AI beyond the hype

Insights from ~ 100
AI & Data Practitioners

AI Applications in SMEs

Why RAG and Agentic AI do not reach production - and how to change that

Open Research Report
by PandasAI

Location: Munich, Germany

Date: July 7, 2025

Creator's Name: Sinaptik GmbH

Link to Creator's Website: https://pandas-ai.com

Contributors: see section "2. Acknowledgements & Credits"

Abstract

This report presents insights from approximately 100 AI and data professionals working on real-world AI applications in small and medium-sized enterprises (SMEs), which do not own AI infrastructure and rely heavily on cloud-based AI services and consultancy-driven implementation. This dependency creates a fragile ecosystem where few AI initiatives reach production, hindered by unclear business requirements, inflated expectations, and limited budgets. Consulting firms dominate implementation, typically starting with Proofs of Concept (PoCs) ranging from €40K to €100K, which rarely translate into production-ready systems. There are rare occasions where consultancy-driven PoCs reach prod, when results are rigorously evaluated before deployment, leading to satisfactory performance and acceptable KPIs in end-user satisfaction (e.g. NPS, CSAT).

Functional areas have high yet unclear expectations for AI outcomes, demanding agentic functionalities (e.g. web search, API connections), but find model output unreliable and unpredictable. AI startups continue pursuing aggressive growth strategies despite financial losses, while freelance and boutique firms flood the market with LLM wrappers and RAG solutions of variable quality.

For both Agentic AI and retrieval-augmented generation (RAG) systems, challenges in memory, knowledge structuring, and data freshness persist while. Open-weight models are viewed as underperforming in accuracy, with deployment considered risky and technically demanding. Fine-tuning and training smaller models are seen as nearly taboo due to lack of expertise and perceived unpredictability, preventing SMEs from aligning model output and serving application-specific models.

Overall, the report highlights a fragmented, consultant-driven landscape marked by fragile pipelines and systemic barriers to production. Practical AI in SMEs remains a work in progress, with significant technical, organisational, and economic challenges to overcome.

1. Purpose and Methodology

The purpose of this research is to examine AI adoption challenges within SMEs and to identify why advanced AI technologies like Retrieval-Augmented Generation (RAG) and agentic AI systems fail to reach production deployment.

Through in-depth interviews with around 100 AI and data practitioners—engineers, developers, and technical leads directly involved in implementing AI solutions across the EU and UK—we investigated the gap between AI's theoretical potential and its real-world application. Since the vast majority of participants were Italian, the study provides reliable insights for the Italian market. While respondents from other EU countries appear to follow similar patterns, there is insufficient data to confirm that these findings are valid for the entire European landscape.

Interviews were conducted through video calls or in-person meetings using an open-question format that encouraged candid dialogue about technical challenges, business constraints, and implementation failures. Detailed notes were systematically clustered and organised thematically to identify recurring patterns and common pain points across different organisations.

This qualitative analysis approach was specifically chosen to surface the nuanced, often unspoken struggles that quantitative surveys might miss, providing an authentic picture of the practical barriers to AI production deployment in the SME environment.

As a company developing open source dev tools, our goal was product discovery—learning the challenges of AI engineers and gathering insights to build products and dev tools that deliver business outcomes at scale. Since open source is core to our philosophy, this product research is also "open" and part of our "open discovery" initiatives, with all findings shared with the AI community and AI builders.

Ultimately, we aim to contribute to the EU AI ecosystem towards a future where AI is owned and not just rented, avoiding a dangerous monopoly. By understanding the real barriers to AI deployment and developing solutions that address these challenges, we seek to empower European businesses with the tools and knowledge needed to ensure technological sovereignty.

2. Acknowledgements & Credits

This work is fundamentally collective, made possible thanks to the many professionals who generously accepted to be interviewed and shared their experiences, challenges, and insights. The interviewer, Della Corte Giuseppe, *Founding Product Researcher* at PandasAI (YC W24), performed only the work of clustering and organising these insights into coherent themes and findings.

Not everyone participating in this research chose to make their name public, and we deeply thank these anonymous participants whose contributions are equally valuable to the overall findings. Their willingness to share honest perspectives, often about sensitive topics like project failures and organisational challenges, has been essential to the authenticity of the research.

Hereby follows the list of AI and data professionals who took part in this open research and allowed to be listed as contributors, listed in alphabetical order: *Alberton, Stefano; Amadio, Riccardo; Atrio R, Àlex; Borelli, Lorenzo; Bruni, Lorenzo Maria; Cerza, Silvano; Colombo, Maria Cristina; Etta, Gabriele; Goyal, Kanika; Guglielmi, Daniele; Iaquinta, Tommaso; Lewis, Damian; Lombardi, Tommaso; Mariani, Giacomo; Maidana, Facundo Nicolas; Melissari, Paolo; Morelli, Pierandrea; Navas Alejo, Irean; Reitano, Stefano; Saleem, Arslan; Sanges, Adriano; Sebastianelli, Nicola; Tran, Dat; Trovato, Matilde*

3. AI Projects Externalization

The vast majority of SMEs rely on third-party providers for their AI projects, creating a dependency that fundamentally shapes how AI initiatives are conceived and executed. Within corporate structures, organisations tend to merge "innovation" and "data/AI" teams, leading to an approach where AI projects emerge from a hybrid of bottom-up and top-down directives. On one hand, Boards of Directors mandate finding AI use cases for competitive advantage, while on the other, innovation managers launch campaigns to gather use cases from functional areas across the organisation.

This organisational gap presents a prime entry point for consultancy firms, which position themselves as bridges between strategic directives and operational needs. These firms typically sell education services to functional areas while simultaneously supporting data and innovation teams in gathering insights and identifying opportunities. For consultancy firms, this educational phase often serves as an entry point to subsequently develop and sell Proofs of Concept (PoCs). However, these PoCs rarely transition to production environments, creating a cycle of investment without measurable business outcomes.

Most SMEs lack the internal competencies to evaluate and select appropriate AI providers, resulting in two common decision-making patterns: either choosing the cheapest available option or, when existing trust relationships exist, requesting comprehensive data and AI strategy services from established strategy firms. The quality of strategic advisory services varies drastically depending on the competence of the consulting team assigned to the project. A recurring complaint from SMEs is that strategic advice rarely translates into tangible products or deployable solutions. In exceptional cases, medium enterprises establish partnerships with consultancy firms to actually deploy AI products after PoC, but even these collaborations typically yield long-term impacts that remain difficult to measure and quantify.

A notable exception to this consulting-dominated landscape comes from specialised AI startups offering vertical-specific products, particularly in domains like legal services, where ready-to-deploy solutions provide more immediate and measurable value to organisations.

Overall, this externalisation pattern creates a fundamental disconnect between AI investment and business outcomes. SMEs find themselves caught in a cycle where strategic mandates drive AI initiatives, but the resulting implementations rarely align with operational realities or deliver measurable value. The dominance of consultancy firms in this space, combined with the lack of internal AI competencies, perpetuates a system where organisations remain dependent on external providers while struggling to build sustainable, production-ready AI capabilities that truly serve their business objectives.

4. Centralisation without Ownership

Medium enterprises, especially in highly regulated industries, attempt to divide core AI development from external providers—a trend that appears counterintuitive given section 3 of this report. On one hand, consultancy firms dominate the development of AI PoCs, while on the other, top-down plans to create internal AI projects keep emerging. These internal, long-term and large-scale projects are championed by internal tech leads and AI engineers, but are often perceived as unaligned with the needs of business areas.

This misalignment creates a paradoxical situation where business decision makers tend to prefer externalisation (section 3) for most use cases, while internal technical decision makers gravitate toward creating something they consider more impactful and technologically challenging. However, these ambitious long-term projects often end up being useless as the teams struggle to keep pace with rapidly evolving AI technologies, resulting in underperforming outcomes that fail to meet initial expectations.

Compounding this issue, most internal teams lack the specialised skills required to build application-specific AI models, ultimately resorting to limited ownership approaches. AI models are served via major cloud providers—primarily Azure, Google Cloud, or AWS—creating a dependency that undermines the original goal of internal control. This reliance on external infrastructure makes it difficult, if not impossible, to load balance LLM requests effectively, implement real-time correction of poor model outputs, serve application-specific models tailored to business needs, or optimise inference speed for production requirements.

The result is a centralised approach that provides neither the agility of external solutions nor the true ownership and control that internal development was meant to achieve, leaving organizations trapped in a middle ground that delivers limited value while consuming significant internal resources.

AI engineers consistently report significant struggles with fine-tuning or distilling models for specific applications, citing lack of expertise, unpredictable outcomes, and resource constraints as primary barriers. Faced with these limitations, teams resort to extensive prompt engineering and orchestration techniques applied to third-party models, attempting to coax application-specific behaviour from general-purpose systems. While these workarounds can yield some improvements, they represent a fundamental compromise that prevents organisations from achieving the performance, cost efficiency, and control that properly customised models could provide. This reliance on external model orchestration rather than internal model development further reinforces the dependency cycle that centralisation was meant to break.

5. AI Startups at a Loss

Behind the dismissive "Is your startup yet another GPT wrapper?" lies a more nuanced reality where AI startups predominantly focus on the application layer, building specialised solutions that leverage large language models as foundational infrastructure rather than core differentiation. These startups typically identify specific use cases and develop either agentic applications or RAG-based systems, with the real value creation happening in user experience design, domain expertise, and workflow optimisation rather than in the underlying AI models themselves.

Agentic applications represent one clear category, with examples like Cursor, Windsurf, and Lovable demonstrating how LLMs can be equipped with tools, instructions, and knowledge access to execute specific tasks. The core differentiation in these applications lies in UX design—targeting developers in the case of Cursor and Windsurf, or non-technical users with Lovable. These systems provide LLMs with access to file systems, code execution environments, and other tools necessary to perform complex tasks like website creation or code generation, with success heavily dependent on the quality of the interface and workflow design rather than the underlying model capabilities.

Knowledge-based AI startups lean toward RAG architectures, with companies like Glean focusing on enterprise knowledge retrieval and synthesis. However, the distinction between RAG-based and agentic applications is becoming increasingly blurred, with most successful implementations incorporating blended functionalities. Customer care agents and vertical AI solutions for specific domains like legal or tax services exemplify how RAG systems can be enhanced with agentic capabilities to provide more comprehensive and interactive user experiences.

The startup landscape reveals stark realities about sustainability and dependency. While a few startups achieve notable success, the majority struggle with fundamental business model challenges. Even successful companies typically depend heavily on OpenAI or Anthropic, operating at losses while pursuing aggressive growth strategies and battling high churn rates. The focus on growth metrics helps attract venture capital, but this approach creates precarious dependencies on a small number of LLM providers who control pricing and availability.

The fragility of this ecosystem became starkly apparent in incidents like Anthropic restricting Claude access for Windsurf following competitive pressures from OpenAI applications. Such events highlight how even exceptional teams with sophisticated agent-building capabilities remain vulnerable to decisions made by foundation model providers, creating systemic risks that threaten the entire application layer ecosystem.

6. RAG Market Flood

The AI application market has become saturated with providers offering RAG solutions, creating a chaotic landscape where traditional software houses, consultancy firms, and freelancers compete aggressively for small use cases from innovation and data departments. These projects typically target either consumer-facing applications or internal use cases, with providers often employing aggressive sales tactics that oversell the capabilities of what are essentially OpenAI wrappers, showing little consideration for actual client requirements beyond winning the initial Proof of Concept contract.

This market dynamic creates a perverse incentive structure where corporate stakeholders' lack of AI knowledge penalises more transparent software houses that honestly communicate limitations and realistic timelines. The resulting two-tier system sees transparent providers serving fewer but more qualified clients with decent project margins, while less scrupulous competitors chase volume through promises they cannot fulfil. These aggressive providers inevitably deliver PoCs that fail to meet expectations and are never renewed, creating a cycle of disappointment that damages the entire market's credibility.

The situation is exacerbated by SMEs where functional areas maintain unrealistically high expectations coupled with constantly changing requirements, leading to friction and struggles throughout PoC development. This environment has fostered increasing cynicism among software houses, with many choosing to live "PoC by PoC" rather than investing in sustainable product development. The flood of dubious-quality offerings contributes to growing skepticism toward AI applications, particularly after multiple failed PoC experiences leave organisations disillusioned with AI's practical potential.

The actors who successfully monetise this chaotic environment are typically those selling AI education services, gradually building relationships that eventually convert into project opportunities. However, none of these players possess the capabilities required for extremely complex AI applications, limiting their solutions to basic prompt engineering, rudimentary chunking strategies, and minimal agentic functionalities. This technical ceiling means that even successful providers remain fundamentally controlled by the few companies serving LLMs via APIs, who dictate model behaviour, capabilities, and pricing structures.

Among the limited "success cases" in this landscape, customer-facing chatbots represent the most common production deployment. Despite being widely disliked by end users, these systems provide sufficient cost reduction by decreasing customer service headcount to justify their implementation. When properly designed with adequate freshness systems and well-implemented RAG architectures, these applications manage to deliver enough measurable value to survive the transition to production environments.

However, more complex AI applications consistently fail due to the fundamental unpredictability of current systems and the lack of performant small language models that can be reliably deployed in production environments. This technical limitation creates a ceiling effect where only the most basic AI applications can achieve production stability, while more ambitious implementations remain trapped in the PoC phase, unable to deliver the reliability and consistency required for critical business operations.

7. Agentic AI: Ineffective and Expensive

One of the root causes of AI deployment failures in SMEs relates to shifting goals and unrealistic expectations from functional areas, who demand agentic features without explicitly mentioning them in project requirements or understanding their complexity. These stakeholders routinely expect capabilities like web searching, document downloading and parsing, summarisation, database connections, and automated updates—essentially requesting full agentic workflows while treating them as basic features. This expectation mismatch has driven an emerging trend toward Agentic AI solutions marketed as workplace productivity boosters, but the reality proves far more challenging than the marketing suggests.

Agentic AI remains extremely difficult for developers to build effectively. While some developers initially find frameworks like LangChain, LlamaIndex, or LangGraph fascinating, they frequently abandon these tools in favour of direct function calling and custom-built frameworks. The complexity of these established frameworks often outweighs their benefits, leading developers to create simpler, more controllable solutions that sacrifice sophistication for reliability and maintainability.

The fundamental challenge is that Agentic AI implementations are typically limited to PoC environments, with AI engineers consistently avoiding the complexity required for production-scale deployment. Setting appropriate guardrails for agents proves exceptionally complex, while building reliable tools that agents can consistently use remains a significant technical challenge. The unpredictable nature of agent behaviour makes it difficult to ensure consistent, business-appropriate outcomes across different scenarios and use cases.

To address these deployment difficulties, specialised development tools like Mistral AI Agents API and Panda Agentic Intelligence SDK have emerged as solutions designed to eliminate boilerplate code and deployment complexity. These tools attempt to provide either vertical agents via API or general AI agent customisation capabilities, focusing on removing DevOps barriers through features such as WebSocket communication for real-time bidirectional connections, pluggable tool handlers that provide different capabilities, isolated environment integration for secure operations, and event streaming for real-time transmission of agent thoughts and actions.

Despite these technical advances, agentic solutions face fundamental economic challenges when deployed with cloud-based LLM providers. The cost per interaction becomes prohibitively expensive for most organisations, while the unpredictable nature of agent behaviour makes it difficult to justify costs against measurable business outcomes. The initial "wow effect" of sophisticated AI agents quickly vanishes when organizations calculate the actual cost of each generated dashboard or completed workflow, revealing an unsustainable economic model for most use cases.

Simpler "agents"—more accurately described as automation workflows that incorporate LLM capabilities—appear to offer a more practical path toward productivity improvements for internal teams. These solutions typically function as internal tools rather than production-ready projects, focusing on specific, well-defined tasks where LLM capabilities can provide clear value without the complexity and cost of full agentic systems.

The fundamental issue of shifting goals and failing projects cannot be solved simply by deploying more advanced AI agents. Effective and cost-efficient agentic AI requires two critical components that most organisations lack: small language models owned and controlled by the organisation itself, and clear, stable business requirements that allow specialised models to outperform general-purpose large language models on specific tasks. Without these foundational elements, agentic AI remains an expensive experiment rather than a practical business solution, perpetuating the cycle of failed AI implementations that characterises the current market landscape.

8. Memory and Knowledge Layer Struggles

One of the key challenges in deploying AI solutions for SMEs lies in the struggle with the memory and knowledge layers of the system. These two dimensions—knowledge, representing static domain information such as documents and structured resources, and memory, referring to dynamic inputs provided by users over time—require careful coordination and ongoing maintenance. In practice, both layers tend to expose structural weaknesses in SME implementations, often due to the fragmented and inconsistent nature of available data and the operational overhead of keeping systems up to date.

On the knowledge side, AI applications frequently depend on ingesting large volumes of unstructured, heterogeneous documents. These documents often include scanned PDFs, handwritten forms, or exported files with inconsistent layouts and formatting. The success of downstream AI tasks depends heavily on the quality of document parsing and structuring. There is no universal parser that fits all cases; instead, a variety of parsing technologies are typically tested and applied based on document type. A common approach involves organising documents into type-specific folders and applying different parsers accordingly. In this multi-parser setup, the aim is usually to convert the inputs into usable plain text or markdown.

However, given the variability in document structure and the presence of long, irrelevant sections, LLMs are increasingly being used to extract and synthesise content directly. While these models introduce the risk of hallucinations, they also offer significant advantages in terms of summarisation and the structuring of loosely formatted information. In some cases, additional mapping is required across document sets, necessitating the use of third-party data or custom logic to link related data points. Teams that invest early in robust document analysis processes and strong data pipelines typically deliver higher-quality proofs of concept, yet many SMEs struggle to maintain these pipelines over time. Changes in document formats or lapses in data governance often lead to broken parsers and degraded performance.

The memory layer introduces a different set of difficulties. Capturing and managing the evolving context of user inputs remains a complex task. Some open-source tools, such as Zep Memory, as well as emerging graph-based systems, attempt to store and update user-provided inputs over time. These systems aim to create a persistent, retrievable structure of user knowledge that can support contextual responses. However, triggering memory in a precise and timely manner, especially in production environments, remains elusive.

In many cases, production systems revert to simpler, rule-based memory updates, preserving only a few critical user attributes and retrieving them under predefined conditions. This limits the potential of memory-driven personalisation but remains one of the few reliable strategies currently viable for SMEs, where infrastructure constraints and production stability often outweigh experimental complexity.

Overall, the interplay between knowledge structuring and memory persistence remains a defining challenge in the development of effective AI systems for SMEs. Success often depends not only on model performance but also on the quality, stability, and adaptability of the underlying data and memory infrastructure.

9. Freshness Layer Fragility

A further critical yet often underestimated challenge in AI implementations for SMEs is the fragility of the freshness layer. Freshness, in this context, refers to the ability of the system to keep its knowledge base updated with current, accurate information while systematically removing outdated or obsolete content. This requirement becomes particularly pressing in domains where business processes, documentation, or product information evolve rapidly. Despite its importance, maintaining freshness is a fragile operation that is frequently undermined by the structural limitations of typical SME data workflows.

Updating the knowledge layer is rarely a matter of simply appending new information. In most cases, new data either modifies or supersedes existing entries, and failing to reconcile these changes can lead to contradictions, duplicated facts, or the persistence of outdated insights. This undermines trust in the system's outputs and creates cognitive dissonance for end users. Yet, the mechanisms required to ensure freshness—automated version control, deduplication logic, change detection, and relevance filtering—are rarely in place in SME environments. As a result, knowledge bases tend to accumulate noise over time, reducing their effectiveness and clarity.

Even when updates are manually or periodically integrated, the absence of a rigorous strategy for identifying and pruning obsolete data results in drift between the AI system's outputs and the current state of the business. For example, if an AI assistant continues referencing a deprecated policy or an outdated process flow, its responses become not only unhelpful but potentially damaging. Moreover, the brittleness of document parsers and extractors often compounds the problem. When updated documents are ingested, changes in formatting or structure can silently cause extraction failures, leading to partial or missing updates without immediate visibility into the issue.

Efforts to automate freshness checks are often limited by both technical and operational constraints. LLMs can assist in surfacing inconsistencies or flagging outdated references, but their reliability varies, and they introduce new risks related to hallucination or false positives. Consequently, teams must often rely on brittle rule-based systems or manual oversight, both of which scale poorly. In practice, only organisations with a deliberate commitment to maintaining freshness—through change detection pipelines, content lifecycle management, and version-aware ingestion strategies—achieve consistently accurate and relevant knowledge layers.

For SMEs, this level of infrastructure maturity is often aspirational. The result is a fragility that affects not just the accuracy of outputs but also the long-term viability of the AI solution itself. Without freshness, even well-designed systems degrade over time, as the underlying knowledge diverges from the evolving reality they are meant to support.

10. AI Guardrails Struggle

One of the most persistent obstacles in building AI systems for SMEs is the inadequacy of current guardrail mechanisms—tools and methods designed to prevent undesired or dangerous outputs from AI models. While the importance of guardrails is widely recognised, especially for production environments, the reality is that most existing solutions are fragile, superficial, or operationally impractical. Their failure often stems from the inherent unpredictability of large, general-purpose language models, which remain difficult to constrain despite advances in prompting, filtering, and behavioural control.

A common guardrail technique is prompt conditioning—embedding rules, roles, and behavioural constraints into the prompt itself. While cheap and flexible, prompt engineering remains unreliable: models frequently ignore instructions, especially when instructions conflict or become long and complex. Another widespread method is content filtering, using external moderation APIs (e.g. OpenAI's Moderation API, Google's Perspective API) to detect and block outputs that are toxic, biased, or off-policy. These tools are reactive, applied after generation, and thus cannot prevent the model from internally reasoning with or propagating unsafe logic. They are also prone to both false positives and false negatives, and often lack context awareness.

More advanced solutions like Guardrails AI, TruLens, or Rebuff attempt to define behavioural expectations using schemas, assertions, and post-processing rules. These frameworks offer more structured oversight, but they still operate around black-box models, lacking deep integration into the generation process. They are also difficult to generalise, often requiring extensive configuration and maintenance for each new task. Function calling and tool-use APIs offer some determinism by outsourcing specific tasks to external systems, but they introduce complexity and fail to guard the parts of the model that still perform open-ended reasoning or content generation.

Even Retrieval-Augmented Generation (RAG) systems, which ground outputs in curated external documents, cannot fully prevent hallucinations or undesired reasoning. The model can still misinterpret or invent information based on retrieved content, and RAG pipelines add substantial infrastructure and latency that SMEs often cannot support. Confidence estimation and fallback strategies, another common pattern, also suffer from a core limitation: the model itself is being asked to judge whether it should be trusted—an unreliable self-referential process.

Ultimately, these mechanisms try to contain something that is not inherently designed to be contained. General-purpose language models are probabilistic, open-ended systems trained for fluency, not for correctness or constraint. They will always exhibit edge-case behaviours, produce unexpected outputs, and defy rigid boundaries, especially when given diverse, real-world inputs. No combination of filters, prompts, or post-hoc rules can guarantee safety or determinism at scale.

In contrast, small, task-specific models offer a fundamentally different trade-off. By reducing generalisation and focusing on narrowly scoped tasks—such as classification, extraction, summarisation, or domain-specific question answering—these models become more reliable, interpretable, and controllable. They are easier to test and validate, exhibit less emergent behaviour, and can be directly aligned with ground truth data. When fine-tuned or trained from scratch on domain-relevant examples, these models can achieve high performance within their intended scope without requiring complex guardrails. Their smaller size also allows for faster inference, easier deployment on local infrastructure, and greater transparency in failure analysis.

However, despite the clear advantages, neither SMEs nor the consultancy firms that support them are currently positioned to develop such models independently. Most SMEs lack the internal machine learning expertise, model training infrastructure, or strategic clarity to invest in building their own models. Even consultancy teams, though often skilled in system integration and rapid prototyping, rarely have the deep machine learning know-how or access to specialised talent required to curate datasets, select architectures, fine-tune effectively, and monitor model drift over time.

Furthermore, many projects suffer from unclear or overly broad use cases, making it difficult to scope a task narrowly enough for a small model to be viable. The result is a continued over-reliance on general-purpose models paired with fragile guardrails—despite a growing recognition that smaller, purpose-built models would offer greater reliability and cost-effectiveness if the right capabilities were in place.

While tooling and open-weight models are improving access, the full transition to small-model deployments remains stalled by the combined limitations in domain-specific expertise, MLOps maturity, and product clarity—a gap that remains one of the biggest structural challenges for deploying robust AI in SME contexts.

11. AI and Data Analysis

The promise of conversational data analysis and self-service BI tools has captured the imagination of SMEs seeking to democratise data insights across their organisations. However, the reality of implementing AI-driven data analysis reveals significant complexity, particularly when dealing with substantial data volumes and the messy, incomplete datasets that characterise most real-world business environments.

Organisations with existing data platforms, including those using Databricks, frequently express dissatisfaction with commercial solutions like Genie AI/BI, driving them to seek alternatives in the open-source ecosystem. This search intensifies as the number of tables and query complexity increases, revealing fundamental limitations in current AI data analysis approaches. The challenge becomes particularly acute when dealing with the non-gold-standard data quality typical of SME environments—datasets plagued by missing parameters, inconsistent formatting, and incomplete records that confound even sophisticated AI systems.

End users compound these technical challenges by asking questions that far exceed the capabilities of normal SQL queries, expecting AI systems to perform complex analytical reasoning that simple LLM integration cannot deliver. The naive approach of "plugging in an LLM" consistently fails, leading to hallucinations and unreliable results that undermine user confidence in AI-driven analytics.

The most promising approaches emerging from successful implementations involve a two-pronged strategy. First, organisations develop simple but comprehensive descriptions of each table, including data glossaries that explain table headers and their business context to the LLM. Second, they invest in semantic layers that create multiple curated views of their data tables, pre-processing complex relationships and business logic into more manageable structures. This approach shifts the AI's role from inferring complex data relationships to selecting appropriate pre-built views and generating targeted queries against those views.

Three distinct trends have emerged in AI data analysis applications: text-to-SQL generation for accelerating query development, self-service BI for exploratory data analysis, and the ambitious goal of enabling natural language requests for linear regression and predictive analysis. However, these experiments consistently struggle to reach production environments, primarily because they require fundamental changes to existing BI infrastructure that organizations are reluctant to implement.

The typical enterprise request focuses on integrating chatbot-like functionality within established tools like PowerBI or Tableau, rather than replacing existing systems. This integration requirement creates additional complexity, as AI components must work seamlessly within legacy architectures while maintaining familiar user interfaces and workflows.

The substantial effort required to prepare semantic data layers for successful AI data analysis represents a significant barrier for SMEs. Unlike large enterprises with dedicated data engineering teams, smaller organisations often lack the resources and expertise needed to properly structure their data for AI consumption. This preparation work—involving data cleaning, relationship mapping, and semantic layer construction—requires significant upfront investment that many SMEs cannot justify given uncertain returns.

The result is a persistent gap between the theoretical capabilities of AI data analysis and the practical realities of implementation in resource-constrained environments, where the combination of poor data quality, complex integration requirements, and limited technical expertise continues to prevent successful production deployments.

12. AI and Data Engineering

Data engineers represent one of the most skeptical professional groups regarding AI integration, primarily due to their deep-seated need for deterministic, reliable data pipelines. The probabilistic nature of LLMs fundamentally conflicts with the precision and predictability that data engineering demands, where even minor inconsistencies can cascade through entire analytical ecosystems and compromise business-critical reporting.

Despite this skepticism, a growing segment of data professionals has begun cautiously adopting LLMs for specific, non-production tasks that leverage AI's pattern recognition capabilities while maintaining human oversight. The most common application involves using AI to accelerate data glossary creation, particularly when confronting unfamiliar datasets or legacy systems lacking proper documentation. In these scenarios, LLMs analyse data structures, field names, and sample values to generate preliminary descriptions of what the data represents, creating initial glossaries that data engineers then validate and refine.

This approach proves particularly valuable during data discovery phases, where engineers must quickly understand large volumes of poorly documented data. Rather than manually examining hundreds of tables and thousands of fields, teams can use AI to generate initial hypotheses about data meaning and relationships, dramatically reducing the time required for initial data assessment. However, these AI-generated descriptions are invariably treated as starting points rather than definitive documentation, with experienced engineers maintaining strict validation protocols.

A more cautious but emerging trend involves using LLMs to generate boilerplate code for data infrastructure components, particularly dbt models, YAML schemas, and semantic layer definitions. Data engineers leverage AI to create initial templates for common data transformations, aggregation patterns, and view definitions based on existing data structures. This application addresses one of the most time-consuming aspects of data engineering: writing repetitive configuration code that follows predictable patterns.

The generation of semantic layers represents the most sophisticated application, where LLMs analyse existing data schemas to suggest potential aggregation strategies, dimensional models, and analytical views. AI can identify common business metrics, propose fact and dimension table structures, and generate initial semantic layer configurations that data engineers can then customise for specific organisational needs.

However, the deterministic requirements of production data pipelines create a clear boundary for AI involvement. Data engineers consistently avoid deploying AI-generated code directly into production environments, instead using these tools exclusively for initial development phases, exploratory work, and boilerplate generation. The critical nature of data pipeline reliability means that any code touching production systems undergoes rigorous human review and testing, regardless of its origin.

This cautious approach reflects a mature understanding of AI limitations within the data engineering context. While LLMs excel at recognising patterns and generating initial frameworks, they cannot guarantee the consistency, performance, and error handling requirements essential for production data systems. The result is a pragmatic adoption pattern where AI serves as an intelligent assistant for development tasks rather than an automated replacement for engineering judgment.

The integration of AI in data engineering workflows thus represents a measured evolution rather than a revolutionary change, with professionals leveraging AI's strengths while maintaining the deterministic control essential for reliable data infrastructure.

13. Evaluation Avoidance

The combination of limited budgets, high expectations, and complex input requirements creates a systematic avoidance of formal evaluation processes in SME AI projects. Most organisations skip structured evaluation entirely, defaulting to end-user "gut feeling" assessments during ad-hoc testing phases.

This evaluation gap stems from several interconnected challenges. Organisations typically demand highly customised solutions tailored to their specific domain, making it difficult to establish standardised evaluation metrics or golden datasets for comparison. Even when technical teams attempt to implement automated evaluation frameworks, functional areas and end users consistently prefer manual testing, wanting to personally examine outputs and validate results against their domain expertise.

The absence of dedicated evaluation budgets exacerbates this issue. SMEs often view evaluation as an optional overhead rather than a critical component of AI system development. Consequently, improvements are driven by specific user complaints and iterative adjustments rather than systematic performance measurement. While development teams may privately utilise monitoring tools like LangFuse and implement basic evaluation metrics, the official assessment process remains dominated by subjective manual reviews from business stakeholders.

When technical teams do implement evaluation frameworks, they face a fragmented landscape of tools and libraries with no clear winner emerging. Options include DeepEval, RAGAS for RAG system evaluation, Evidently AI for model monitoring, LangSmith for tracing alongside LangFuse for observability, Phoenix Arize for evaluation workflows, MLflow with custom LLM metrics, Weights & Biases for performance monitoring, and HELM for standardised benchmarking. +

However, many teams find that evaluating this knowledge layer—the retrieval and structuring of information before it reaches the LLM—requires extensive manual parsing and validation, often proving more important than assessing the LLM response itself. This knowledge layer evaluation involves checking document relevance, chunk quality, metadata accuracy, and retrieval precision, tasks that resist automation and demand domain expertise. The complexity of validating these upstream components often overshadows LLM output evaluation, as incorrect or irrelevant knowledge retrieval can render even perfect language generation useless for business applications.

For such reasons, automatic and scientific evaluation of LLM responses is avoided, and evaluation avoidance creates a feedback loop where AI systems are optimized for immediate user satisfaction rather than measurable performance criteria, contributing to the broader challenges of moving AI initiatives from proof-of-concept to reliable production systems.

14. Evaluation-Driven Success

In stark contrast to the prevalent evaluation avoidance that characterises most AI implementations, there exist notable exceptions where enterprises and their consulting partners embrace comprehensive assessment methodologies as the cornerstone of successful deployment. These rare collaborative engagements actively pursue rigorous evaluation frameworks, viewing systematic testing not as an obstacle to rapid deployment but as the essential foundation for sustainable AI integration. Rather than optimising for immediate user satisfaction at the expense of measurable performance criteria, these partnerships prioritise long-term system reliability through structured validation processes that bridge the gap between proof-of-concept demonstrations and production-ready solutions.

The distinguishing factor in these successful partnerships lies in the establishment of evaluation-first methodologies where both consulting teams and enterprise stakeholders commit to systematic performance measurement from project inception. This approach fundamentally reframes the relationship between technical implementation and business requirements, moving away from technology-driven solutions toward evidence-based development processes. Key performance indicators emerge through collaborative dialogue, encompassing not only traditional user satisfaction metrics such as Net Promoter Score (NPS) and Customer Satisfaction Score (CSAT), but also operational metrics including task completion rates, error reduction percentages, user adoption velocity, and system reliability scores.

Requirements alignment becomes a data-driven process in these evaluation-centric engagements, where both parties invest significant effort in establishing measurable success criteria before technical development begins. This iterative refinement process prevents the common scenario where sophisticated AI solutions fail to address actual business needs by ensuring that every system capability directly corresponds to validated performance requirements. Teams establish clear boundaries for AI decision-making authority based on empirical testing results rather than theoretical capabilities, creating sustainable frameworks for human-AI collaboration.

Perhaps most critically, these partnerships view comprehensive evaluation as a competitive advantage rather than a compliance burden. Multi-phase evaluation protocols include controlled pilot testing, systematic user feedback integration, and rigorous performance benchmarking against established baselines. This evaluation-centric approach allows teams to identify and optimise system performance before production deployment, while also generating concrete evidence of business value that supports broader organisational adoption. The commitment to thorough evaluation processes creates a positive feedback loop where continuous measurement drives iterative improvement, ultimately producing AI systems that demonstrate measurable impact on organisational objectives.

The resulting implementations prove that when both consultancy firms and client organisations prioritise evaluation rigor over deployment speed, AI initiatives achieve sustained performance levels and maintain reliable operation in production environments. These collaborative approaches demonstrate that successful AI deployment depends fundamentally on the willingness to invest in systematic assessment methodologies, where evidence-based decision-making creates the foundation for scalable, measurable business value.

15. Open Models Accuracy Concerns

Business experts and functional areas consistently dismiss benchmark performance metrics when evaluating AI models, preferring manual testing based on their specific use cases and domain requirements. This subjective evaluation approach systematically favours commercial API-based models over open-weight alternatives, creating a perception that proprietary solutions deliver superior accuracy and reliability.

Open-weight models have increasingly become positioned as a budget-conscious choice rather than a strategic decision for technical sovereignty. This shift reflects a fundamental gap in expertise rather than inherent model limitations. While some open-weight models can deliver cost-effective solutions with faster inference times compared to commercial alternatives, realising these benefits requires substantial fine-tuning expertise that most SMEs lack.

The technical barrier to effective model customisation has created a self-reinforcing cycle. Average developers often lack the specialised knowledge needed to properly fine-tune or distill open-weight models for specific business applications. This knowledge gap is particularly pronounced in areas like prompt engineering, parameter-efficient fine-tuning techniques, and domain adaptation strategies.

Consulting firms, operating within tight budget constraints, typically avoid investing in the extensive fine-tuning work required to optimise open-weight models. The upfront effort and technical risk associated with model customisation often exceed the allocated project budgets, making plug-and-play commercial APIs the safer choice for meeting client expectations within defined timelines.

This dynamic perpetuates the misconception that open-weight models are inherently inferior, when in reality they often require more sophisticated implementation approaches that SMEs and their consulting partners are currently ill-equipped to execute effectively.

16. LLM APIs Dependency and Cost Frustration

SMEs and AI startups find themselves trapped in an uncomfortable dependency on commercial LLM APIs, where initial proof-of-concept costs appear manageable but production-scale expenses quickly become prohibitive. This creates a persistent tension between technical feasibility and economic sustainability that undermines long-term AI adoption strategies.

The cost structure of API-based models presents a fundamental mismatch with both SME business models and startup economics. While development teams can demonstrate impressive capabilities during limited testing phases, scaling to handle real-world usage volumes often reveals monthly costs that exceed entire IT budgets. Organisations frequently discover that processing their actual document volumes or supporting their full user base would require API expenses ranging from €5,000 to €50,000 monthly—figures that dwarf the original project budgets.

AI startups face a particularly acute version of this challenge, often losing money on individual customers due to high API costs while pursuing aggressive growth strategies. Many startups knowingly operate at negative unit economics, prioritising revenue growth and user acquisition while hoping to address the fundamental dependency on expensive LLM APIs later through optimisation, volume discounts, or eventual migration to self-hosted solutions. This "grow now, optimise later" approach creates significant financial risk and forces startups into a race against time to achieve sufficient scale before running out of funding.

This economic reality forces SMEs and startups into uncomfortable compromises. Many resort to artificial usage restrictions, limiting the number of queries per user or reducing the scope of AI-enabled features to control costs. Others implement complex caching strategies or pre-processing workflows to minimise API calls, adding technical complexity that partially negates the simplicity advantage of cloud-based solutions.

The unpredictability of API costs compounds these challenges for both SMEs and startups. Unlike traditional software licensing with fixed annual fees, LLM API expenses fluctuate based on usage patterns, model updates, and provider pricing changes. This variability makes budget planning extremely difficult for organisations operating on tight financial margins, creating resistance from financial stakeholders who prefer predictable technology investments.

Privacy and regulatory compliance add another layer of complexity to API dependencies for both SMEs and startups. Organisations increasingly face pressure to comply with regulations like GDPR, the EU AI Act, and sector-specific requirements such as DORA (Digital Operational Resilience Act) for financial services. Sending sensitive business data to external API providers raises significant privacy concerns and potential regulatory violations, particularly when data crosses jurisdictional boundaries or lacks adequate processing transparency.

Many organisations discover that their chosen API providers cannot guarantee data residency within EU borders or provide sufficient audit trails for regulatory compliance. This creates a paradox where the simplest technical solution becomes the most legally risky option, forcing SMEs and startups to choose between operational efficiency and regulatory safety.

Vendor lock-in concerns further exacerbate the dependency issue. Organisations worry about sudden price increases, service discontinuation, or changes in model performance that could disrupt their operations. API providers maintain complete control over model behaviour, can discontinue services unilaterally, and often implement changes that affect system performance without customer consent. Recent court decisions, including rulings requiring OpenAI to retain even deleted chat conversations, have highlighted how API providers may store and retain user data regardless of deletion requests, creating additional privacy and compliance risks for SMEs handling sensitive business information.

However, the technical complexity of migrating between providers or transitioning to self-hosted alternatives often feels insurmountable given current internal capabilities.

The result is a fragile ecosystem where AI initiatives remain perpetually vulnerable to external cost pressures and vendor decisions, preventing both SMEs and startups from building confident, long-term AI strategies and contributing to the high failure rate of production deployments.

17. The Training and Fine-Tuning Taboos

Fine-tuning and training custom models have become virtually taboo subjects in SME AI projects, treated with the same apprehension typically reserved for experimental or high-risk technologies. This aversion stems from a combination of technical intimidation, resource constraints, and deeply ingrained misconceptions about the complexity and unpredictability of model customisation.

The perception of fine-tuning as an arcane, unpredictable process dominates conversations with business stakeholders and even technical teams. Many developers view model training as a black box requiring PhD-level expertise, despite the availability of user-friendly tools and frameworks that have significantly lowered the technical barriers. This psychological barrier is reinforced by consultant firms who, operating under tight budgets and delivery timelines, prefer the perceived safety of pre-trained models over the uncertainty of customisation efforts.

Resource constraints compound these perceptions. SMEs often lack the computational infrastructure traditionally associated with model training, and cloud-based training costs appear prohibitive when compared to simple API calls. The upfront investment required for data preparation, training infrastructure, and experimentation feels overwhelming, particularly when weighed against the apparent simplicity of plug-and-play commercial solutions.

The unpredictability stigma surrounding fine-tuning reflects a fundamental misunderstanding of modern techniques. Parameter-efficient methods like LoRA (Low-Rank Adaptation), QLoRA, and adapter layers have dramatically reduced both the computational requirements and the technical complexity of model customisation. However, awareness of these advances remains limited among SME development teams, who continue to associate fine-tuning with the resource-intensive full-parameter training approaches of earlier years.

This knowledge gap creates a self-perpetuating cycle where the lack of fine-tuning expertise prevents organisations from developing the domain-specific models that could solve their accuracy and cost challenges. Instead, SMEs remain dependent on generic commercial models that may be poorly suited to their specific use cases, languages, or business contexts.

The taboo extends to training smaller, specialised models that could deliver superior performance for specific tasks while reducing operational costs. Domain-specific models trained on curated datasets often outperform larger general-purpose models for particular applications, while consuming significantly fewer computational resources. However, the expertise required to design appropriate training strategies, select optimal model architectures, and implement effective evaluation frameworks remains largely absent from the SME ecosystem.

Consulting firms actively avoid proposing fine-tuning solutions, citing the additional complexity and time investment required. The preference for delivering quick, demonstrable results within constrained budgets makes fine-tuning appear risky and uneconomical, even when customised models could provide better long-term value and reduced dependency on external API providers.

This systematic avoidance of model customisation represents a critical missed opportunity for SMEs to develop AI solutions that are truly aligned with their business needs, cost structures, and technical sovereignty requirements.

18. Reliability & Programmable AI

In the SMEs landscape, business leaders have long operated with systems that offer predictable outcomes and deterministic behaviours. Traditional software follows explicit instructions, producing consistent results when given identical inputs. This reliability has been the cornerstone of business operations for decades. However, the emergence of Large Language Models (LLMs) presents a paradigm shift that challenges this fundamental expectation. While LLMs offer unprecedented capabilities in natural language understanding and generation, their probabilistic nature introduces a tension between innovation and reliability that organisations must navigate.

The power of LLMs stems from their ability to generalise across diverse contexts, drawing from vast training data to handle unforeseen scenarios. Yet this same characteristic—their statistical approach to language—introduces variability in outputs that can be problematic in business-critical applications. A financial institution cannot afford hallucinations in regulatory compliance reporting; a healthcare provider cannot risk inconsistent medical recommendations. The unpredictability that makes LLMs versatile simultaneously renders them challenging to deploy in environments where errors carry significant consequences.

The industry has responded to this challenge through the development of AI agents—LLMs augmented with tools that constrain their actions within more deterministic boundaries. By channelling model outputs through well-defined tools with predictable behaviours, organisations can mitigate some of the inherent variability. This approach represents a compromise between leveraging the generative capabilities of LLMs while imposing guardrails that increase reliability. The tool-calling paradigm effectively narrows the action space of the model, reducing the surface area for potential errors while maintaining flexibility for complex reasoning.

Interestingly, there are indications that commercial LLM providers are pursuing another strategy to enhance reliability: intentional dataset curation that guides models toward consistent behaviours in specific domains. Unlike traditional machine learning where overfitting is avoided, these providers appear to be selectively "overtuning" models to produce more deterministic responses for high-value business scenarios. This represents a departure from general-purpose optimisation, instead prioritising predictability in exchange for some degree of generalisability.

This approach suggests a promising direction for developing more reliable AI systems: purpose-built models that sacrifice some generality for greater determinism within their intended domains. Smaller, specialised models—trained on carefully curated datasets that emphasise consistent behaviours for specific business functions—could offer several advantages: reduced computational requirements, faster inference times, lower operational costs, and most importantly, more predictable performance within their designated scope. A model designed exclusively for contract analysis, for instance, might lack the versatility of a general-purpose LLM but could deliver superior reliability for its intended purpose.

The challenge, however, lies in democratising the expertise required to develop such specialised models. Currently, the knowledge and resources needed to effectively balance generalisability with determinism remain concentrated among a handful of technology giants. Smaller organisations lack access to the training methodologies, computational resources, and domain expertise necessary to create reliable, task-specific AI systems. Without broader dissemination of these capabilities, we risk creating a two-tiered AI landscape where only the largest enterprises can deploy truly reliable AI solutions.

Moving toward a future of more programmable AI thus requires not only technical innovation but also knowledge sharing across the industry. Open research initiatives, collaborative development frameworks, and accessible tools for model specialisation will be essential to ensure that organisations of all sizes can benefit from more deterministic AI systems. The path forward involves recognising that absolute determinism may remain elusive for generative AI, but through thoughtful design constraints and purpose-built specialisation, we can develop systems that strike an appropriate balance between flexibility and reliability for business applications.

Conclusions and Future Work

The research presented in this report has exposed the profound structural and operational barriers that prevent AI projects in SMEs from reaching production. From externalised consultancy-driven implementations to a culture of evaluation avoidance and fine-tuning taboos, the ecosystem remains fragile, fragmented, and overdependent on commercial LLM APIs. A combination of technical gaps, budget constraints, and inflated expectations from non-technical stakeholders has led to a landscape where even the most promising AI applications rarely achieve long-term sustainability or business impact.

Yet, beneath these challenges lies a path forward—one that involves redefining what reliable, scalable AI looks like for SMEs. A key insight emerging from this study is that the future of SME-friendly AI is not necessarily bigger or more general, but rather smaller, more deterministic, and domain-specific. Reliability, cost-efficiency, and explainability demand a shift from probabilistic general-purpose models toward programmable AI systems tailored for well-scoped business functions. These models, if built correctly, offer consistent outputs, better integration with traditional software paradigms, and dramatically lower operational costs.

Crucially, the next phase of research and development must focus on enabling SMEs not just to consume AI but to own it. This means giving them the knowledge, tools, and infrastructure to train, fine-tune, and deploy dedicated small language models that are aligned with their data, workflows, and compliance needs. Democratising ownership will require open-source frameworks, user-friendly developer tools, and community-driven methodologies that bridge the gap between general AI capabilities and specific business requirements. Only by making programmable AI truly accessible can we break the cycle of dependency and enable SMEs to build AI systems they can understand, trust, and evolve over time.

Future work will therefore concentrate on three foundational areas: (1) simplifying the tooling and DevOps required to host and iterate on small models, (2) designing fine-tuning workflows that are accessible to software teams without deep ML backgrounds, and (3) establishing best practices and reference patterns for model evaluation, governance, and domain alignment. By anchoring innovation in ownership and control, we believe it is possible to move beyond PoCs and unlock a new generation of reliable, SME-owned AI applications.

About PandasAI
Open Discovery Initiatives

PandasAI (YC W24) is focused on building developer tools for real-world AI applications. Its *Open Discovery Initiatives* are collaborative product discovery and open research projects that gather insights from AI engineers and practitioners to guide problem discovery and openly share findings with the whole AI community. Aligned with the goal of strengthening EU technological sovereignty, these initiatives aim to reduce dependence on foreign AI providers by empowering local builders. To contribute, check for open initiatives on our website (https://pandas-ai.com) or email pm@sinaptik.ai.