How to Calculate Business Value and Cost for Generative AI Use Cases

12 February 2024 - ID G00805323 - 34 min read

By Rita Sallam, Bern Elliot, **and 7 more**

Data and analytics leaders must assess the potential benefits and cost of new generative AI investments. Experimentation can be done inexpensively for most use cases, but this research provides a decision framework for assessing cost and realizing value from enterprise generative AI initiatives.

Overview

Key Findings

- Organizations demonstrating early business value from generative AI (GenAI) are building a portfolio of use cases focused on defending their competitive position with incremental improvements, extending current processes for differentiation or upending their industries, core process and business models, but the vast majority of Gartner client-reported GenAI investments currently fall into the “defend” and “extend” categories.
- Productivity gains are the dominant initial benefits reported by early adopters. Most outcomes from productivity impact leading indicators of future value as opposed to direct financial benefit (such as immediate cost reduction). This requires a higher tolerance for indirect, future financial results versus immediate ROI.
- How benefits from higher productivity will contribute to short-term versus longer-term financial outcomes depends on how newfound productivity is used and proactively managed by leaders.
- Identifying and measuring what users intend to do with the benefits during pilot, rollout and production helps overcome value attribution challenges reported by early adopters.
- Continuous value and cost monitoring of GenAI efforts, including ongoing tracking of cost-impacting market innovations and pricing, ensures expected value realization and controls uncertain costs and risks.
- Collaborating with HR, finance, legal and corporate strategy as early partners optimizes benefits realization by ensuring efforts are in place to manage change, strategically use time savings from productivity improvements and minimize risk from negative impacts of AI.

Recommendations

Data and analytics leaders tasked with assessing and realizing the value of generative AI should:

- Identify the organization's AI ambition and investment focus by working with business leaders to determine strategic intent to use AI to defend, extend or upend your competitive position and industry. Align your GenAI portfolio to your ambition.
- Realize improved productivity, cycle time and quality for specific tasks by investing in "defend" use cases such as GenAI productivity assistants.
- Extend and improve existing business processes to create differentiation and competitive advantage by embedding GenAI in applications and leveraging unique enterprise data.
- Disrupt or upend your industry or create new markets by investing in new strategic GenAI products, services, core processes and business models. These require more aggressive investments and a higher tolerance for risk and complexity and for strategic as opposed to direct, tactical benefit.

Strategic Planning Assumptions

- Through 2025, 30% of generative AI projects will be abandoned after proof of concept (POC) due to poor data quality, inadequate risk controls, escalating costs or unclear business value.
- By 2026, more than 80% of independent software vendors (ISVs) will have embedded generative AI capabilities in their enterprise applications, up from less than 1% today.
- By 2028, more than 50% of enterprises that have built their own large models from scratch will abandon their efforts due to costs, complexity and technical debt in their deployments.

Introduction

Data and analytics leaders have become key strategic advisors and players in developing and executing on their organization's GenAI strategy. In partnership with other business leaders, they have the potential to play a major role in realizing unprecedented productivity improvements, competitive differentiation and business transformation for their organizations.

Earlier adopters across industries and business processes are reporting a range of business improvements that vary by use case, job type and skill level of the worker. [1] A large majority of business executives who are implementing or actively planning to implement GenAI have anticipated or realized benefits from their implementations, according to the Gartner Generative AI 2024 Planning survey of 822 business leaders. [2] On average, survey respondents report:

- 15.8% revenue increase
- 15.2% cost savings, 4.6% through reduction in headcount
- 22.6% productivity improvement

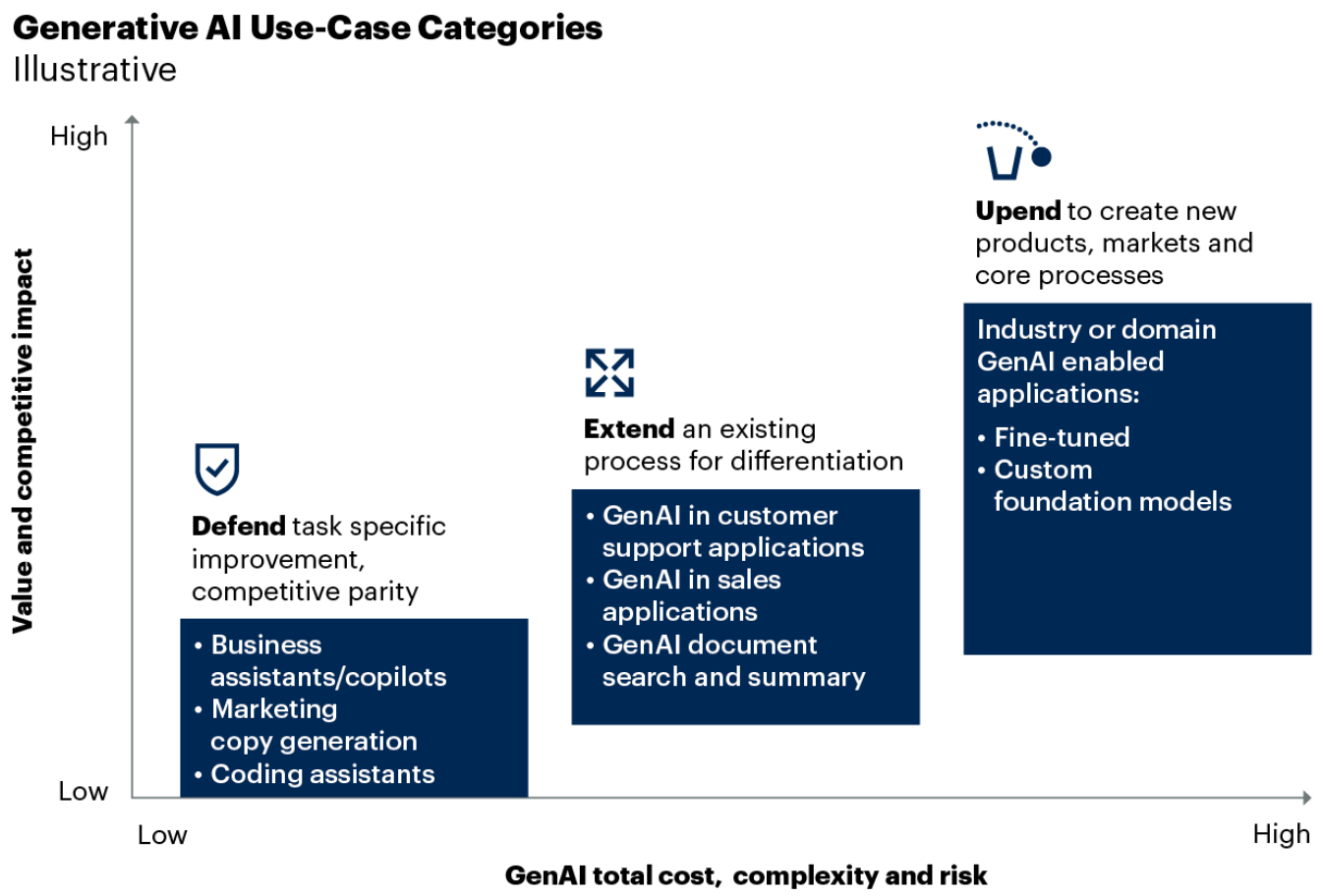
Worker productivity has broadly been found to improve when GenAI is employed:

- ChatGPT has been shown to improve worker productivity by 37%. [1]
- GenAI coding assistants can result in 7% to 55% worker productivity improvements. [3]
- GenAI conversational assistants can improve customer service and support agents' productivity. (studies show a range of 14% to 35% improvement). [4,5]

2024 will be the year organizations scale their piloted use cases. D&A leaders must scale early results both from a value and cost perspective, as they play a critical role by delivering AI-ready data, people and governance.

This report provides a framework for building ballpark estimates of the total cost and value of eight of the most common GenAI use cases Gartner clients have invested in, falling into the three competitive impact categories (see Figure 1). To demonstrate the approach, the analysis includes sample calculations and assumptions in downloadable Excel and PowerPoint files that can be refined through further analysis and deployment experience. You can replace the sample assumptions and calculations with your own.

Figure 1: Generative AI Use-Case Categories



Links to these estimates for each use case are below and include upfront and recurring costs, estimated value potential, sample key performance indicator (KPI) categories impacted as well as sample case studies.

Your AI Ambition

Defend Use Cases

Extend Use Cases

Upend Use Cases

The assumptions and calculations used are detailed in the Microsoft Excel and PowerPoint downloads attached.

To access the Excel and PowerPoint downloads for this document, click on the links at the end of this Introduction or select "Downloads" from the menu along the left side of the screen.

Organizations demonstrating early business value from GenAI are building a portfolio of use cases. While a few are aiming to upend their industries, the majority of investments are focused on defending a competitive position with incremental improvements or extending current processes for differentiation.

Calculate Business Value and Cost for GenAI Use Cases

GenAI Value and Cost Drivers

Analysis

Identify Your AI Ambition With New Risk-Reward Equations and Metrics for Success

The road to realizing the value of generative AI should start with identifying the organization's AI ambition. D&A leaders, in partnership with other technology and business leaders, must assess their strategic intent to use AI to defend, extend or upend their competitive position and industry (see **Gartner AI Opportunity Radar: Set Your Enterprise's AI Ambition**).

A Higher Tolerance for Future Financial Gain Is Needed

Regardless of AI ambition (defend, extend, upend), GenAI requires a higher tolerance for indirect, future financial investment criteria versus immediate ROI. Historically, many CFOs have not been comfortable with investing today for indirect value in the future. This reluctance can skew investment allocation to tactical versus strategic outcomes.

Value measurement and realization for GenAI (and, more broadly, AI or any technology) are very specific to a use case, domain or industry. The vast majority of improvements will accrue to leading indicators of future financial value (productivity, cycle time, customer experience, brand, quality, faster upskilling of junior people, etc.). Unless these benefits translate into immediate headcount reduction and other cost reduction, financial benefits accrue over time depending on how the generated value is used (e.g., to do more with less as demand increases, to use fewer senior workers, to lower use of service providers, to improve customer and employee value that leads to higher retention, etc.). These leading indicators of future value can and should be measured to represent tactical and strategic business impact.

Manage Upending, Transformative Investments Like a Venture Capitalist

GenAI has the potential to completely transform industry dynamics, create new market winners and losers or create new industries. These initiatives should be evaluated and managed as a venture capitalist would do — on strategic, competitive and market-level metrics with a higher tolerance for risk. Strategic metrics, such as size of new market created or percent revenue from new GenAI/AI products, are more appropriate for "upending" initiatives. Completely new metrics may be needed (for example, annual recurring revenue, or ARR, emerged as a way to measure cloud performance versus revenue per PC for on-premises software vendors).

Strategic competitive considerations (your AI ambition) may warrant a more aggressive risk tolerance and investment decision criteria that prioritize long-term strategic benefit over short-term ROI.

Benefits Realization Depends on Executing Flawlessly on People and Process Changes

Beyond executing on the technical deployment, realization of these potential benefits will depend on strategic alignment and managing and investing in the people and process changes necessary to deliver the expected business outcomes (see **Quick Answer: How to Assess Your Readiness for the AI Disruption**).

Strategic considerations may include:

- Aligning investments to your organization's AI ambition in the context of your business strategy (see **Gartner AI Opportunity Radar: Set Your Enterprise's AI Ambition**)
- Managing change for new processes and ways of working
- Absorbing and strategically managing productivity gains and create new ways to measure impact
- Upskilling or acquiring the right technical talent — investing in enterprisewide AI literacy
- Scaling deployments responsibly using repeatable AI engineering processes
- Defining new roles specific competencies required to get the most from GenAI investments, such as prompt engineering, "validation" and other human-in-the-loop activities
- Designing the right human augmentation versus automation balance to leverage the strengths of both humans and AI
- Managing risk, security and governance

Costs Can Be Uncertain and a Major Barrier to Realizing Value at Scale

Unlike previous computing paradigm shifts, the process of training and running large language models (LLMs) has high structural costs. Even when LLMs with billions of parameters are trained, they continue to require massive computing power to run because they do billions of calculations on specialized GPU hardware every time they respond to a prompt. These costs are reflected in vendor pricing models, which most buyers find confusing and are currently in flux, with frequent changes likely to continue. To manage uncertainty, identify and simulate lower and upper bounds of model and total costs as part of a GenAI value and risk assessment. The framework provided in this research and accompanying files gives you a starting point to conduct and refine this analysis over time.

In addition to model training and running, assess additional (potentially hidden) cost drivers during proof of value and cost pilots. Although experimentation and pilots can be done relatively inexpensively, assessing the competitive impact and value of use case options and estimating the future total costs of enterprise deployments at scale can be a challenge. Usage patterns are not yet fully understood, which drives token-based pricing and compute costs.

Token-based pricing, a model that has been around for some time, is now used by LLM vendors to attach pricing to the amount of text entered into prompts and the output of GenAI queries. These vary by use case and can be a significant cost, depending on pricing model and buy vs. build decisions.

Current vendor pricing models that pass on the high cost of innovation and developing, training and running LLMs may result in a negative ROI for many seemingly high productivity-saving use cases when deployed at scale — even when pricing is subsidized by vendors attempting to gain early market share. Higher costs, which become clearer as the solutions are rolled out at scale, can be due to:

- Usage patterns
- Deployment models
- Accuracy needs
- Initial and ongoing model training and inference
- Token pricing
- GPU pricing
- Cloud Infrastructure
- Cost optimization techniques as they evolve
- Incremental investments in application development, cloud and AI infrastructure
- Data management
- New talent and skills
- New systems to support changes in work and processes
- Risk management

Pricing models and techniques that lower costs are already evolving as the market matures. OpenAI, for example, has lowered prices several times and introduced several versions since the

launch of ChatGPT in November 2022.

Realize Task-Specific Benefits With “Defend” Use Cases

> **Defend Use Cases**
>
> **Typical pricing models:** Per-user SaaS
>
> **Build vs. buy:** Buy
>
> **Sample vendors/products:** Adobe Firefly, Amazon CodeWhisperer, Amazon Q, Gemini for Google Workspace, Jasper.ai, Microsoft Office 365 Copilot, OpenAI ChatGPT Enterprise
>
> **Time to value:** Short (less than one year)

Many organizations have begun the process of identifying and prioritizing GenAI use cases by assessing business value, urgency, cost and risk. A recent Gartner survey showed that 55% of the 1,419 organizations surveyed in September of 2023 plan to increase their investment in AI. This is up from 45% of 2,544 survey respondents asked the same question in March and April of 2023. [6] Most early use cases being explored by Gartner clients fall into the “defend” and “extend” categories.

Incremental, task-specific benefits from GenAI productivity assistants should be measured on time savings spent for those specific tasks, across aggregate tasks related to specific processes. While time savings value could translate into immediate headcount reduction, most Gartner clients are not initially taking that approach. For now, the vast majority are planning to *use* the additional productivity to do more with the same headcount in the future (cost avoidance) with more junior, lower-salaried staff. They are also assessing how this extra time can improve metrics such as time to market, quality, and customer and employee experience, which impacts churn and cost to acquire in the future, and so on. Reducing developer churn, for example, can save the recruiter fee (three to four month’s salary), plus increase productivity during the time to onboard a new developer.

Early adopter experiences at large scale are limited. However, they suggest that GenAI embedded in productivity applications can result in double-digit productivity gains and significant improvements to user experience for specific tasks. Given these applications’ low barriers to adoption and broad democratization that is possible with a focus on upskilling, productivity gains can quickly become competitive table stakes, resulting in minimal competitive differentiation. However, the depth and duration of the competitive moat enterprises can realize from defend (and extend) types of investments depends on:

- How well an organization manages how users spend their newfound productivity to positively impact the business
- How GenAI capabilities are integrated with other business processes

- How small micro innovations are aggregated into larger synergistic enterprise benefits
- How well risks are mitigated
- How skilled users become at fully exploiting their potential to impact their work

Moreover, incremental costs for implementation, data management, risk management and security, training and change management, and content and benefit auditing should be assessed as part of the pilot. For example, additional investment may be required for integrating with software development applications, like Jira for coding assistants or creating a knowledge architecture that enables relevant content and data to be exposed to GenAI for business productivity use cases.

Table 2 shows total cost and value ballpark estimates for three common defend GenAI use cases that most Gartner clients have either in pilot or early rollout phases.

Table 1: Estimated Total Cost of Ownership and Value Calculation Ranges for "Defend" Use Case Examples

Estimated total cost of ownership and value components	Coding Assistants	Business Productivity	Marketing Content Creation
Initial pilot and rollout, development, deployment, integration, training	**~$100,000 to $200,000** Costs include developers, platform engineering, security, risk and governance for three months Plus user training	**~$400,000 to $500,000** Costs include Office 365 Specialists doing SharePoint optimization, risk management, security, governance and audit; change management and training for 12 months	**~$200,000 to $250,000** Costs include developers, data engineers, data scientists, security, risk and governance, product management working for three months Plus user training

Recurring costs	**~$280 to $550 per user per year** includes SaaS application pricing and 10% of initial deployment costs for 200 users; no incremental AI and data management licenses	**~$220 to $410 per user per year** includes SaaS application pricing and 10% of initial deployment costs for 1,000 users; no incremental AI and data management licenses	**~$1,000 to $2,100 per user per year** includes SaaS application pricing and 10% of initial deployment costs for 100 users plus AI and data management licenses
Value	**~$1,600 to $6,000 per user per year** for 7% to 30% productivity improvement on 25% of time spent writing code with a "productivity leak"* of 20% to 30%.	**~$7,000 to $16,000 per user per year** for 25% to 59% productivity improvement assuming 30% of time spent on creating business documents with a 20% to 30% productivity leak	**~$2,000 to $4,000 per user per year** for 25% to 59% productivity improvement, assuming 90% of time spent on marketing content creation with a productivity leak of 20% to 30%.

Sample KPIs impacted	• Stories or story points/sprint • Stories completed • Deployment frequency • Number of changes having a customer impact • Number of commits • Change lead time • Number of changes per week • Pull request iteration time • Lines of code per developer • Developer retention • Time to upskill junior developers • Ratio of junior staff to senior staff • Average salary per developer • Time to proficiency/ onboard	• Time saved to create business documents • Employee retention • Percent of time spent with customers • Quicker time to proficiency • Time spent on email per day • Time to create content • Time spent analyzing data • Return on employee	• Content output per marketer — do more with the same resources • Quality of output — time for rework • Average skill level per marketer: Junior marketers perform at expert levels • Average time to create videos, blogs, marketing copy • Click-through rate
Sample case studies	Case Study: Piloting GitHub Copilot to Boost Developer Productivity	**Video: How AI-Ready Data Drives Generative AI Innovation** **Vodcast: AI-Ready Data Provides the Foundation for Generative AI Success**	**Novo Nordisk Delivers Diabetes Care One AI-Optimized Message at a Time** (Phrasee)

Note: See downloadable files for basic assumptions and calculations that you can modify for your specific case.

* “Productivity leak” refers to the percentage of time saved that is not used to improve the business. For example, a developer could save one hour a day, but may choose to get an extra cup of coffee with 10 minutes of that time savings.

Source: Gartner (February 2024)

Recommendations

Rank, prioritize and pilot:

- Collect use cases from employees across the enterprise and conduct a strategic top-down analysis with top business stakeholders and the strategy team to identify potential use cases that align to the business strategy and AI ambition. Rank, prioritize and vet by potential impact and risk. (See **Toolkit: Discover and Prioritize Your Best AI Use Cases With a Gartner Prism**.)
- Once you have determined your AI ambition with senior leadership, identify and prioritize high-value GenAI use cases by competitive impact, business value, urgency, cost and risk by calculating incremental upfront and ongoing costs versus the business outcomes generated over time.
- Launch low-cost pilots (see **How to Pilot Generative AI**) to gain a better understanding of which roles are realizing specific productivity gains, the associated financial benefit and the additional costs required when deploying at scale. Also identify outcomes that can’t easily be quantified in financial terms. Simulate upper and lower bounds for benefit and cost to account for uncertainty.
- Build a portfolio of defensive, differentiating and upending GenAI use cases that combine initiatives with hard ROI and those delivering benefits and competitive advantage that are difficult to initially quantify directly in financial terms.
- Use fast-cycle innovation approaches to assess ongoing value realization and costs in addition to technical feasibility.

Assess cost and value:

- Identify financial and nonfinancial (financial influencing, strategic) success metrics for each pilot or initiative and create a process for measuring actual costs and benefits. This includes:
  - Measuring ‘intent’ for use of time from newfound productivity
  - Incorporating GenAI costs and benefit measurement into the FinOps processes
  - A/B testing of impact

  - Using automated productivity metrics in existing enterprise and software development applications and platforms where possible
- Account for total costs that may be required to seize benefits. Such costs might include investments in new types of data repositories (e.g., embeddings/vector databases), updating the knowledge architecture or knowledge graph and/or data labeling. The need for high-quality data doesn't go away with LLMs.

Measure and manage outcomes:

- Understand who will use the capability for what, and determine how it will impact their work. Add new metrics or deprioritize less relevant ones as deployments progress based on actual benefits realized (see **How to Optimize Enterprise Value From Data and Analytics**).
- Partner with finance to develop a robust AI finance capability and cost-benefit analysis process. This should include a set benefits realization window (e.g., 12 months) to compare investment alternatives; role identification for finance support to approve benefit and attribution calculations; and the reporting process to track, validate and report financial benefits.

Ensure value realization:

- Invest in upskilling and change management for users. Proactively manage the use of newfound productivity through new performance goals for users.
- Obtain or upskill resources who can identify and assess business value impacts, implement valid measurement systems and conduct multivariate analysis to demonstrate benefits with confidence.
- Draft acceptable use guidelines with general counsel to address the particular IP and confidentiality risks associated with public LLMs (see **Generative AI: 4 Decisions to Make When Creating a Policy**).

Extend and Differentiate By Applying GenAI to Existing Processes

> **Extend Use Cases**
>
> **Typical pricing models:**
>
> *Buy:* Per-user SaaS and/or per user plus consumption
>
> *Build:* Consumption of LLM APIs for custom GenAI applications grounded using RAG.
>
> **Build vs. buy:** Buy or build
>
> **Sample vendors/products:**

> *Buy:* Dialpad, EinsteinGPT add-ons to Salesforce cloud services, Forethought, Microsoft Copilot for Dynamics 365
>
> *Build:* Amazon Bedrock with SageMaker, Cohere, Databricks MosaicML, Google Vertex AI, Microsoft Azure OpenAI Service, OpenAI
>
> **Time to value:** Medium (between one and two years)

Investments in differentiating use cases that leverage GenAI embedded in enterprise, domain and industry applications or custom applications have the potential to improve specific existing business processes. Differentiating use cases can leverage enterprise data in unique ways for more defensible competitive advantage than “defend” use cases, but come with higher and more unpredictable costs and risk at scale.

“Extend” differentiating GenAI use cases will require higher investment than defensive ones. The costs will also be more unpredictable and must be assessed to determine whether they can be offset by financial benefit resulting from productivity gains and future financial gains from use of productivity gains, as well as the potential for revenue generation.

Beyond the low-level models, realizing competitive advantage and full benefits from GenAI will in large part be driven by how effectively an organization can redesign work/processes and manage change and risk.

Table 3 shows ballpark total cost and value estimates for three common extend GenAI use cases that Gartner clients have either in pilot or early rollout phases.

Table 2: Total Cost of Ownership For “Extend” Use-Case Examples

Estimated total cost of ownership components	**GenAI-assisted support**	**Personalized sales content creation**	**Document search and summarization**
Initial pilot and roll-out, development, deployment, integration, training	**~$750,000 to $1 million** Costs include developers; data engineers; data scientists; security, risk and governance; product management working for six months Plus user training	**~$750,000 to $1 million** Costs include developers; data engineers; data scientists; security, risk and governance; product management working for six months Plus user training	**~$750,000 to $1 million** Costs include developers, data engineers; data scientists; security, risk and governance; product management working for six months Plus user training
Recurring costs	**~$2,000 to $11,000 per user per year** includes: • SaaS plus API consumption licenses and varies call volume at the upper end • Application and model maintenance of 15% of initial deployment costs for 500 agents • AI and data management licenses	**~$1,300 to $11,000 per user per year** includes: • SaaS plus API consumption licenses on upper end • Application and model maintenance of 15% of initial deployment costs for 500 reps • AI and data management licenses	**~$790 to $1,200 per user per year** includes: • API consumption fees • 3 to 1 GenAI query to search usage ratio • Three API calls per GenAI query • GPUs for embeddings • Application and model maintenance of 20% of initial deployment costs for 1,000 users • AI and data management licenses

Estimated total cost of ownership components	**GenAI-assisted support**	**Personalized sales content creation**	**Document search and summarization**
Value	**~$7,000 to $15,000 per agent per year** for 14% to 35% productivity improvement with a productivity leak of 20% to 30%.	**~$6,000 to $16,000 per year per rep incremental contribution to EBITDA** for 30% to 59% productivity improvement on 9.5 hours per week spent on creating sales content with a productivity leak of 20% to 30%, with 50% of saved time used for selling and EBITDA range of 20.3% to 33%	**~$4,000 to 10,000 per knowledge worker per year** for 20% to 59% productivity improvement with a 20% to 30% productivity leak, assuming 30% of time spent on searching for information

Estimated total cost of ownership components	**GenAI-assisted support**	**Personalized sales content creation**	**Document search and summarization**
Sample KPIs impacted	• Time to resolution • Response time • Net Promoter Score (NPS) • Call satisfaction scores • Agent retention • Productivity of junior staff • Time to expert skill level • Time to productivity • Number of escalations • Manager's time to train new staff • Incremental revenue upsell/cross-sell • Level of detail of call center analytics (topic ID on all calls rather than a sample)	• Revenue per sales rep — Higher sales productivity leads to more deals per rep in the funnel • Average deal size • Percent cross-sell/upsell revenue • Size of pipeline • Close rate • Customer retention • Cost to acquire • Customer satisfaction	• Percentage of knowledge worker time spent on the highest-value tasks of their role (respond to customers, build products, selling, managing suppliers, contracts, claims, legal documents) • Time to answer client questions • Content quality/error rates

Estimated total cost of ownership components	GenAI-assisted support	Personalized sales content creation	Document search and summarization
Sample case studies	**Upwork Reduces Time to Resolution by 50% With Forethought** (Forethought)	**How Uniform Used AI to Build a Video Marketing Strategy** (HeyGen)	**Meet Lilli, Our Generative AI Tool That's a Researcher, a Time Saver, and an Inspiration** (McKinsey)

EBITDA: earnings before interest, taxes, depreciation and amortization

Note: See downloads for basic assumptions and calculations that you can modify for your specific case.

Source: Gartner (February 2024)

GenAI-Assisted Support

Organizations can directly use commercial applications that have generative AI capabilities embedded within them. For example, CRM vendors are offering GenAI virtual agent capabilities embedded in customer support applications, and native GenAI customer support application vendors have emerged. Over the next 12 to 18 months, productivity from GenAI customer service use cases is expected to increase by 17.2% on average, per Gartner's Generative AI 2024 Planning Survey. [2]

Erik Brynjolfsson's study on GenAI's impact on customer support [5] showed the following:

- Productivity increase of 35% for junior staff — reduction was extremely high among most junior staff
- Exposes lower-skill workers to the best practices of higher-skill workers
- Reduced time of handling issue by 9%
- Reduced issue resolution by 14% per hour
- Reduced escalations to speak to a manager by 25%
- Time spent by managers training junior staff was decreased
- GenAI reduced agent attrition by 40%

Other benefits observed with Gartner clients in early prototyping and rollout include:

- Making the frontline agents smarter, reducing need for managers conducting quality assurance checks and resolving escalations
- Improved self-service, leading to fewer human agents
- Improved call analytics, helping identify where and how to improve service and, thereby, customer satisfaction

Personalized Sales Content Creation

An average 15.8% of revenue increase is expected from GenAI deployments, per Gartner's Generative AI 2024 Planning Survey. [2] The personalized sales content creation use case would leverage an LLM natural language user experience and content creation, text analytics and public data combined with enterprise data such as customer and transaction data to prospect, prepare for meetings and create hyperpersonalized sales pitches and communications. This use case could be purchased as part of the CRM application GenAI add-on, similar to the GenAI customer support use case, or built as a custom application using a similar approach to the document search and summarization use case.

Productivity improvements would reduce the amount of sales rep time spent on prospecting and customer meeting preparation and postmeeting activities. Sales reps could use a portion of the additional time to generate more revenue. Higher-quality and more personalized sales content could also translate into potentially higher close rates and higher average deal sizes. Higher sales productivity would lead to more deals per rep in the funnel, and higher-quality interactions with customers could lead to higher customer retention.

Document Search and Summarization

Putting a conversational interface in front of a document repository is intended to make it faster for a knowledge worker to find and summarize the textual information they need to do their jobs. For example:

- A mortgage underwriter condensing the time to close a loan
- A consultant creating a client deliverable or responding to a query faster
- A job recruiter reducing the time it takes to write job advertisement descriptions.

With newfound time, knowledge workers can do more of their core work, such as selling, creating client content, making decisions, identifying additional account or prospect opportunities or even prioritizing account engagements, etc.

One approach to creating this type of application is to embed model APIs into a search- and LLM-based application. RAG techniques are the current state of art for incorporating additional information into the query response beyond what was in the foundation model's training data. Extending models via a RAG approach can provide an appropriate balance between bringing organizational context into foundation models with the complexity and cost of modifying the

underlying models (fine-tuning or building models from scratch; see **How to Choose an Approach for Deploying Generative AI**). Other techniques will continue to emerge.

Model how usage patterns will be different with GenAI queries versus enterprise search. For example, a user may do one or two searches, then scroll down and click on links. With GenAI, a user may do multiple iterations of the search to refine their output. The number of interactions and sub-API calls to improve accuracy will drive the ultimate cost per GenAI query.

Given that the use-case deployment approach is to embed model APIs into a custom application, there will be incremental cloud infrastructure costs/GPUs for inference, including vector database embeddings. Depending on current capabilities, additional technical full-time equivalents (FTEs) may be needed to deploy and maintain the application, including:

- Change management
- Data management (for example, knowledge graph or vector database enhancements/embeddings to improve accuracy)
- AI infrastructure costs
- Risk management
- Changes to underlying systems to support new work processes and flows

Recommendations

- Conduct scenario analysis and planning as part of the proof of concept and pilot to assess the upper limits of licensing tokens and other inference costs at scale. Usage patterns for search will likely be different than for GenAI queries.
- Assess potential incremental costs for implementation, data management, running and maintaining the models, risk, security, systems to support new processes, training on new skills, and change management. Refine these during beta, phased rollout and go-lives.
- Identify metrics that capture both financial benefits and strategic outcomes — like better user experience, broader access to capabilities previously requiring higher skills, and employee and customer satisfaction — and assess their impact.

Upend the Industry With Game-Changing, Transformative GenAI Use Cases

> **Upend Use Cases**
>
> **Typical pricing models:** FTEs and fine-tuning, model inference and compute and the cost to build, train/retrain, run and maintain a fine-tuned or custom LLM.
>
> **Build vs. buy:** Build

> **Sample vendors/products:** Amazon Bedrock with SageMaker, Cohere, Databricks MosaicML, Google Vertex AI, Microsoft Azure OpenAI Service, OpenAI
>
> **Time to value:** Long (more than two years)

Unique competitive advantage and industry disruption from new GenAI products, business models and core processes will require investments in transformative use cases. These bring higher cost, complexity and risk.

Transformational use cases are new products and services that could create completely new market categories and disrupt current ones. They also serve to retain customers by adding these capabilities to existing products (essentially creating new domain- and industry-specific GenAI applications). For example:

- An insurance company could fine-tune a foundation model with its own policy documents to incorporate knowledge into the model and improve its performance on its specific use cases.
- A financial services organization might create a foundation model trained with financial data, which could then be used for many financial services use cases.
- Examples of domain-specific models include BloombergGPT in financial services, Thomson Reuters for legal and Absci for pharmaceuticals. Others are emerging daily.

Many early adopters are experimenting with building smaller, domain-specific models using pretrained open-source models. By building (or buying) smaller, cost-effective custom LLM pretrained for analytics and/or specific industry vertical or business domain needs, enterprises could achieve comparable or acceptable performance and accuracy while significantly reducing costs associated with training and deploying a generic model. Smaller models with fewer parameters offer potentially faster training and inference times and have lower cloud infrastructure requirements such as memory and storage. They also have smaller compute and energy consumption footprints.

However, the old adage remains true:

> **Open source is free like a puppy is free.**
>
> *— Scott McNealy, Sun Microsystems, 2005*

Consider:

- Initial training and ongoing inference, hosting and compute costs shift to you.

- Leveraging open-source tools requires higher levels of skill and FTEs to build and integrate the different libraries of capabilities with limited documentation compared to more packaged commercial offerings.
- There may be differences in enterprise-grade security features.
- General availability and restrictions on commercial use will vary by open-source model.

Being early in a rapidly changing market introduces the potential for technical debt as new techniques and innovations rapidly emerge in exchange for strategic market position benefits and early learning.

“Upend” investments are best-suited for organizations that have an AI ambition to lead and change their industry where investment criteria prioritizes strategic value over task- or process-specific productivity benefits.

Upend investment measurements will include many of the tactical outcomes in defend and upend use cases but should also assess strategic, competitive and market-level outcomes, but there may be new revenue metrics and business models. ARR, for example, emerged to track cloud business models, whereas on-premises software revenue was measured in terms of license and maintenance and KPIs such as revenue per PC.

An upend investment is transformational: For example, an insurance company could invest in a core process that measures the time it takes to microprice risk, as well as the granularity at which it examines the risk. Such a change would completely change the underwriting game and insurance industry profitability.

Investment decision criteria for transformational use cases should prioritize strategic benefits that may be difficult to quantify in financial terms over immediately identifiable task- or process-specific hard financial benefits.

Ongoing innovations in GenAI are refining models and techniques and bringing down adoption costs. However, until lower-cost options emerge, early innovators may have to accept difficult-to-quantify hard financial returns and higher cost, complexity and risk in exchange for competitive and first-mover advantage.

Executives and boards will need to accept higher risk tolerance for industry-transforming and disruptive innovation from new GenAI products and business models. Investment decisions to move forward should be based on strategic, competitive and market-level impact considerations.

Table 4 shows rough order-of-magnitude total cost of ownership estimates that assume using pretrained open-source models as the foundation. The decision to use either pretrained open-source or closed-source models — along with many, many other variables — changes the cost

parameters and equations substantially. Here we show broad cost categories and some base assumptions upon which you can build estimates. However, the details can be quite complex and vary widely based on many factors not detailed here and will evolve as you pilot and roll out your application.

Table 3: Total Cost of Ownership for Upend Use Case Examples

Use case	New domain application fine-tuned LLM for insurance underwriting	New domain application custom LLM for drug discovery
Initial pilot and rollout, development, deployment, integration, training	**~$5 million to $6.5 million** Costs include initial fine-tuning of pretrained open-source 13B parameters; plus developers; data engineers; data scientists; security, risk and governance; product management working for 9 months Plus user training	**~$8 million to $20 million** Costs include initial model training of pretrained open-source 13B parameters, plus developers; data engineers; data scientists; security, risk and governance; product management working for 12 months Plus user training
Recurring costs	**~$8,000 to $11,000 per user per year** Includes ongoing LLM inference costs per GenAI query, GPUs for embeddings, application and model maintenance 25% of initial deployment costs for 1,000 users, plus AI and data management licenses	**~$11,000 to $21,000 per user per year** Includes ongoing LLM inference costs per GenAI query, GPUs for embeddings, application and model maintenance of 25% of initial deployment costs for 1,000 users, plus AI and data management licenses
Value	Use-case-specific	Use-case-specific

Sample KPIs impacted	• Market share • Percent of revenue from GenAI products • Underwriting losses percent • Size of new market created • Reduced time to assess risk from complex mix of contracts and insurance documents • Percent in improvement in claim cost • Percent of claims agent time saved in generating personalized content • Higher claims processing rates • Lower time to process a claim • Higher customer retention • Higher customer lifetime value • Brand as innovator • Percent of services that are now free to drive business in others	• Reduce time for clinical testing of new drugs • Time to market for new drugs • Percent of productivity improvement in document generation such as contracts, briefs and pleadings • For legal, productivity improvements to contract analysis and negotiation, due diligence, discovery, dispute resolution, litigation support, and audit support • Higher-quality diagnosis and treatment; better healthcare outcomes • Reduce error rates in medicine and law • Predict risks and opportunities in cases
Sample Case Studies	**Meet Intuit Assist, a New AI Assistant That Can Do More Than Just Answer Questions** (VentureBeat) **BloombergGPT: A Large Language Model for Finance** (Arxiv) **Thomson Reuters Launches Generative AI-Powered Solutions to Transform How Legal Professionals Work** (Thomson Reuters) **Absci Achieves a Breakthrough in AI Drug Creation** (Absci)	
Note: See downloadable files for basic assumptions and calculations that you can modify for your specific case.		

Source: Gartner (February 2024)

Recommendations

- Create a tiger team, including executive leadership, corporate strategy, and business and technology thought leaders to identify and assess new GenAI-based products, services, and delivery and business models options for the organization that could disrupt the industry.
- Align investments to your organization's AI ambition in the context of your business strategy (see **Gartner AI Opportunity Radar: Set Your Enterprise's AI Ambition**).
- Socialize with the board and executive leadership the need for a higher risk tolerance and new investment criteria that prioritize strategic value over task- or process-specific productivity benefits for transformational GenAI opportunities.
- Build a portfolio of quick wins as well as differentiating and transformation use cases. Use an options-based approach that combines initiatives with hard ROI with loss leaders and those delivering transformation benefits and competitive advantages that are difficult to initially quantify directly in financial terms.
- Monitor and reassess deployment approaches as new market options and processing innovations lower compute costs and as vendor pricing models evolve to drive demand.
- Assess costs that may be required to optimize benefit realization (such as new systems and channels to support new processes and business models). Also assess foundational investments in data quality, data labeling, embeddings in vendor databases, knowledge graphs, new types of data repositories (such as graph databases), and additional security and governance.

Evidence

[1] **Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence**, MIT.

[2] **Gartner Generative AI 2024 Planning Survey:** This survey was conducted to examine generative AI's use case implementation and impact by business function. The survey was conducted from September through November 2023. In total, 822 business executives who lead corporate functions outside IT and who indicated will begin or continue to implement Generative AI across the next 12 months qualified and participated. The research was collected via online surveys in English. The sample was equally split across the following eight corporate functions: finance; HR; marketing; sales; customer service; supply chain; procurement; and legal, risk and compliance. The sample mix by location was North America (n = 536), Europe (n = 176) and Asia/Pacific (n = 110). The sample mix by size was $50 million to less than $500 million (n = 119), $500 million to less than $1 billion (n = 129), $1 billion to less than $10 billion (n = 374) and $10 billion or more (n = 200). *Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.*

In the Gartner Generative AI 2024 Planning Survey, 1,040 business executives that lead functions outside of IT were screened about their plans for Generative AI in 2024 and we found that 79% of these business executives will begin or continue to implement Generative AI across the next 12 months. Those that are currently not planning to implement Generative AI are working to improve their understanding of the technology.

[3] **McKinsey Says "About Half" of Its Employees Are Using Generative AI**, Ventur

eBeat. Gartner inquiries show average productivity improvements of 7% to 20% for coding in English.

[4] **Will Generative AI Make You More Productive at Work? Yes, But Only If You're Not Already Great at Your Job**, Stanford University Human-Centered Artificial Intelligence.

[5] Erik Brynjolfsson, **Generative AI at Work | NBER**

[6] Source: **Generative AI Realities: Proactive Approaches for Quantifiable Business Results** Webinar Polling September 2023; Source: **Beyond the Hype: Enterprise Impact of ChatGPT and Generative AI** Polling March and April 2023.

Contributors