Gartner Research

Pragmatic Guide to Scoping and Delivering Generative AI Capabilities in Tech Product Offerings

Radu Miclaus, Jim Hare

8 September 2023

Gartner

Pragmatic Guide to Scoping and Delivering Generative AI Capabilities in Tech Product Offerings

Published 8 September 2023 - ID G00798693 - 10 min read

By Analyst(s): Radu Miclaus, Jim Hare

Initiatives: Product/Service Design and Creation

As the generative AI market disrupts both technology builders and users, the opportunities abound for designing and delivering impactful GenAI features. Product managers need to uncover and amplify the potential of GenAI to augment user experience and decision making.

Overview

Key Findings

- Existing enterprise technology usage patterns are being disrupted by generative AI (GenAI) due to the change in how users interface and "converse" with software. Consequently, enterprises are in the process of experimentation and prioritization around what capabilities to adopt.
- Early productivity applications for GenAI in the enterprise include communications, knowledge management, creative design and code generation.
- Enterprises experience a mix of excitement and caution when it comes to upcoming widespread adoption of GenAI capabilities. The exact effect on productivity in enterprise-focused tasks is not yet clear; however, enterprise-grade GenAI applications are starting to move from beta and preview to general availability.

Recommendations

Product managers looking at designing and creating differentiated experiences with GenAI need to focus on the following, depending on the target application:

- Target content consumption opportunities by enabling enterprise knowledge access (enrichment, management and retrieval) for grounding generative outputs in defendable factual information in the enterprise.

- Boost productivity through content creation by focusing on building co-pilots and assistants for communications, creative design, business processes and low-code/no-code development environments.
- Deliver robust support for building applications by supporting the emerging field of large language model ops (LLMops) and integrating GenAI models in custom applications.

Strategic Planning Assumption

By 2026, more than 80% of independent software vendors (ISVs) will have embedded GenAI capabilities in their enterprise applications, up from less than 5% today.

Introduction

Target These Three Software Capabilities to Accelerate Enterprise GenAI Readiness

GenAI technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content.

In How TSPs Can Accelerate the Generative AI Readiness of Their Enterprise Customers, we describe content consumption, content creation and technology creation as the main areas of disruption in the enterprise, and how each one is met with different readiness levels by the enterprise. Based on a survey conducted from a Gartner webinar series on GenAI, 70% percent of respondents are exploring, 19% are piloting and 4% have GenAI applications in production. With only one in four organizations either piloting or in production, the readiness curve is still evolving. New possibilities in the user experience and emerging needs for building custom applications powered by GenAI are opening a wide area of opportunity for ISVs.

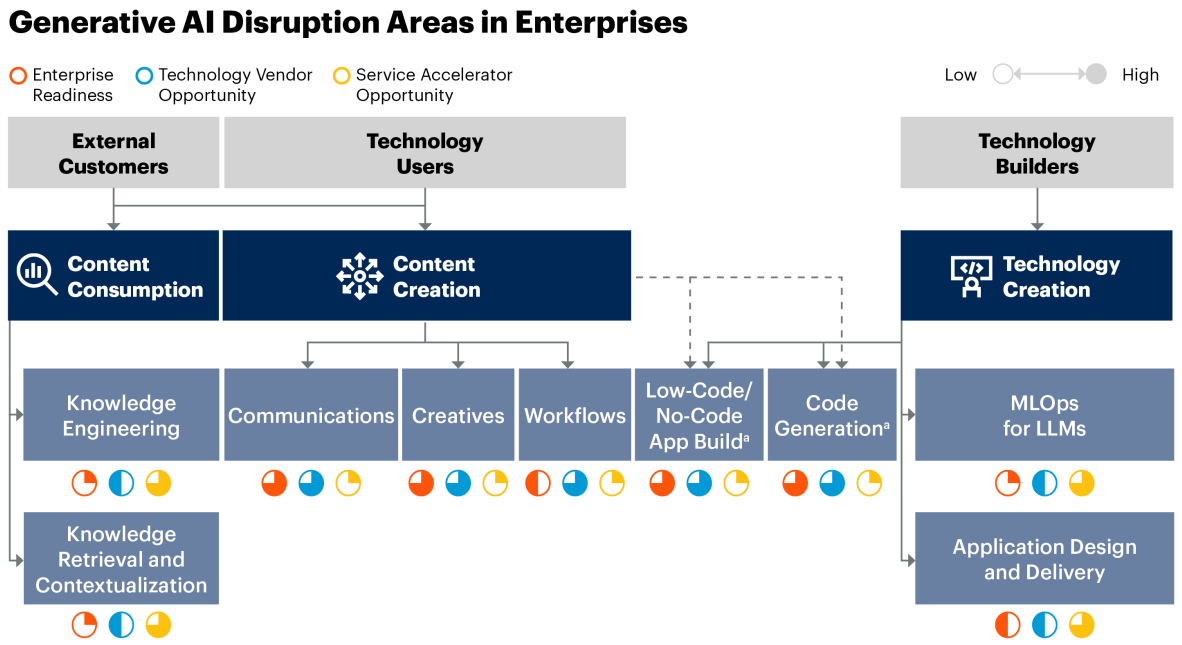
In this document, we dive deeper into the practical capabilities that product managers can focus on to capitalize on the changes in the following three areas (as depicted in Figure 1).

Content consumption in the context of enterprise knowledge manifested mostly in unstructured and semistructured data (policies, documentation, emails, contracts, images, codebases and other repositories) requires special considerations for acceptable usage. Generative content consumption needs to be grounded and informed by facts. This assumes a new focus on knowledge retrieval based on role-based access and building applications that use a hybrid called “retrieval augmented generation.” The new roadmap milestones for ISV are the features and workflows that accelerate enterprises toward configuring and deploying content consumption applications on their proprietary data for internal or external use.

Content creation deals with the usage of productivity tools that are augmented by generative capabilities to speed up and scale output. The use cases encompass communications, creative and design, and business processes and workflows, as well as low-code/no-code development. ISVs, both new entrants and incumbents, will compete heavily in this space with features such as conversational AI via co-pilots/assistants, enhanced prompting experience, automation, decision support and others.

Technology creation powered by GenAI is a growing opportunity for ISVs focusing on data science and app development activities. Enterprises that want to build their own custom LLM and embedded generative applications will need guided workflows and features for rapid model refinement, evaluation, deployment and monitoring. Targeted features and workflows for support of LLMops as well as design, delivery and scaling of GenAI applications will enable ISVs to maintain competitive advantage and stay close to value creation applications in the enterprise.

Figure 1: GenAI Disruption Areas in Enterprises



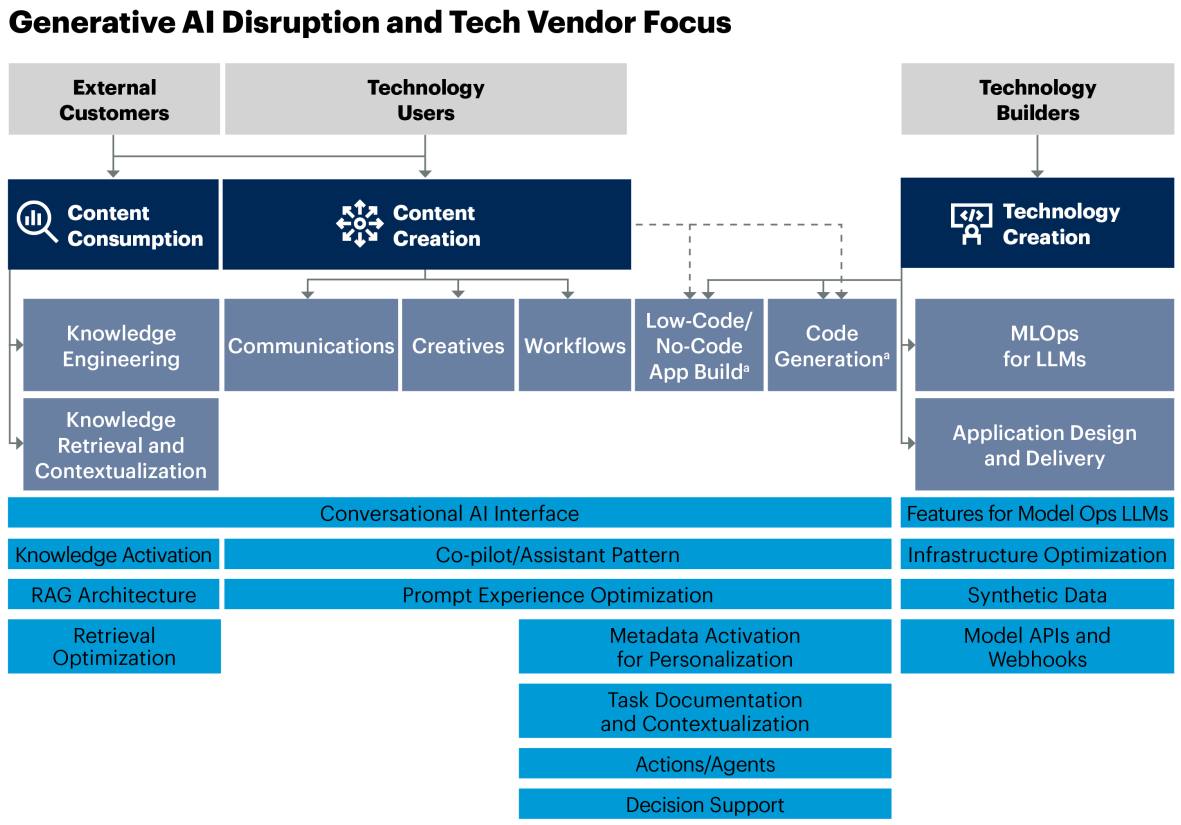
Source: Gartner
[a] Low-code/no-code app build and code generation are content creation tools used by technology builders.
798693_C

Gartner

Analysis

ISVs are the ones pushing the envelope and reinventing the user experience of both consumers and workers. In terms of enterprise audiences, ISV product leaders need to consider how users will interact with embedded generative features within the context of business tasks. But they also need to consider how builders and administrators of technology can efficiently design, build, configure and maintain these enterprise applications. Figure 2 describes the high-level set of capabilities that ISVs should explore, mapped to each area of disruption.

Figure 2: GenAI Disruption and Tech Vendor Focus



Source: Gartner

[a] Low-code/no-code app build and code generation are content creation tools used by technology builders.

798693_C

Gartner

Content Consumption

Usage Pattern: The new expectation for enterprise content consumption is both robust knowledge retrieval and generative synthesis through an intuitive conversational AI interface. The difference between popular consumer tools, such as ChatGPT, Bard or Claude, is that, for enterprises, security, factual validation, role-based access to information and cost-efficiency are essential. Capabilities from vendors enabling content consumption in the enterprise need to map to these needs closely.

Use Cases: Enterprise search, specialized research, investigation tasks, regulatory compliance, consumer self-service and customer support agent augmentation

ISV Profiles: Insight engines, personalization vendors, hyperscalers with search and GenAI services, vector databases, and conversational AI vendors.

Recommendation: Product managers in ISVs who want to support content consumption within their applications should build the capabilities and features shown in Table 1.

Table 1: Capabilities, Descriptions and Features for Content Consumption

(Enlarged table in Appendix)

Capabilities	*Descriptions*	*Features*
Conversational AI Interface	Conversational-based interface as a prompting or chatbot experience.	▪ Natural language understanding ▪ Voice support (for input and output consumption) ▪ Multilingual support ▪ Bot orchestration ▪ Back-end service integration
Knowledge Activation	Activities that prepare enterprise knowledge bases for retrieval and support of generative synthesis.	▪ Knowledge-base ingestion and storage ▪ Data indexing and index management ▪ Data enrichment: entities, topics, sentiment, embeddings transformation ▪ Data staging: vector database support, knowledge graph building and expansion ▪ Role-based access control (RBAC) access configuration
RAG Architecture	Hybrid architecture useful for grounding generative output by allowing information retrieval to inform the prompting of GenAI services.	▪ Staging a search engine/capability with GenAI service in optimal network location ▪ Configuration of search capability on top of activated knowledge base ▪ Options for configuration and integrations between the two services (APIs and documentations) ▪ Performance testing options ▪ Observability features
Retrieval Optimization	To improve output relevance and optimize costs incurred in the GenAI application, retrieval optimization techniques can be facilitated through the search capability configuration.	▪ Ability to configure semantic search capabilities (augmented by vector databases) ▪ Ability to manage and customize search pipelines by search engineers ▪ Infrastructure configurations for optimal performance ▪ Observability

Source: Gartner (September 2023)

Content Creation (Productivity Co-Pilots and Assistants)

Usage Pattern: The new expectation for enterprise users is for GenAI assistants to be served directly and in context of the tasks they need to perform. The experience needs to be informed by the specific content and task characteristics and workflows designated by the activity. The co-pilot and assistant will have back-end components for content and model serving in the experience as well as front-end plug-ins within the context of the task flows.

Use Cases: Types of applications for content creation span communications (internal and customer-facing communications and messaging), creative (design related to images, videos, voice and sound), workflows (business-IT processes like CRM, ERP and data management), and low-code/no-code and code development.

ISV Profiles: Productivity tools, meeting solution providers, digital media design and management tools, development tools, CRM, ERP, supply chain management, data and analytics tools, and PaaS platforms focused on application development.

Recommendation: Product managers in ISVs who want to support content creation within their applications should build the capabilities and features shown in Table 2.

Table 2: Capabilities, Descriptions and Features for Content Creation

(Enlarged table in Appendix)

Capability	*Description*	*Features*
Conversational AI Interface (for Productivity Tools)	Conversational-based interface as a prompting or chatbot experience design in context of the task flow supported by the productivity tool.	▪ Natural language understanding ▪ Voice support ▪ Multilingual support
Co-pilot Design Pattern	Co-pilots/assistants are designed as plug-ins into productivity tools and have a general pattern on deployment that offers versatility in use cases.	▪ Integrate with conversational AI interface as a plug-in ▪ Optimizes the data flow and compute needed for the generative tasks at hand ▪ The plug-ins can be used to bridge the applications with more complex GenAI workflows and agents for multitask completion ▪ User feedback loop fully integrated
Prompt Engineering Experience	The prompt engineering experience is optimized in the context of the task, making use of templates, personalization and recommendations.	▪ Multilanguage support ▪ Prompting templates ▪ Autocomplete recommendations included in prompting experience ▪ Prompting help and suggestions (best practices integrated in the prompting experience)
Content Documentation and Contextualization	The ability to generate documentation and descriptions of tasks completed for productivity support in more complex business processes dealing with multiple artifacts, workflows and tasks.	▪ User prompting for summarization of tasks ▪ User prompting for documentation creation ▪ User prompting for narratives and description generation for artifacts such as data tables, insights and shareable outputs.
Metadata Activation and Personalization	Metadata becomes the input into generative capabilities that can personalize the experience based on past behavior and preferences.	▪ Activation of metadata and user activity tracking throughout the capabilities ▪ Personalize data elements used into prompting recommendations and generative narratives in the application ▪ Contextualized narratives with tone and terminology specific to roles and functions
Actions/Agents	Ability for co-pilots/assistants to complete complex chains of tasks based on prompt-based goal definition and instructions.	▪ Ability for co-pilots/assistants to triage prompt instructions to generative agents for complex task completion ▪ Ability for agents to continuously learn about the preferences of users and optimize similar consecutive requests
Decision Support	In applications that require analysis for decision support, besides the ability to present analytics to users, generative simulations can present variations of outcomes to support decision makers.	▪ Ability for users to prompt the system for a simulation of key performance indicators (KPIs) with various constraints ▪ Ability for users to ask for optimal outcome description of the simulation results to support the decision

Source: Gartner (September 2023)

Technology Creation

Usage Pattern: Besides productivity AI co-pilots/assistants for development tools, enterprises will expect support for building and refining their own LLMs as well as the integration hooks and architectural best practices for integrating them in custom applications. Custom applications will likely be focused primarily on content consumption, and they may have content creation elements, as well as an emphasis on the specialized domains that enterprises want to refine.

Use Cases: LLMops, enterprise architecture, knowledge management, ethics and responsible AI compliance.

ISV Profiles: Data science and machine learning (DSML) engineering platforms, specialized model API services and marketplaces, DevOps tools, app development vendors, computational infrastructure providers, vector database vendors.

Recommendation: Product managers in ISVs who want to support the enterprise ability to build custom LLMs and applications should build the capabilities and features shown in Table 3.

Table 3: Capabilities, Descriptions and Features for Technology Creation

(Enlarged table in Appendix)

Capability	Description	Features
Model Ops for LLM Support	LLMops is the processes, methods and tools used for building, integrating and operating LLMs in production environments.	▪ Infrastructure staging, including compute configurations and vector management facilities (vector libraries or databases) ▪ Options for base model selection mapped to use cases and commercial considerations ▪ Data preparation for training and fine-tuning (Q&A pairs, labeled data) ▪ Prompt engineering and management ▪ Evaluation, bias detection and responsible AI techniques and reports ▪ Process for using reinforcement learning with human feedback (RLHF) ▪ Ability to build LLM chains or pipelines ▪ Documentation options ▪ Deployment packaging and handoff
Synthetic Data	Synthetic data is needed when training data is not enough or is not representative of all variations of outcomes or occurrences.	▪ Ability to build synthetic data using various generative methods based on usage scenarios needed ▪ Use of various data labeling techniques for LLM refinement
Infrastructure Orchestration	The processes of architecture, build and maintenance for the supporting infrastructure of generative services in production.	▪ Access (direct or via integration) to GPU clusters for training and inference ▪ Orchestration of stateful and stateless services for best configuration needed for the use case ▪ Options for managing a vector store or using a managed vector database ▪ Observability across all pertinent services
Model API and Webhook Options	The ability to call/integrate with an external GenAI service via an API or webhook.	▪ Provide or enable access to external specialized GenAI models (hosted or heterogeneous marketplaces) ▪ Ability to call to an external (specialized) hosted generative service via API ▪ Ability to use automated webhooks to different generative services ▪ API life cycle management

Source: Gartner (September 2023)

Evidence

Beyond the Hype: Enterprise Impact of ChatGPT and Generative AI. This webinar was held on 30 March and 21 April with 1,465 and 1,079 respondents to the polling, respectively, for a total of 2,554 responses. Results of these polls should not be taken to represent all executives as the survey responses come from a population that had expressed interest in AI by attending a Gartner webinar on the subject.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Innovation Guide for Generative AI Technologies

Emerging Tech: Top Use Cases for Generative AI

Assessing How Generative AI Can Improve Developer

Experience How to Pilot Generative AI

Table 1: Capabilities, Descriptions and Features for Content Consumption

Capabilities ↓	*Descriptions* ↓	*Features* ↓
Conversational AI Interface	Conversational-based interface as a prompting or chatbot experience.	■ Natural language understanding ■ Voice support (for input and output consumption) ■ Multilingual support ■ Bot orchestration ■ Back-end service integration
Knowledge Activation	Activities that prepare enterprise knowledge bases for retrieval and support of generative synthesis.	■ Knowledge-base ingestion and storage ■ Data indexing and index management ■ Data enrichment: entities, topics, sentiment, embeddings transformation ■ Data staging: vector database support, knowledge graph building and expansion ■ Role-based access control (RBAC) access configuration

Capabilities ↓	*Descriptions* ↓	*Features* ↓
RAG Architecture	Hybrid architecture useful for grounding generative output by allowing information retrieval to inform the prompting of GenAI services.	■ Staging a search engine/capability with GenAI service in optimal network location ■ Configuration of search capability on top of activated knowledge base ■ Options for configuration and integrations between the two services (APIs and documentations) ■ Performance testing options ■ Observability features
Retrieval Optimization	To improve output relevance and optimize costs incurred in the GenAI application, retrieval optimization techniques can be facilitated through the search capability configuration.	■ Ability to configure semantic search capabilities (augmented by vector databases) ■ Ability to manage and customize search pipelines by search engineers ■ Infrastructure configurations for optimal performance ■ Observability

Source: Gartner (September 2023)

Table 2: Capabilities, Descriptions and Features for Content Creation

Capability	*Description*	*Features*
Conversational AI Interface (for Productivity Tools)	Conversational-based interface as a prompting or chatbot experience design in context of the task flow supported by the productivity tool.	■ Natural language understanding ■ Voice support ■ Multilingual support
Co-pilot Design Pattern	Co-pilots/assistants are designed as plug-ins into productivity tools and have a general pattern on deployment that offers versatility in use cases.	■ Integrate with conversational AI interface as a plug-in ■ Optimizes the data flow and compute needed for the generative tasks at hand ■ The plug-ins can be used to bridge the applications with more complex GenAI workflows and agents for multitask completion ■ User feedback loop fully integrated

Capability ↓	*Description* ↓	*Features* ↓
Prompt Engineering Experience	The prompt engineering experience is optimized in the context of the task, making use of templates, personalization and recommendations.	■ Multilanguage support ■ Prompting templates ■ Autocomplete recommendations included in prompting experience ■ Prompting help and suggestions (best practices integrated in the prompting experience)
Content Documentation and Contextualization	The ability to generate documentation and descriptions of tasks completed for productivity support in more complex business processes dealing with multiple artifacts, workflows and tasks.	■ User prompting for summarization of tasks ■ User prompting for documentation creation ■ User prompting for narratives and description generation for artifacts such as data tables, insights and shareable outputs.

Capability ↓	*Description* ↓	*Features* ↓
Metadata Activation and Personalization	Metadata becomes the input into generative capabilities that can personalize the experience based on past behavior and preferences.	■ Activation of metadata and user activity tracking throughout the capabilities ■ Personalize data elements used into prompting recommendations and generative narratives in the application ■ Contextualized narratives with tone and terminology specific to roles and functions
Actions/Agents	Ability for co-pilots/assistants to complete complex chains of tasks based on prompt-based goal definition and instructions.	■ Ability for co-pilots/assistants to triage prompt instructions to generative agents for complex task completion ■ Ability for agents to continuously learn about the preferences of users and optimize similar consecutive requests

Capability ↓	*Description* ↓	*Features* ↓
Decision Support	In applications that require analysis for decision support, besides the ability to present analytics to users, generative simulations can present variations of outcomes to support decision makers.	■ Ability for users to prompt the system for a simulation of key performance indicators (KPIs) with various constraints ■ Ability for users to ask for optimal outcome description of the simulation results to support the decision

Source: Gartner (September 2023)

Table 3: Capabilities, Descriptions and Features for Technology Creation

Capability	*Description*	*Features*
Model Ops for LLM Support	LLMops is the processes, methods and tools used for building, integrating and operating LLMs in production environments.	■ Infrastructure staging, including compute configurations and vector management facilities (vector libraries or databases) ■ Options for base model selection mapped to use cases and commercial considerations ■ Data preparation for training and fine-tuning (Q&A pairs, labeled data) ■ Prompt engineering and management ■ Evaluation, bias detection and responsible AI techniques and reports ■ Process for using reinforcement learning with human feedback (RLHF) ■ Ability to build LLM chains or pipelines ■ Documentation options ■ Deployment packaging and handoff

Capability	*Description*	*Features*
Synthetic Data	Synthetic data is needed when training data is not enough or is not representative of all variations of outcomes or occurrences.	■ Ability to build synthetic data using various generative methods based on usage scenarios needed ■ Use of various data labeling techniques for LLM refinement
Infrastructure Orchestration	The processes of architecture, build and maintenance for the supporting infrastructure of generative services in production.	■ Access (direct or via integration) to GPU clusters for training and inference ■ Orchestration of stateful and stateless services for best configuration needed for the use case ■ Options for managing a vector store or using a managed vector database ■ Observability across all pertinent services

Capability	*Description*	*Features*
Model API and Webhook Options	The ability to call/integrate with an external GenAI service via an API or webhook.	■ Provide or enable access to external specialized GenAI models (hosted or heterogeneous marketplaces) ■ Ability to call to an external (specialized) hosted generative service via API ■ Ability to use automated webhooks to different generative services ■ API life cycle management

Source: Gartner (September 2023)

Actionable, objective insight

Position your organization for success. Explore these additional complimentary resources and tools for tech product management leaders:

eBook

3 Lessons From High-Growth Companies to Build a Successful Product Strategy

Find out how to avoid common product strategy pitfalls.

Download Now

Report

Top Trends for Tech Providers for 2024

Explore the top trends product leaders must evaluate across all business dimensions.

Download Now

eBook

Leadership Vision for Technology Product Managers

Learn the top three strategic priorities for tech product managers.

Download Now

Tool

Gartner Product Decisions

Inform your product strategy and roadmap with this tool.

Learn More

Already a client?
Get access to even more resources in your client portal. Log In

Connect With Us

Get actionable, objective insight to deliver on your mission-critical priorities. Our expert guidance and tools enable faster, smarter decisions and stronger performance. Contact us to become a client:

U.S.: 1 844 466 7915

International: +44 (0) 3330 603 501

Become a Client

Learn more about Gartner for Product Teams
gartner.com/en/industries/high-tech/product-management-leaders

Stay connected to the latest insights

Attend a Gartner conference
View Conference

Gartner