involve evaluating two key factors: whether the vendor's licensed data can be reverse-engineered from the output, and whether the output data could function as a material substitute for the vendor's original data. If the output cannot be readily reverse engineered and cannot be used as a substitute for the vendor's original data, usage and ownership of the output should not be restricted. This framework attempts to balance vendor IP protection concerns with investment firms' operational requirements for utilizing insights derived from licensed content.

Contractual provisions may address ownership rights to inputs and outputs generated through AI applications, and could include restrictions preventing AI providers from using such information to train their own models or incorporating firm data into outputs provided to other users. It is important to emphasize that buy-side firms are not sharing vendor data externally through these applications - the processing occurs within controlled environments for internal analytical purposes only. The specific terms would depend on the particular circumstances and risk tolerance of both parties.

**Security and Confidentiality Considerations**

Data vendors' security concerns can be addressed through appropriate contractual provisions regarding data access, storage, and processing parameters. Investment firms may describe their AI implementations and security measures, which often include solutions such as private cloud environments, encryption keys, and other enterprise-grade security controls.

Contractual language may specify whether the vendor's original data will be confined within the firm's IT environment or transferred to, stored, or processed within the AI provider's systems. Nearly all enterprise GenAI providers offer solutions subject to SOC2 and other relevant information security frameworks, which can provide additional assurance to vendors.

## Conclusion

The integration of GenAI technologies into investment workflows represents a natural evolution of how licensed data is used to generate insights and enhance decision-making capabilities. While data vendors' concerns about IP protection are legitimate, overly restrictive contractual terms that fail to distinguish between legitimate enhancement use cases and actual substitution risks can impede innovation and create unnecessary commercial friction.

By understanding the technical distinctions between different AI technologies, recognizing legitimate investment use cases, and implementing balanced contractual approaches that protect IP while enabling innovation, data vendors and investment firms can establish licensing frameworks that serve both parties' interests in the evolving AI landscape.

https://fisd.net/alternative-data-council/

Version 1.0 March 2025

The key to successful licensing relationships in the GenAI era lies in education, clear communication, and contractual precision that addresses actual risks rather than perceived threats. As both industries continue to adapt to technological change, collaborative approaches to licensing will enable the continued growth and innovation that benefits all market participants.

https://fisd.net/alternative-data-council/

Version 1.0 March 2025