Deloitte.

Together makes progress

SFL Scientific

a Deloitte business

Five lessons learned from building GenAI and agentic AI solutions

Introduction

The advent of Generative AI (GenAI) and, more recently, agentic AI has significantly transformed the future business landscape. These advancements have ushered in a proliferation of models, platforms, and vendors that substantially reduce the technical barriers of entry. Features that once took weeks and months to develop can now be showcased within *minutes*.

However, while prototyping has accelerated, overall timelines for deploying fully functional solutions remain largely unchanged. The difference lies in where time is spent—moving from upfront data collection and planning to post-hoc validation, guardrail setup, change management, and traditional infrastructure scaling.

Furthermore, GenAI and agentic AI solutions excel in areas where robotic process automation (RPA) and more traditional machine learning have struggled—namely, with large amounts of *unlabeled and unstructured* text and images. With the potential to unlock vast business value, GenAI and agentic AI are not without unique challenges that we have the opportunity to address.

Implications:

1. Corporate strategy must evolve. Organizations must rethink their approach to drive efficiency, growth, and innovation. To leverage GenAI's potential, businesses must reshape their processes and strategies to align with this new paradigm.
2. GenAI and agentic solutions require new approaches. The development, deployment, and scaling stages of these solutions present unique challenges. Successfully integrating GenAI into business solutions requires a nuanced understanding of its strengths, including task-specific optimizations and engineering best practices.

This white paper primarily focuses on the latter implication of GenAI and agentic AI integration. In what follows, we outline five key lessons our team has learned from developing and deploying scalable GenAI and agentic AI solutions into business verticals. As GenAI is a principal component of agentic AI, we will henceforth refer to both simply as agentic AI for brevity.

1. Agentic AI solutions should start small

Agentic AI has such far-reaching and generalizable applications that it may be challenging to focus the technology into business-specific solutions that generate value. Scope creep and misaligned expectations can derail business strategies. It is therefore essential to start with narrowly focused solutions that can be iteratively refined for increased business value generation over time.

Key considerations:

- **Define explicit use cases and intended users early.** Establish well-scoped use cases early to help maintain focus, prevent scope creep, and ensure business alignment.
- **Educate stakeholders.** Define and manage realistic expectations for the tool's application, anticipated value, and limitations of the technology.
- **Prioritize data quality.** Accurate and useful outputs depend on well-managed, quality data. Prepare to allocate the necessary resources to data curation.
- **Fast feature generation does not necessarily reduce production time.** The ability to quickly showcase new features does not directly correlate to reduced time to production, scalability, or value generation.

Strategy | Development | Change management

2. Evaluating agentic AI systems is open-ended and evolving

Unlike traditional AI, where solutions can be validated against labeled data gathered before model development, agentic AI solutions generally rely on post hoc human feedback in development. However, the evaluation is particularly challenging due to the unstructured and inherent novelty of outputs (i.e., generated text or images that never previously existed). These systems won't discover your goals on their own; Define "what good looks like" up front and wire it into feedback loops, otherwise agentic systems optimize the wrong thing and adoption stalls.

Key considerations:

- **Define business value key performance indicators upfront.** Establish performance benchmarks early, even if they are approximate.
- **Establish technical evaluation metrics.** Define clear criteria for technical accuracy, relevance, and comprehensiveness. These should be aligned with business requirements to help ensure meaningful and reliable assessments. Model metrics may involve rubric-based scoring, human evaluation, or automated benchmarking tools designed for unstructured and non-deterministic outputs.
- **Simulations offer controlled environments for evaluation.** Simulated users and virtual environments can test agentic AI solutions with accepted benchmarks defined through collaboration with subject matter experts.
- **Evaluate entire task trajectories.** For agentic AI, entire trajectories of sequential outputs can be evaluated against expected trajectories. This includes evaluating the sequence of tools that the agentic AI chooses to call and the intermediate natural language responses that guide the user. Tying these trajectory choices to business requirements is a nuanced objective that needs continued refinement.
- **Implement structured feedback loops.** Engage in user feedback sessions and distribute user interface (UI)-embedded interactive surveys or rating systems to capture real-time input. Additional feedback frameworks can be semi-automated using large language models (LLMs) to provide "user" feedback, which is also highly useful for large-scale evaluations.
- **Iteratively review and refine.** Implement a set cadence for deployment, testing, and improvement cycles. Performance improves through iterative testing cycles (rather than one-time validation on static held-out data), error analysis, and systematic refinements such as prompt tuning, system architecture adjustments, and user feedback integration.

Strategy | Development | Change management

3. User experience (UX) and change management (CM) are hugely important

User expectations for GenAI and agentic solutions have risen sharply, and teams are judged on delivering a complete, reliable product rather than a showcase of capabilities. Successful agentic AI deployment goes beyond technical feasibility—it also demands strategic user experience and adoption planning. Due to the novelty of agentic AI solutions, users do not have many anchor intuitions on how applications should function and are often more resistant to change. In many cases, solutions designed to integrate smoothly into end-user workflows, even with significantly lower technical accuracy, tend to have higher adoption and better utility. This makes UX and CM even more critical for businesses in generating value when using agentic AI.

Key considerations:

- **Implement agentic AI-specific UX improvements:**
  - Design intuitive handling of failure modes (e.g., providing retry options for failed queries).
  - Offer transparent support systems (e.g., pre-generated prompts and guidance on improving query formulation).
  - Implement response-ranking for improved clarity and precision.
  - Enable human-in-the-loop workflows and scaffolding for critical tasks (e.g., hardcoding rulesets or prompts that users can remove/adjust over time).
  - Integrate in-app documentation, active suggestions during usage, and tutorial features for users to improve over time.
- **Integrate CM into the development process.** In addition to standard change management best practices, such as establishing change networks, providing ongoing training, and implementing roll-out strategies, utilize necessary development and validation cycles to educate, gather feedback, and create engagement and buy-in with the solution. This will also assist in refining the business case and ensuring stakeholder alignment.

Strategy | Development | Change management

4. Shipping AI solutions means tackling uncertainty and hallucinations

Because LLMs are probabilistic, next-token predictors, the appearance of fluency can outrun reasoning or correctness. The nature of this technology as the underpinning of GenAI and Agent solutions underscores the need to account for hallucinations and errors in a comprehensive solution build. To make answers dependable, all new solutions necessitate wrapping LLM modeling technology with use-case-driven reliability constraints like RAG, citations, ReAct, etc. This work reduces uncertainty and moves systems from sounding right to being right—and performing consistently at scale.

Key considerations:

- **Build resilience against non-deterministic AI outputs.** The interplay of embeddings, LLM outputs, and vector retrieval creates unpredictability, making fine-grained control difficult. Organizations should accept this ambiguity and design resilient workflows around this limitation. Invest in frameworks like ReAct and autonomous agents to enhance reliability, while carefully managing the associated increased costs and latency trade-offs.
- **Optimize cloud scaling for stability.** Latency, throughput, and costs fluctuate as cloud providers adapt to support agentic AI's growing demand. Something as simple as query response times may vary simply by resource availability variations over the day. Organizations should consider multi-cloud strategies, caching mechanisms, and dynamic resource allocation to reduce disruptions and maintain performance stability.
- **Enhance LLMs' ability to generate meaningful insights.** Even with larger models and huge context windows, LLMs do not inherently know what makes a summary or retrieval "good." Thus, single-shot prompts often yield passable but suboptimal results. To bridge the gap, organizations must implement system-level designs, from retrieval augmented generation (RAG) to agentic AI solutions.
- **Use RAG systems to improve performance.** Use automatic (semantic) information retrieval and incorporate it into the context to improve accuracy and reduce hallucinations. The RAG framework is a useful baseline that adds minimal complexity over simple LLM generation.
- **Expand beyond basic GenAI capabilities with agentic AI.** Simple GenAI use cases, such as retrieval, summarization, and explanation, should generally be tackled using RAG and direct LLM generation. Use agentic AI for more complex tasks like knowledge synthesis, reasoning, decision-making, self-correction, refinement, and adaptive workflows. Multi-step agentic approaches like reranking, metadata filtering, hybrid search, or GraphRAG (which models relationships between data points) are likely required to further refine performance for real-world solutions.

Strategy | Development | Change management

5. The field of agentic AI is moving fast

There have been many developments in agentic AI, ranging from developing protocols and governance best practices to solution patterns that improve performance across all benchmarks. This creates new areas of sophistication and value but also introduces additional complexity and cost for businesses to consider.

Key considerations:

- **Leverage both open- and closed-source models for optimal performance.** Frontier models across different providers now offer comparable performance, and a balance must be struck between the business's tech stack, implementation complexity, and managed service costs.
- **Anticipate significant (but decreasing) infrastructure costs.** While infrastructure costs remain significant, they are expected to decrease as the agentic AI technology stack matures and open-source alternatives gain traction. This means that projected costs per user will likely decrease over time.
- **Build modular solutions.** Agentic AI design patterns and applications are evolving in real time. Therefore, it is imperative to develop modular and systematic approaches to building scalable and extensible solutions.
- **Model Context Protocol (MCP).** MCP provides new opportunities for scalability concerning agent-to-API communication and interactions.
- **Responsible/Ethical AI must keep up with new use cases.** While the AI landscape is evolving rapidly, it is vital to maintain ethical considerations related to AI governance, regulatory compliance, and societal responsibility. Agentic AI requires heightened guardrails and limitations around privacy and data security.

Strategy | Development | Change management

Bonus: Rethink chatbot interfaces

While chatbots are intuitive frameworks, they are not always optimal. The pace of interaction is often constrained by human input, and the overall effectiveness of knowledge and value extraction is limited by a user's ability to express their intent clearly. Chatbots generally induce a slow, manual, and iterative exploration of underlying data, with limited differentiation over out-of-the-box LLM solutions. While useful in certain situations (e.g., in a co-pilot-type function), these solutions tend to have limited differentiation and generally favor the buy side of the "build vs. buy" debate.

Key considerations:

- **Use chatbots for exploration.** Leverage chatbots where conversational AI is right for the use case (e.g., augmenting data discovery tasks, such as interacting with internal documentation or as a customer interface to query information).
- **Consider the user experience.** Users often have misaligned expectations for how chatbots should behave, with overinflated expectations for humanlike behavior. This often leads to frustration during use, as the interface, usage, and outcomes do not align with user intuition. Ensure the user experience matches human intuition for seamless adoption.
- **Transition toward agentic AI for high-impact end-to-end applications.** Automated agentic AI workflows that minimize human intervention can unlock greater efficiency and scalability. Instead of keeping humans in the loop, businesses should focus on strategically placing them "on the loop" for oversight and governance before actions taken by an agent are finalized.

Strategy | Development | Change management

Conclusion

The evolution of GenAI and agentic AI necessitates a shift in how enterprises approach AI strategy and development. While it is easier than ever to prototype AI solutions, success at scale requires rigorous validation, thoughtful change management, agile development, and user-centric design. By integrating strategic governance, robust feedback mechanisms, and intuitive UX, businesses can maximize the impact of AI solutions in real-world environments.

Authors

Michael Luk
Managing Director
Deloitte Consulting LLP
miluk@deloitte.com

Celia Ludwinski
Specialist Leader
Deloitte Consulting LLP
cludwinski@deloitte.com

Ian Thompson
Principal
Deloitte Consulting LLP
iathompson@deloitte.com

SFL Scientific, a Deloitte business, is the artificial intelligence/machine learning (AI/ML) technical arm of Deloitte Consulting LLP. We pair research-grade science with production engineering to turn frontier methods in agentic large language models (LLMs), mixture of experts, computer vision, and robust evaluation into systems capable of standing up to real-world constraints. Our scientists design the methods and our engineers harden them for cloud, graphics processing unit (GPU), and edge. SFL Scientific also enables client teams through hands-on delivery, platform benchmarking, and applied learning via Deloitte's AI Academy. Together, we help organizations establish AI capabilities that can scale efficiently, withstand change, and create lasting value.

Deloitte.
Together makes progress

SFL Scientific
a Deloitte business