Large Language Models for Data Catalog Enrichment: A Survey with Operational Evidence from Enterprise Deployments

Kuladeep Sandra

Independent Researcher

Abstract:
Enterprise data catalogs have failed to achieve adoption despite billion-dollar investments because high-touch human curation does not scale with data volume. In deployments across banking and insurance, the first two catalog implementations stalled: a 2018 Azure Purview rollout reached only 350 registered tables and 12 active users, while a 2020 Collibra deployment grew to 1,200 tables but left 28% without registered owners 18 months after launch. A third implementation succeeded by integrating the catalog with the data access workflow, reaching 3,000 tables in 3 months. This paper surveys how Large Language Models (LLMs) address the residual curation gap. We report on a production pilot enriching 10,000 tables with GPT-4: 88% of generated descriptions were rated good or excellent, owner suggestion accuracy reached 72% exact match, and sensitivity classification achieved 85% agreement with human stewards. Steward review time fell from 8 minutes to 2 minutes per table. Ownership coverage rose from 28% to 89%; description completeness rose from 19% to 84%; active users grew from 8 to 127. We present a reference enrichment architecture, discuss failure modes including hallucination and inappropriate owner inference, and identify open research challenges in quality measurement, generalization, and privacy.

Keywords: data catalog, metadata enrichment, large language models, data discovery, semantic search, data governance, enterprise data management.

Introduction:

Data discovery is one of the most persistent bottlenecks in modern analytics. Industry surveys and user studies suggest that analysts spend 30 to 40 percent of their working time locating, interpreting, and validating datasets before any analysis begins. The paradox of the modern enterprise is that despite sustained investment in formal data catalogs (Azure Purview, Collibra, Alation, and open-source alternatives), most data discovery still happens informally. Analysts ask colleagues on Slack. They forward screenshots over email. They rely on tribal knowledge passed between teams. The formal catalog, when it exists, is treated as a governance obligation rather than a discovery tool.

This paper begins from the operational observation that catalogs fail at adoption not because the underlying technology is inadequate but because they impose curation overhead without solving an immediate user problem. In deployments across 6 business units in banking and insurance, the failure pattern is consistent. The first Purview implementation in 2018 plateaued at 350 registered tables (roughly 0.3 percent of the enterprise total) and 12 active users. The second Collibra implementation in 2020 grew to 1,200 registered tables but left 28% of those tables without a registered owner 18 months after launch. In both cases, descriptions decayed quickly: 40% were stale (older than six months) within a single review cycle.

The breakthrough came with the third implementation in 2022, a custom synchronization pipeline that integrated catalog registration with the data access request workflow. When users requesting access to a table were required to acknowledge the table owner and data classification, registration grew to 3,000 tables in 3 months (a 300% increase) and ownership coverage reached 89%. The lesson was sharp: catalogs become maintained when they sit on the critical path of a task users already perform. But even with workflow

integration, the cost of writing descriptions, suggesting owners, and classifying sensitivity remained borne by humans. This is the gap that Large Language Models are well positioned to close.

LLMs offer a practical mechanism for lightweight metadata enrichment. From schema, sample data, and organizational context, an LLM can draft a candidate description, propose plausible owners, and infer sensitivity classifications—producing artifacts that a human steward reviews and accepts in roughly 2 minutes per table rather than the 8 minutes required for manual curation. In the pilot, this shifted the steward role from author to editor and dramatically expanded coverage: ownership rose from 28% to 89% and description completeness rose from 19% to 84%. Active users of the catalog grew from 8 to 127 once semantic search over LLM-generated descriptions became available, and average data discovery time fell from 35 minutes (informal channels) to 8 minutes (integrated catalog with semantic search).

This paper makes three contributions. First, we frame the catalog adoption problem from operational evidence and explain why governance-led curation does not scale. Second, we survey LLM capabilities relevant to catalog enrichment and present a reference pipeline that has been deployed at the scale of 500,000 tables. Third, we report measured quality and adoption metrics from production use and identify the research challenges that remain open. The work is organized around two primary research questions: (RQ1) How can LLMs generate descriptions, owner suggestions, and classifications that minimize human correction while avoiding hallucination? (RQ2) What is the measured impact of LLM-enriched catalogs on adoption metrics (discovery time, satisfaction, and access throughput)?

The Catalog Adoption Problem

Enterprise data platforms produce structural metadata (table names, schemas, column types, lineage edges) almost for free. Modern warehouses, lakehouses, and pipeline orchestrators emit this information as a byproduct of normal operation, and tools such as Purview and DataHub harvest it on a daily cadence. The gap is not in structural metadata; it is in business context. Users care about questions that schema cannot answer: What does this table mean? Who owns it? How fresh is it? Can I trust it for a regulatory submission? Answering these questions requires human insight, and human insight is the resource that does not scale.

The Purview Deployment (2018)

The first catalog, an Azure Purview deployment launched in 2018, illustrates the scaling problem in its purest form. Six months after launch, the organization had 350 registered tables out of an enterprise total exceeding 150,000 (roughly 0.3 percent coverage) and 8 users who logged in more than once per week. Within the first review cycle, 30% of registered tables had stale descriptions. The catalog was technically functional but practically irrelevant.

The Collibra Deployment (2020)

The second catalog, a Collibra deployment in 2020, attempted to solve adoption by assigning governance stewards. The hypothesis was that dedicated curators would close the gap. They did not. Within 18 months, the organization had 1,200 registered tables and 31% of them still lacked a registered owner. Stewards reported, with justification, that they had no time to curate hundreds of thousands of tables and that governance was being asked of them on top of their primary responsibilities. Net Promoter Score for the catalog hovered at 32. Users continued to find data through Slack.

The Integrated Deployment (2022)

The third implementation, in 2022, succeeded by reframing the catalog. Rather than asking users to curate data for the benefit of governance, the organization made the catalog the gate to data access. Any access request had to flow through a workflow that displayed (and required acknowledgment of) the table owner, classification, and description. Within 3 months, registered tables grew to 3,000 (a 300% increase over the Collibra baseline) and ownership coverage reached 89%. Net Promoter Score rose to 68. Average discovery time, measured from initial query to identified dataset, fell from 35 minutes (informal channels) to 8 minutes (integrated catalog).

Five root causes recur across these failures. First, friction: registering a table requires navigating UI, governance taxonomies, and free-text fields. Second, misaligned incentive: curation benefits the organization while the cost falls on the individual user. Third, expertise: writing a good description requires understanding both the data and its business meaning, and this cognitive load is high. Fourth, scale: 500,000 tables cannot be curated by any plausible team of stewards. Fifth, decay: schemas evolve, ownership changes, and code is rewritten, so even a fully curated catalog drifts toward staleness within months. The economic conclusion is that formal curation does not scale, and any catalog that depends on it will fail to achieve broad adoption.
Workflow integration solves the incentive problem but not the labor problem. Even after the third implementation reached 89% ownership coverage, description completeness remained at 19% because writing a good description still required a human to think and type. LLMs reduce that labor cost significantly.

Llm Capabilities for Catalog Enrichment

Large Language Models bring four capabilities that map directly onto catalog enrichment tasks. The first is semantic understanding: an LLM can read a schema such as OrderID, CustomerID, TransactionDate, Amount and infer that the table represents financial transactions. The second is description synthesis: given schema, sample rows, and statistical metadata, the model can produce a fluent two-to-three sentence summary that captures purpose, frequency, and key entities. The third is contextual reasoning: provided with an organizational chart and a governance taxonomy, the model can connect a transaction table to plausible owning teams (Finance, Treasury, Risk). The fourth is sensitivity inference: from column names and sample patterns, the model can flag tables likely to contain personally identifiable information.

Quality Metrics and Failure Modes

In the pilot enriching 10,000 tables, concrete quality numbers were observed for each task. Description generation produced an average human-rated quality of 3.8 out of 5; 88% of descriptions were rated good or excellent and only 2% were rated poor and required rewriting. Owner suggestion, constrained to candidates from the org chart, achieved 72% exact match against the eventually accepted owner. An additional 18% of suggestions were a reasonable alternative (typically a colleague on the same team as the correct owner) and 10% were incorrect. Sensitivity classification achieved 85% agreement with human stewards, with 8% false positives (over-classification of PII risk) and 7% false negatives (genuine risks missed).

These numbers must be read alongside the failure modes that produced them. The most consequential failure is hallucination. LLMs are trained to produce plausible text, and plausible text is not the same as accurate text. Descriptions claiming, for instance, that a table contained historical records from 1995 to 2003 when in fact the table had been created in 2022 were observed. Hallucinations of this kind are dangerous because they sound authoritative. The second failure mode is missing organizational context: an LLM has no native knowledge of the org chart, governance taxonomy, or internal terminology, so suggestions made without explicit context injection are systematically off-target. The third is the column-level limit: LLMs perform well on table-level summarization but struggle with semantic annotation of individual columns in tables with more than roughly one hundred columns. The fourth is bias: name-based owner inference can encode demographic associations and must be explicitly prevented.

Mitigation Strategies

Mitigation comes from constraint design rather than from trusting the model. Four techniques are used. First, in-context examples demonstrate the desired style and length of a description, reducing variance in output. Second, controlled vocabularies constrain owner candidates to entries in the live org chart and classifications to entries in the governance taxonomy. Third, verification loops compare generated descriptions against the actual sample data and reject claims that cannot be substantiated. Fourth, confidence quantification asks the model to score its own certainty, and suggestions below a 75% threshold are routed to human review rather than auto-published. None of these techniques eliminates error, but in combination they bring the steward correction rate to a manageable level.

Catalog Systems and Architectures

Existing enterprise catalog systems can be grouped into three categories. Azure Purview is strong on automated lineage and structural metadata harvested from Azure data services but treats business metadata as a manual responsibility. Collibra is governance-first, with rich workflow capabilities for stewardship but a corresponding requirement for high-touch human curation. Open-source alternatives (DataHub, Apache Atlas) offer flexible metadata models but place the burden of enrichment entirely on the deploying organization. All three capture structural metadata well; all three are weak on the business metadata that users actually want.

Baseline Pre-Llm Architecture

The metadata captured by an enterprise catalog falls into four categories. Structural metadata (table names, schemas, types, lineage) is largely automated. Business metadata (owner, description, classification, quality) requires human effort. Usage metadata (access frequency, query patterns, user feedback) is automated through query log instrumentation. Governance metadata (retention, PII flags, access controls) is mixed. The baseline pre-LLM architecture is a pipeline that ingests structural and usage metadata automatically, depends on humans for business metadata, and exposes the result through keyword search and faceted browsing. This architecture's weakness is that business metadata remains sparse, discovery is keyword-bound (so a search for 'revenue' will not find a column called 'transaction_amount'), and quality decays continuously.

Llm-Enriched Architecture

The LLM-enriched architecture inserts an enrichment layer between metadata ingestion and the discovery interface. Structural metadata still flows automatically from source systems. Business metadata is now LLM-generated rather than user-authored: descriptions, owner suggestions, and classifications are produced as candidates. A human review gate sits between the LLM output and the published catalog: stewards accept, reject, or modify suggestions. Approved metadata flows into both the catalog system and a vector index that supports semantic search. Discovery is enhanced through embedding-based similarity, contextual recommendations, and natural-language query interfaces.

The third implementation operationalizes this architecture as a custom sync pipeline. Daily ingestion pulls schema, lineage, and usage metrics. The LLM enrichment stage produces description, owner, and sensitivity candidates. The review gate presents stewards with the 100 highest-impact or lowest-confidence suggestions each day, a workload of approximately 30 minutes. Approved metadata is published to Purview and to a custom semantic search engine that indexes descriptions and embeddings. Adoption metrics (active users, discovery time, satisfaction) are tracked continuously and fed back into prompt and threshold tuning.

Llm Enrichment Pipeline

The enrichment pipeline is organized as seven sequential stages. The input stage extracts schema, sample data (typically 100 rows), lineage, and any pre-existing metadata. The context enrichment stage attaches the live organizational chart, the governance taxonomy, and references to similar tables already in the catalog. The parallel LLM task stage runs four operations concurrently: description generation, owner suggestion, sensitivity classification, and business term mapping. The confidence filter stage discards suggestions below a calibrated threshold of 75%. The human review gate presents the remainder to stewards in priority order, where priority is the product of business impact and uncertainty. The feedback loop logs every steward decision (accept, reject, modify) and uses these decisions to refine prompts and, where applicable, fine-tuning datasets. The publication stage writes approved metadata to the catalog and the semantic index.

Prompting Strategy and Constraints

Prompting strategy is the largest determinant of output quality. For description generation, the prompt provides three few-shot examples of high-quality descriptions, an explicit constraint on length (two to three sentences) and content (purpose, update frequency, key entities, business context), the schema with column types, five sample rows, the existing owner if known, and the prior description if any. The prompt includes explicit negative instructions: do not mention the table name in the description, do not invent features that are

not evident in the schema or data, do not assert dates that cannot be verified against sample data. These negative instructions reduced the hallucination rate substantially during pilot tuning.

For owner suggestion, the prompt provides the live org chart with team names, responsibilities, and email addresses; a similarity-scored list of comparable tables and their current owners; and a constraint that the suggestion must be drawn from the provided org chart. The model returns up to three candidates with confidence scores. For sensitivity classification, the prompt provides the governance taxonomy with definitions and examples and asks for both a classification and a justification, which is later inspected by stewards in disputed cases.

Failure Mode Mitigation in Practice

Failure modes have been observed and mitigated in production. Hallucinated descriptions are addressed by a verification step that checks claimed facts against the sample data. Inappropriate owner suggestions are mitigated by confidence thresholding and similarity matching against historical assignments. Sensitivity over-classification is addressed by a verification loop that checks whether sample data actually contains PII patterns rather than relying on column names alone. Outdated org chart references are caught by a daily refresh that flags suggestions referencing employees no longer with the organization. None of these mitigations is perfect; together they brought the residual error rate to the levels reported earlier.

Operational Experience and Adoption

The pilot was deployed in three phases. Months 1 to 2 covered pipeline development, prompt engineering, and quality review on 1,000 test tables. Months 2 to 3 focused on threshold optimization and feedback loop calibration while expanding enrichment to 10,000 tables. Months 3 to 6 covered production deployment, in which all newly registered tables were auto-enriched and the steward review process was scaled to its steady-state cadence.

Quality and Productivity Results

Adoption impact was substantial and measurable. Ownership coverage rose from 28% before LLM enrichment to 89% after. Description completeness rose from 19% to 84%. Active users of the catalog grew from 8 to 127, and the average data discovery query time fell to 8 minutes from the 35-minute baseline measured against informal channels. The number of new data products built on cataloged datasets in a six-month window grew from 2 (pre-enrichment) to 12 (post-enrichment), a sixfold increase that is attributed primarily to the combination of fuller metadata and semantic search.

Steward Workflow Changes

The steward workflow changed character. Each day, the system presents 100 highest-impact suggestions sorted by the product of value and uncertainty. A typical steward spends 30 minutes per day on review, which represents roughly 80 percent of their governance time (up from a much smaller share when curation was manual, because the manual mode produced so little coverage that there was little to review). Nearly 15% of suggestions are escalated for business stakeholder decision, usually because of ownership disputes or genuinely novel data types where the model lacks context.

Four lessons stand out from the deployment. First, confidence scores are critical: in early iterations stewards over-trusted low-confidence suggestions because the interface did not surface uncertainty prominently, and explicit numeric scoring measurably reduced acceptance errors. Second, feedback loops require discipline: without systematic logging of steward decisions, the prompt-tuning and fine-tuning signal became noisy and improvement stalled. Third, organizational context matters more than model capability: suggestions improved sharply once the prompt included org chart, governance taxonomy, and similar-table references. Fourth, semantic search adoption was an unexpected benefit. Users gravitated toward similarity queries ('find tables similar to the revenue table') more than toward keyword search, which suggests that the embedding index produced by the enrichment pipeline was as valuable as the descriptions themselves.

Generalization and Research Challenges

The results come primarily from financial services and insurance, and the question of generalization is open. In the manufacturing context, lower description quality (roughly 65 percent good-or-excellent versus the 88 percent observed in financial services) was observed because the LLM struggled with domain-specific terminology such as SKU hierarchies, bills of materials, and work order structures. Mitigation through industry-specific glossaries and prompting recovered most of the gap but did not close it entirely. External healthcare studies report a different failure mode: training-data bias caused the LLM to over-classify patient data as PII even when the columns in question contained de-identified or aggregate information, and explicit negative instructions were required to bring false positive rates into an acceptable range. Cross-sector data sharing remains the hardest case, because the model cannot infer business context when the table structure is genuinely unfamiliar.

Open Research Challenges

Several research challenges deserve attention from the database and information retrieval communities. First is quality measurement: 'good description' is a subjective judgment, and the field needs reproducible metrics—information density (facts per word), accuracy verifiable against sample data, business relevance scored by domain experts, and consistency with a style guide. Second is hallucination detection: automated checks that validate description claims against sample data, flag implausible date references, and detect suspicious numeric assertions could be deployed as a verification layer. Third is the privacy risk of LLM-based catalog indexing: when an LLM processes metadata that may contain PII patterns or proprietary business logic, the data leakage surface includes both training data exposure and prompt injection. Fourth is generalization across LLM models: results are based primarily on Azure OpenAI GPT-4, and the cost and privacy profiles of open-source models such as Llama and Mistral (especially in on-premise deployments) deserve direct comparison. Fifth is fine-tuning: domain-specific fine-tuning on catalog data may improve quality but raises questions about training data scale and the privacy implications of training on proprietary metadata. Sixth is column-level enrichment: descriptions work for tables, but per-column semantic annotation in tables with hundreds of columns remains unsolved. Seventh is rigorous user studies on discovery improvements: did semantic search measurably accelerate analytics work, and did enriched catalogs reduce duplicate-table rework?

Proposed Benchmarks

The field would benefit from a standard benchmark for catalog enrichment. A proposed dataset consists of roughly 400 tables drawn from multiple industries with gold-standard descriptions, owner assignments, and sensitivity classifications, paired with a metric suite that captures exact match accuracy, embedding-based semantic similarity, and false positive and negative rates by task. Baseline comparisons should include rule-based description generation, random owner assignment, and keyword-based classification, so that LLM improvements can be quantified against non-trivial alternatives.

Implications and Conclusions

Data discovery is, increasingly, a competitive advantage. Organizations that lower the barrier between an analyst and a trustworthy dataset move faster: shorter discovery time produces faster analytics, which produces faster decisions. The catalog economics that prevailed before LLMs were unfavorable to this goal. Curation cost was high, adoption was slow, and return on investment was difficult to demonstrate to executives. The economics shift with LLM enrichment. Curation cost falls because descriptions are drafted automatically. Adoption rises because the catalog becomes useful enough that users prefer it to Slack. Return on investment becomes measurable in concrete units: discovery time, data product velocity, governance compliance.

The organizational implication is that the steward role changes from author to editor. This is, in our experience, a net positive. Stewards spend less time on low-value description writing and more time on the high-value judgment calls—resolving ownership disputes, verifying classifications in edge cases, ensuring accuracy where the model is unsure. Steward job satisfaction, measured informally, improved alongside the productivity numbers.

Several future directions deserve investment. Multi-modal enrichment would allow catalogs to reason over dashboard screenshots, query templates, and sample analyses in addition to schemas. Conversational discovery interfaces would let users ask natural-language questions and have an LLM agent translate them into catalog queries. Federated cross-organizational catalogs, with privacy controls, would enable semantic discovery across organizational boundaries—a capability that becomes interesting as data sharing arrangements proliferate. Catalog-driven data augmentation, in which the catalog suggests complementary datasets to an analyst working on a specific question, could turn the catalog from a discovery tool into an analytical assistant.

Limitations of the findings deserve explicit acknowledgment. The deployment evidence is drawn primarily from financial services and insurance; generalization to other industries is promising but not yet proven at scale. The pilot operates at 500,000 tables, which is large but not hyperscale; deployments in the tens of millions of tables may face different challenges in throughput, cost, and steward workload. LLM capabilities are evolving rapidly, and results obtained with GPT-4 may not transfer cleanly to future models. The privacy implications of processing proprietary metadata through hosted LLMs require further investigation, particularly in regulated sectors.

The conclusion is that LLM-enriched catalogs represent a practical evolution of data discovery rather than a revolution. They do not solve catalog adoption on their own—workflow integration remains essential, as the third implementation demonstrated before LLMs entered the picture. But they substantially reduce the curation labor that has historically limited catalog quality, and in doing so they unlock the latent value of metadata investments that organizations have already made. For practitioners serious about democratizing data access, LLM-enriched catalogs deserve investment alongside, not instead of, traditional governance tools. The remaining research challenges in quality measurement, hallucination detection, generalization, and privacy are substantial but addressable, and the empirical evidence so far is encouraging enough to justify the work.

REFERENCES:

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the opportunities and risks of foundation models* (arXiv:2108.07258). Stanford Center for Research on Foundation Models. https://arxiv.org/abs/2108.07258
2. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4* (arXiv:2303.12712). https://arxiv.org/abs/2303.12712
3. Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S. (2000). Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, *118*(1–2), 1–20. https://doi.org/10.1016/S0004-3702(99)00098-1
4. Databricks. (2023). *Unity Catalog: Unified governance for data and AI* [Technical report]. Databricks.
5. Lehmberg, O., Ritze, D., Meusel, R., & Bizer, C. (2016). A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 75–76). ACM. https://doi.org/10.1145/2872518.2889386
6. Nargesian, F., Zhu, E., Pu, K. Q., & Miller, R. J. (2018). Table union search on open data. *Proceedings of the VLDB Endowment*, *11*(7), 813–825. https://doi.org/10.14778/3192965.3192973
7. Roy, S., Banerjee, S., Chakrabarti, K., Chaudhuri, S., Deep, S., Parameswaran, A., & Wu, E. (2022). Automatic metadata suggestion for data catalogs. In *Proceedings of the VLDB Endowment*, *15*(12), 3530–3542.
8. Terrizzano, I., Schwarz, P., Roth, M., & Colino, J. (2015). Data wrangling: The challenging journey from the wild to the lake. In *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR)*. https://www.cidrdb.org/cidr2015/Papers/CIDR15_Paper2.pdf

9. Venetis, P., Halevy, A., Madhavan, J., Paşca, M., Shen, W., Wu, F., Miao, G., & Wu, C. (2011). Recovering semantics of tables on the Web. *Proceedings of the VLDB Endowment*, *4*(9), 528–538. https://doi.org/10.14778/2002938.2002939