HFS

When the lipstick no longer works….

Fixing our Legacy Debts to invest in our Future

HFS Summit, Cambridge University, September 26, 2024

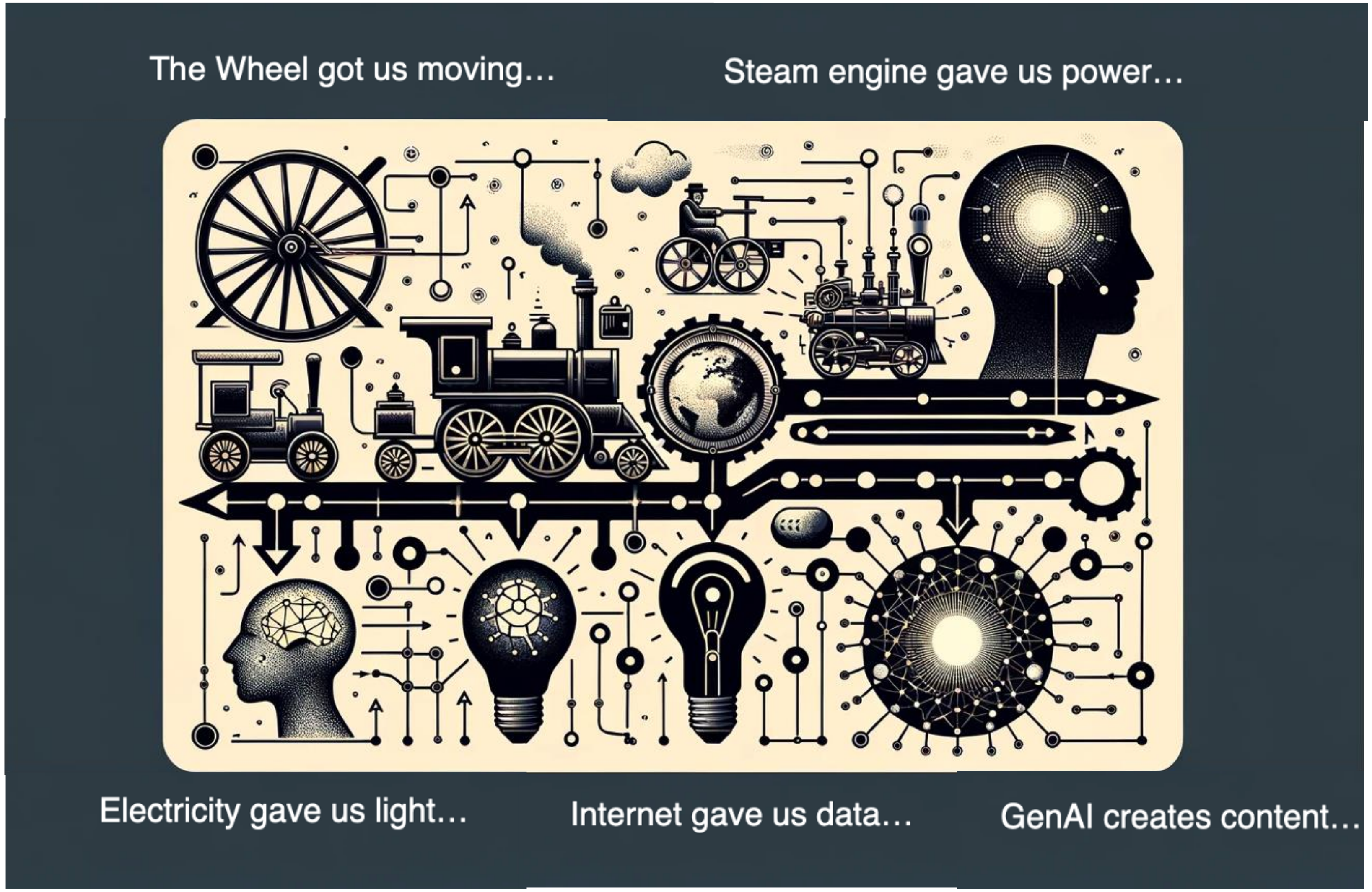
PHIL FERSHT

CEO and Chief Analyst, HFS Research

Five Seismic Human-made Disruptions

Disruption…

“Radical change to an existing industry or market due to technological innovation"



The Wheel got us moving…

Steam engine gave us power…

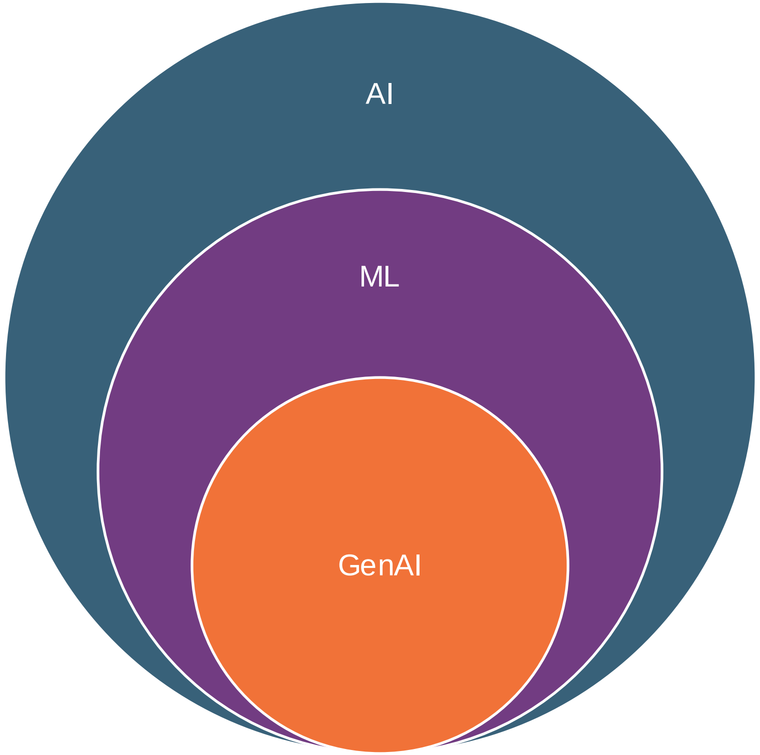
Electricity gave us light…

Internet gave us data…

GenAI creates content…

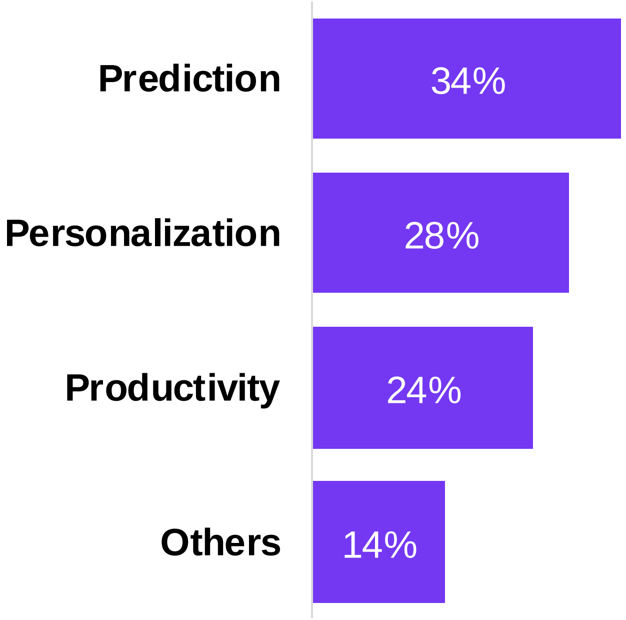
Source: HFS Research, 2024

2023 was about the WHAT

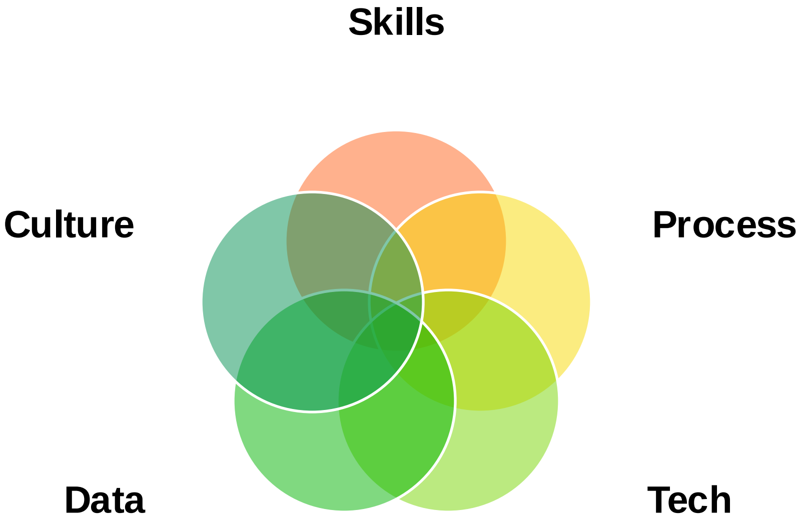


2024 is about the WHY

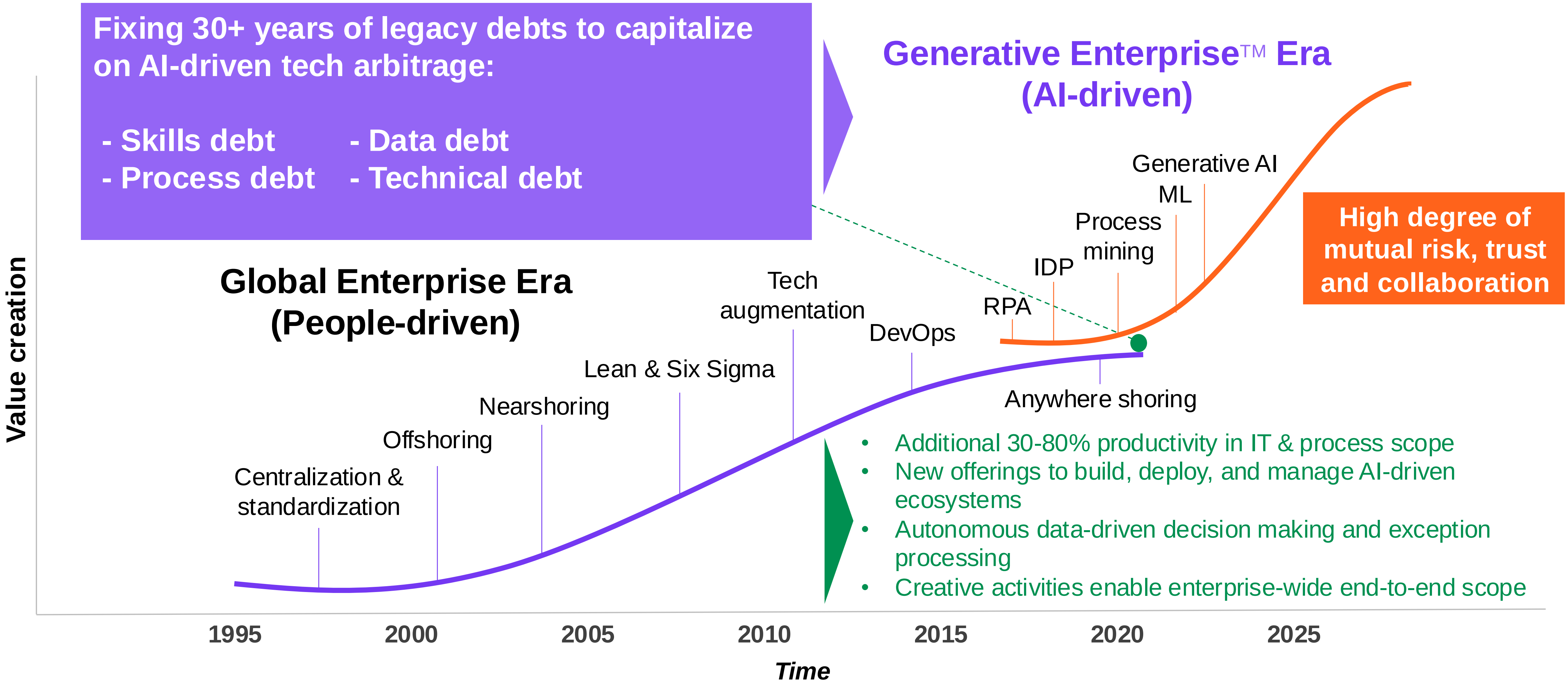
Enterprise GenAI use cases
% use cases



2025 will be about the HOW



The Generative Enterprise™ is driving The Great Services Transition



Fixing 30+ years of legacy debts to capitalize on AI-driven tech arbitrage:

- Skills debt
- Process debt
- Data debt
- Technical debt

Generative Enterprise™ Era (AI-driven)

Global Enterprise Era (People-driven)

High degree of mutual risk, trust and collaboration

- Additional 30-80% productivity in IT & process scope
- New offerings to build, deploy, and manage AI-driven ecosystems
- Autonomous data-driven decision making and exception processing
- Creative activities enable enterprise-wide end-to-end scope

The “Scale of Technology Partnerships” becomes critical in the Generative Era

Services Ecosystem Orchestration

accenture · Capgemini · IBM · Mphasis The Next Applied · Persistent · firstsource · SONATA SONATA SOFTWARE · LTIMindtree · Infosys Navigate your next · TECH mahindra · wipro · UST

cognizant · EY · KPMG · Deloitte. · AKKODiS · Hitachi Digital Services · publicis sapient · KPMG · pwc · tcs TATA CONSULTANCY SERVICES

Applications

Consumer uses

Entertainment	character.ai Midjourney
Productivity	OpenAI ChatGPT neeva
Other	trigo waabi

Enterprise stack

General productivity	ADEPT tome	glean AlphaSense
General and administrative	Ironclad eightfold.ai	synthesia Copilot
Sales and customer support	GONG Clari	RevComm PolyAI
Marketing	Jasper	WRITER
EPD, IT, security	Moveworks VECTRA	Abnormal GitHub Copilot

Industry verticals

Law firms	Harvey.
Creative	runway · Midjourney · imagen · descript
Health	viz.ai · BAYESIAN HEALTH · insitro · PathAI · UNLEARN
Defense	ANDURIL · Shield AI · SLINGSHOT AEROSPACE · VANNEVAR Labs
Agriculture and climate	Pachama · FarmWise
Construction	CANVAS

Enterprise apps

Adobe
PEGA
salesforce
SAP
servicenow
workday.

Infrastructure

Deploy and monitor

watsonx · Hugging Face · arize

Train and fine-tune models

Weights & Biases · mosaicML · PyTorch · watsonx · JAX

Open-source models & frameworks

Hugging Face · LLAMA · Stanford Alpaca · Gemini · GitHub

Full-stack large language models

OpenAI · ANTHROPIC · cohere · character.ai · Inflection

Store and compute

Label and process data	Data warehouses or lakehouses	Cloud service providers
Snorkel · scale · surge · COACTIVE	snowflake · databricks	Google Cloud · aws · Azure

Hardware

NVIDIA · AMD · intel

HFS Services and Ops Tech Vision 2030

Human | Machine

Staff augmentation

- Allows companies to quickly fill skill gaps, scale teams up or down as needed, and maintain control over project execution without the long-term commitments associated with permanent hires.
- **Key Features:**
  - **Flexibility:** Easily adjust team size based on project needs.
  - **Expertise:** Access specialized skills not available in-house.
  - **Control:** Maintain direct oversight of projects and processes.
- Typical commercial model: rate card

Technology-enabled services

- Primarily driven by people but supported by proprietary solution accelerators, tools, and software.
- Most service providers use this model to optimize processes and deliver value efficiently, such as Cognizant Neuro, Infosys Topaz, TCS WisdomNext & Wipro Lab45
- **Key Features:**
  - **Human-Centric:** Primarily driven by skilled professionals.
  - **Tool-Supported:** Utilizes a variety of technology tools and accelerators.
  - **Efficient:** Enhances service delivery through tech integration.
- Typical commercial model: FTE-based pricing

Platform-led services

- Leverage built-in delivery platforms to enhance service delivery and efficiency.
- Examples include Accenture SynOps, TCS Cognix, and Cognizant TriZetto, which streamline operations and provide consistent, scalable solutions.
- **Key Features:**
  - **Integrated Platforms:** Uses cohesive platforms for service delivery.
  - **Scalability:** Easily scalable and consistent across various operations.
  - **Efficiency:** Enhances productivity and efficiency through platform support.
- Typical commercial model: Transaction-based pricing

AI-led Agentic services

- Augmenting human capabilities with smart AI agents to optimize processes and decision-making.
- Examples of platforms include Amazon Q, GitHub, Lyzr, Copilot, Replit's Ghostwriter, Google Gemini, Einstein Agent, Mindcorp.
- Organizations like IBM and the Big 4 consulting firms are increasingly adopting this model.
- **Key Features:**
  - AI-Augmented: Combines human expertise with AI agents.
  - Cost-Effective: Achieves lower TCO through optimization.
  - Enhanced Capabilities: Expands service potential with AI-driven insights.
- Typical commercial model: Augmented FTE-based pricing or outcome-driven performance pricing

Service-as-a-Software

- Unlike traditional software-as-a-service (SaaS), this model focuses on delivering services primarily through technology, minimizing human intervention, and maximizing efficiency.
- Examples include startups like rhino.ai, Now Platform, and builder.ai
- **Key Features:**
  - Technology-driven: Primarily led by advanced software solutions.
  - Minimal Human Intervention: Reduces reliance on human resources.
  - Efficient and Scalable: Provides efficient, scalable, and consistent service delivery.
- Typical commercial model: License / Subscription-based

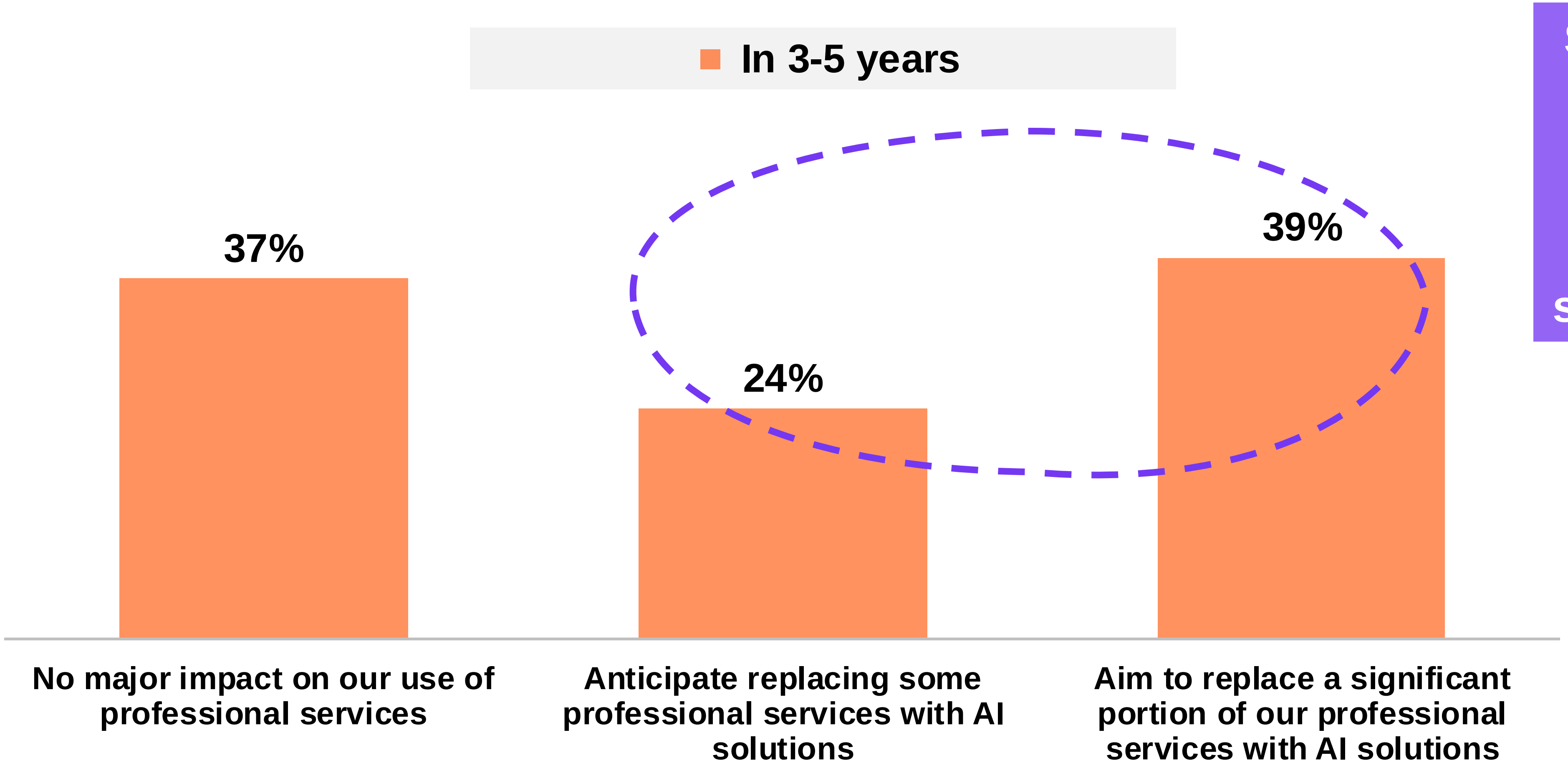
← Current → ← Emerging →

2000-2025 — *2025-2030*

HFS

Organizations are planning a phased implementation strategy to replace services with AI by 2030

Enterprises’ approaches to adopting AI to replace professional services



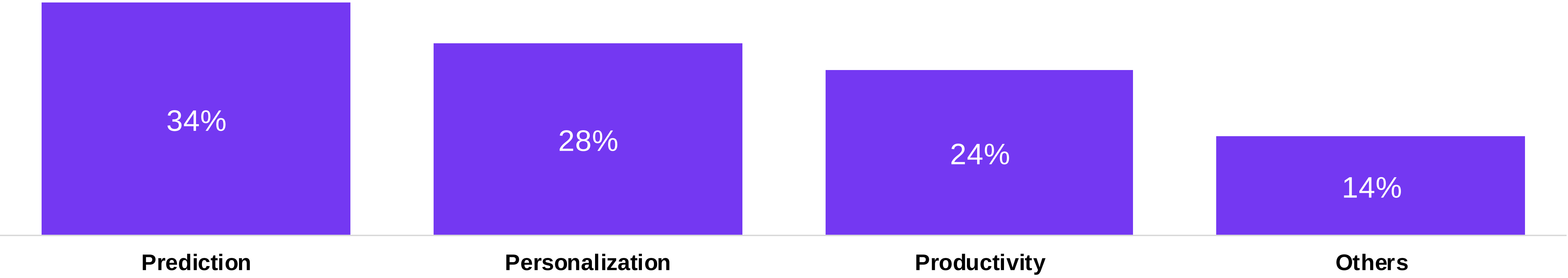
Six out of ten major enterprises plan to replace *people-run* services with *software-run* services before 2030

Sample: 1000 Major Global Enterprises
Source: HFS Pulse, 2024

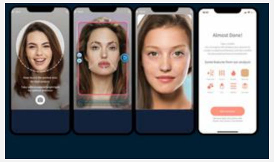
The Generative Era is not only about doing more with less… but also generating actionable business value

Enterprise GenAI use cases

% use cases

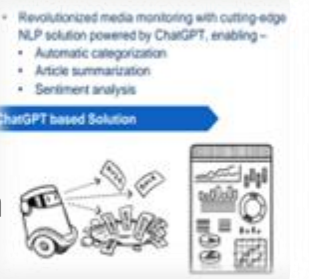


- Facial skin analysis to predict best treatments
- Soil analysis to predict best crop use



- Fast-food order entry, app and menu personalization
- CX in the medical sector.

- Accelerated compliance checking
- Media monitoring in healthcare products



- Revolutionized media monitoring with cutting-edge NLP solution powered by ChatGPT, enabling –
  - Automatic categorization
  - Article summarization
  - Sentiment analysis

ChatGPT based Solution

Sample: 104 enterprise leaders actively exploring and deploying GenAI
Source: HFS Research, November 2023

GenAI isn’t just ChatGPT… a plethora of LLMs are now available. Keeping pace is very challenging!

Most Popular / Widely Used

1.GPT-3 and GPT-4 (OpenAI) - Advanced AI capable of understanding and generating human-like text, widely used in various applications.

2.BERT (Google) - Revolutionized understanding of context in language, essential for improving search engines.

3.T5 (Google) - Converts all text-based language tasks into a unified text-to-text format, facilitating a wide range of NLP tasks.

4.RoBERTa (Facebook AI) - An optimized version of BERT, achieving state-of-the-art results on various NLP benchmarks.

5.ChatGPT (OpenAI) - Designed for conversational AI, providing responses that are contextually relevant.

6.Transformer-XL (Google/CMU) - Introduced a novel way to handle long-range dependencies in text.

7.XLNet (Google/CMU) - Combines the best of BERT and autoregressive models, handling permutation-based training.

8.DeBERTa (Microsoft) - Enhances the BERT and RoBERTa models with disentangled attention mechanism.

9.OPT (Meta/Facebook) - Open-sourced alternative to GPT models, designed for scalable language understanding.

10.LaMDA (Google) - Specializes in generating more sensible and specific responses in dialogues.

Highly Recognized

11.PaLM (Google) - Known for its performance in both language and multimodal tasks using Pathways, a scalable architecture.

12.ERNIE series (Baidu) - Focuses on enhancing model understanding by integrating knowledge graph data.

13.Megatron-Turing NLG (NVIDIA and Microsoft) - One of the largest language models aimed at natural language understanding.

14.LLaMA (Meta/Facebook) - Recognized for providing high-quality performance with fewer parameters, focusing on efficiency.

15.Gopher (DeepMind) - Known for its large-scale and broad coverage of diverse language understanding tasks.

16.CLIP (OpenAI) - Bridges the gap between visual and textual content, enabling models to understand images via text descriptions.

17.DALL-E (OpenAI) - Capable of generating novel images from textual descriptions, demonstrating creativity in AI.

18.Chinchilla (DeepMind) - Optimized for training efficiency by using more data but fewer parameters.

19.Jurassic-1 (AI21 Labs) - Designed to handle a wide variety of language tasks, known for its versatility.

20.BlenderBot (Facebook AI) - Designed for building more engaging and natural long-term conversations.

Notable for Technical Innovation or Niche Applications

21.Wu Dao 2.0 (Beijing Academy of Artificial Intelligence) - Multimodal model capable of understanding both text and images.

22.MoE (Google) - Uses a "Mixture of Experts" to scale efficiently to trillions of parameters.

23.Pangu (Huawei) - A large model focused on Chinese language processing.

24.Anthropic AI's Claude - Designed with a focus on safety and reliability in AI systems.

25.BigBird (Google) - Handles long sequences of data, making it suitable for tasks like document summarization.

26.Switch Transformers (Google) - Introduces a technique to train very large models efficiently by using sparser models.

27.Performer (Google) - Provides an efficient way to scale attention mechanisms in Transformers.

28.Reformer (Google) - Known for processing long sequences using less memory, making it efficient for large datasets.

29.ByT5 (Google) - Treats every input as bytes, simplifying the processing of multilingual text.

30.CTRL (Salesforce) - A conditional language model that can control style, content, and task-specific behavior.

Emerging or Specialized Usage

31.ALBERT (Google) - A version of BERT optimized for lower memory consumption and increased speed.

32.DialoGPT (Microsoft) - Tailored for generating dialogues, simulating conversational exchanges.

33.Codex (OpenAI) - Geared towards understanding and generating programming code, powering tools like GitHub Copilot.

34.ERNIE 3.0 Titan (Baidu) - An iteration that further integrates knowledge graphs for improved semantic understanding.

35.Whisper (OpenAI) - Specialized for speech-to-text tasks, featuring robust performance across languages.

36.Dragon (Baidu) - Focuses on high-performance across various NLP tasks, heavily used in Chinese-language applications.

37.Lucy (OpenAI) - Emphasizes safety and ethical considerations in AI deployment.

38.M6 (Tencent) - A multimodal model designed for diverse applications, including text and image understanding.

39.Z-code (Meta) - Focuses on programming and technical tasks, assisting developers.

40.Alexa Teacher Model (Amazon) - Enhances Alexa's interactions, improving its conversational capabilities.

Additional Models with Specific Contributions

41.Retro (Google) - Incorporates retrieval capabilities into the model, enhancing information access.

42.Music Transformer (Google) - Generates music with long-term coherence.

43.Luminous Base (Luminous AI) - General-purpose model aimed at a wide range of applications.

44.Hive-Cote (Community) - Combines multiple machine learning models for improved predictive accuracy.

45.DeepSpeed (Microsoft) - A library designed to accelerate the training of large-scale models.

46.AdaGram (OpenAI) - Adjusts word meanings based on their use context.

47.Sparse Transformer (OpenAI) - Implements sparsity to scale to larger contexts efficiently.

48.MT-NLG (Microsoft) - Known for its capabilities in natural language generation, particularly in generating coherent long texts.

49.BART (Facebook AI) - Blends the benefits of pre-trained autoregressive models and autoencoders, enhancing both generation and comprehension tasks.

50.Leviathan (Facebook AI) - Designed for complex reasoning and knowledge-intensive tasks, pushing the limits of what AI can understand.

Not Exhaustive

I’ve been informed that GenAI is “Lipstick Looking for a Pig”

Enterprise GenAI use-cases show potential

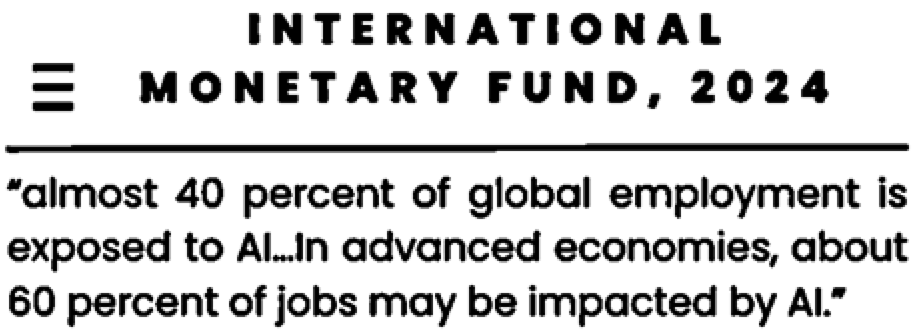
Competitive Advantage

- **Accelerated Medical Diagnostics**: AI in diagnostics streamlines medical assessments, offering a competitive edge with faster and accurate diagnoses.
- **Proactive Equipment Maintenance**: AI analysis enables proactive maintenance, minimizing downtime and maintaining operational efficiency.
- **Product Customization**: Gen AI allows for product customization, meeting diverse market demands and setting the company apart with unique offerings.
- **Personalization in Offerings**: Offering personalized products and services using generative AI gives a competitive edge by meeting individual customer preferences.
- **Predictive Trends Analysis**: Utilizing AI for predictive trend analysis allows companies to stay ahead of market shifts and anticipate consumer needs.
- **Streamlined Supply Chain Operations**: AI optimization of supply chain operations reduces costs and improves efficiency, providing a competitive advantage in operational excellence.
- **Expediting R&D Processes**: GenAI expedites research and development processes, enabling quicker product development and innovation.
- **Personalized Healthcare Solutions**: Providing treatments based on individual genetic profiles offers a competitive edge through personalized healthcare solutions.
- **Market Trends Monitoring**: Employing GenAI to continuously monitor and analyze market trends enables rapid adaptation and staying ahead of competitors.
- **Rapid Prototyping and Testing**: AI-driven rapid prototyping and testing accelerate innovation and reduce time-to-market for new products.

Faster revenue growth and increased market share

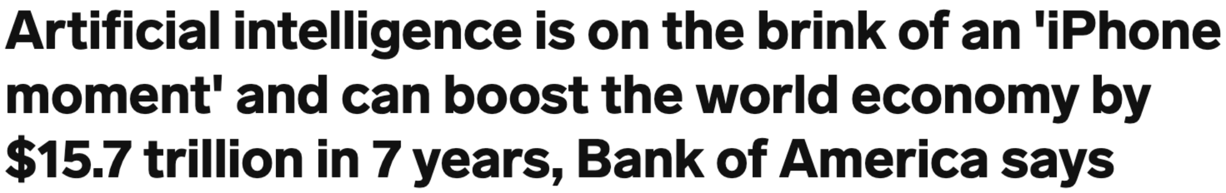
- **Utilizing GenAI for Market Analysis**: Analyze market trends, consumer behavior, and competitive pricing to identify opportunities.
- **Continuous Monitoring of Competitor Activities**: Keep track of pricing, promotions, and product launches to adjust strategies promptly.
- **Optimizing Prices in Real-time**: Adjust pricing strategies dynamically to maximize revenue and capture market share.
- **Streamlining Clinical Trials**: Accelerate research timelines for faster revenue growth.
- **Efficient Lead Qualification**: Streamline the process of identifying and qualifying leads.
- **Enhanced Customer Experience**: Implement GenAI in customer service for personalized and efficient support.
- **Data-driven Pricing Strategies**: Analyze customer purchase histories and preferences for optimizing profitability.
- **Predicting Demand Accurately**: Reduce inventory costs and ensure timely delivery by forecasting demand.
- **Precision Marketing Strategies**: Enhance market share through personalized engagement.
- **Amplifying Production Output**: Use Generative AI to optimize production efficiency and meet market demands.

Fueling the Hype…

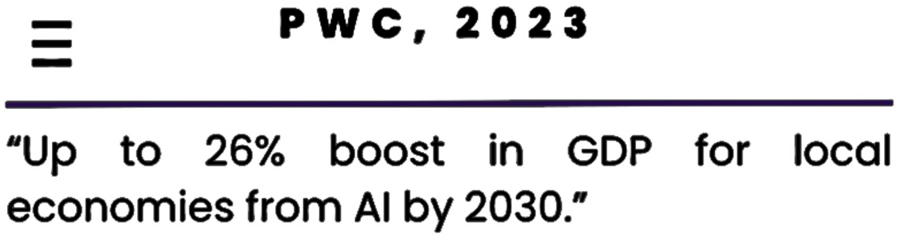


INTERNATIONAL MONETARY FUND, 2024

"almost 40 percent of global employment is exposed to AI…In advanced economies, about 60 percent of jobs may be impacted by AI."

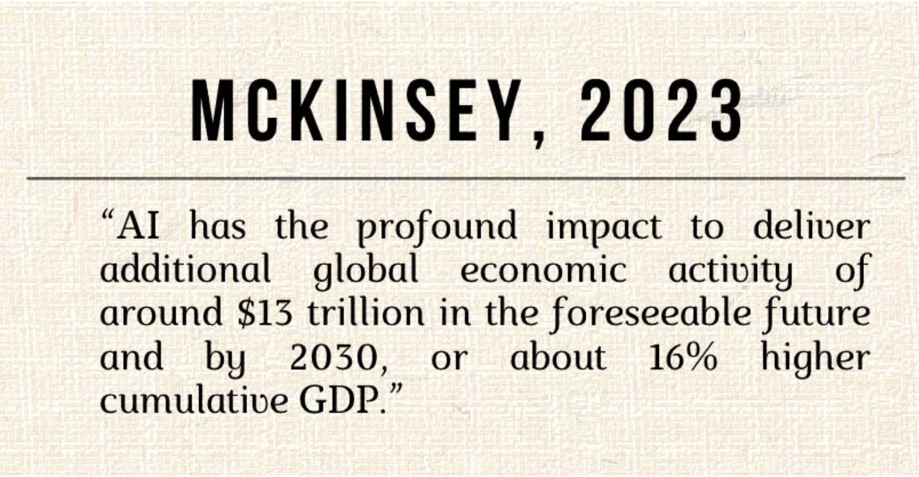


Artificial intelligence is on the brink of an 'iPhone moment' and can boost the world economy by $15.7 trillion in 7 years, Bank of America says



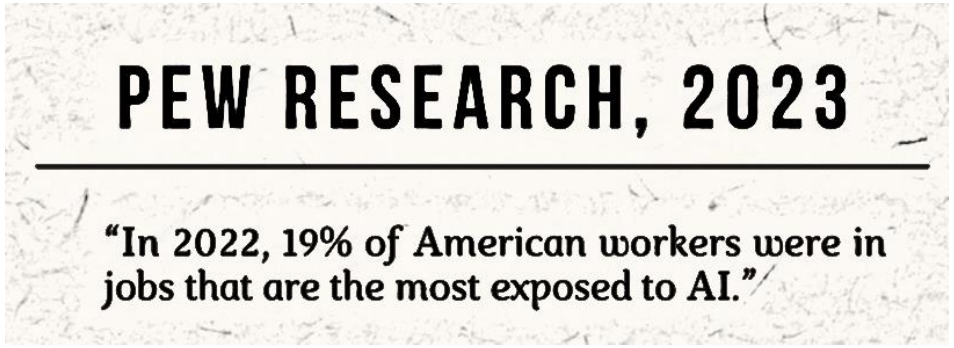
PWC, 2023

"Up to 26% boost in GDP for local economies from AI by 2030."



MCKINSEY, 2023

"AI has the profound impact to deliver additional global economic activity of around $13 trillion in the foreseeable future and by 2030, or about 16% higher cumulative GDP."

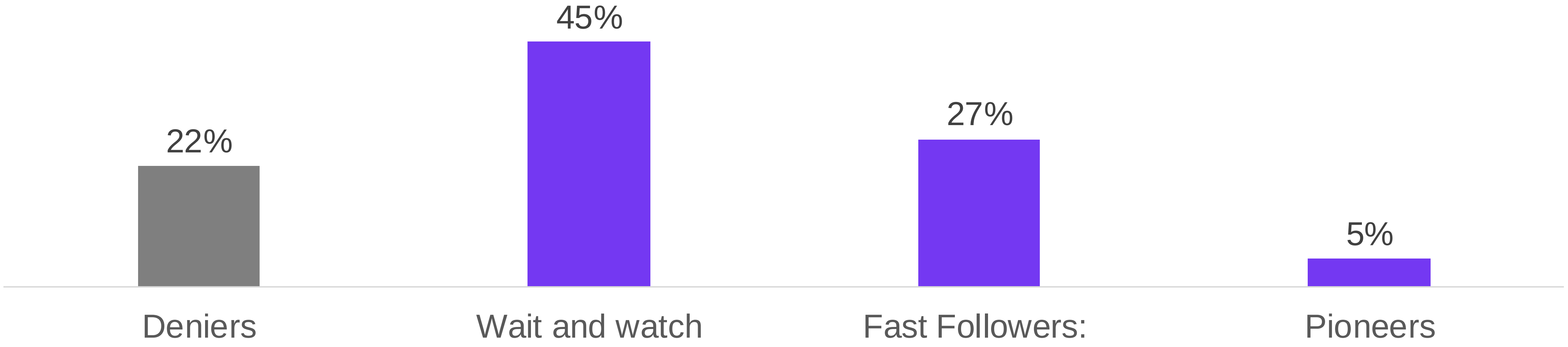


PEW RESEARCH, 2023

"In 2022, 19% of American workers were in jobs that are the most exposed to AI."

Reality-check: GenAI is a smoldering platform, not a burning one

Only **5%** of enterprises have committed significant technology spend on GenAI and successfully deployed GenAI solutions across multiple parts of their business



No GenAI plans or spending

No significant commitment to spend on GenAI and mainly in planning stages

Already committing significant tech spend to GenAI or deploying GenAI solutions across multiple parts of their business

Already deployed GenAI solutions across multiple parts of the business and committing significant technology spend to GenAI

Sample: 550 Enterprise Leaders
Source: HFS Research, 2024

Can Enterprise Middle Management Withstand the C-Suite's Impatience?

C-Suites are bullish on GenAI's benefits, but they are impatiently witnessing their organizations "Wait and Watch"

Enterprise Middle Management is strangled by:

1) Everyone jockeying to show some implementation in GenAI without governance, limiting ability to focus investment on what matters.
2) Traditional funding being tapped out by high fixed costs of legacy IT app development, maintenance, and infrastructure.
3) Slow delivery of meaningful GenAI functionality by many legacy technology software providers, so the business is cobbling together custom implementations
4) Internal skill gaps – its all so new.

The current response is:

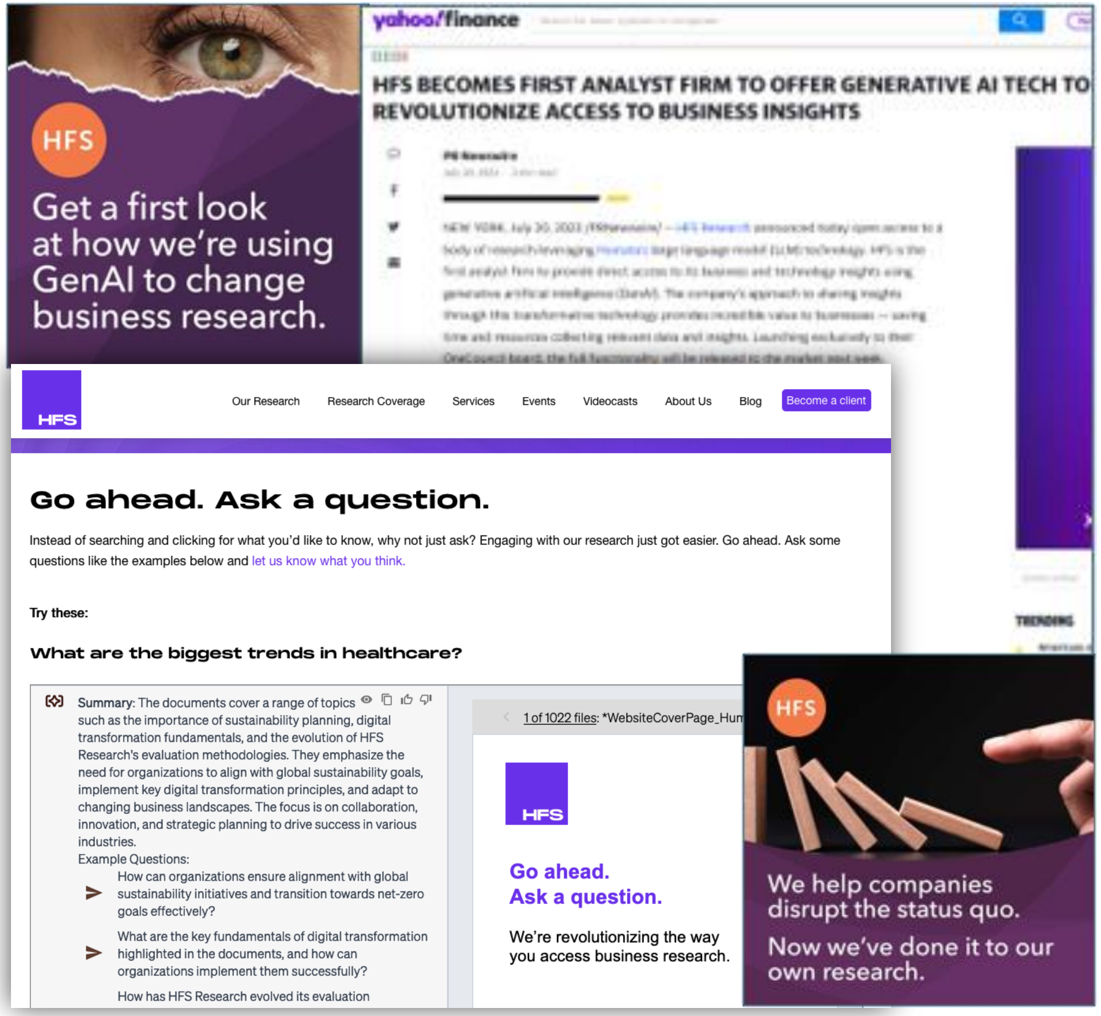
1) Middle management is funding GenAI from leftover change in their budgets.
2) GenAI subjected to "death by a thousand pilots"
3) Few meaningful large scale impacts
4) Enterprises can't get past data privacy rules/controls

HFS is All-in on GenAI:
Research in the Context of YOU!

Impact of adopting GenAI as a game changer:

- **First to market** – launched in June 2023
- HFS has uploaded over **1,300 pieces of unique research**
- Since January 2024 we have had **over 110,000 questions** asked *and answered* by our GenAI tool
- No more document-based searching, our insights now come in the **context of your unique needs**.

> ***"Humata and HFS have created a game changing experience that delivers contextual answers via a secure, private genAI and eclipsing traditional search."***
>
> *– Cyrus Khajvandi, CEO Humata.ai*


HFS
Get a first look at how we're using GenAI to change business research.
yahoo!/finance

HFS BECOMES FIRST ANALYST FIRM TO OFFER GENERATIVE AI TECH TO REVOLUTIONIZE ACCESS TO BUSINESS INSIGHTS

PR Newswire

NEW YORK, July 20, 2023 /PRNewswire/ -- HFS Research announced today open access to a body of research leveraging Humata's large language model (LLM) technology. HFS is the first analyst firm to provide direct access to its business and technology insights using generative artificial intelligence (GenAI). The company's approach to sharing insights through this transformative technology provides incredible value to businesses -- saving time and resources collecting relevant data and insights. Launching exclusively to their OneCouncil board, the full functionality will be released to the market next week.

Go ahead. Ask a question.

Instead of searching and clicking for what you'd like to know, why not just ask? Engaging with our research just got easier. Go ahead. Ask some questions like the examples below and let us know what you think.

Try these:

What are the biggest trends in healthcare?

Summary: The documents cover a range of topics such as the importance of sustainability planning, digital transformation fundamentals, and the evolution of HFS Research's evaluation methodologies. They emphasize the need for organizations to align with global sustainability goals, implement key digital transformation principles, and adapt to changing business landscapes. The focus is on collaboration, innovation, and strategic planning to drive success in various industries.

Example Questions:

- How can organizations ensure alignment with global sustainability initiatives and transition towards net-zero goals effectively?
- What are the key fundamentals of digital transformation highlighted in the documents, and how can organizations implement them successfully?
- How has HFS Research evolved its evaluation

1 of 1022 files: *WebsiteCoverPage_Hum

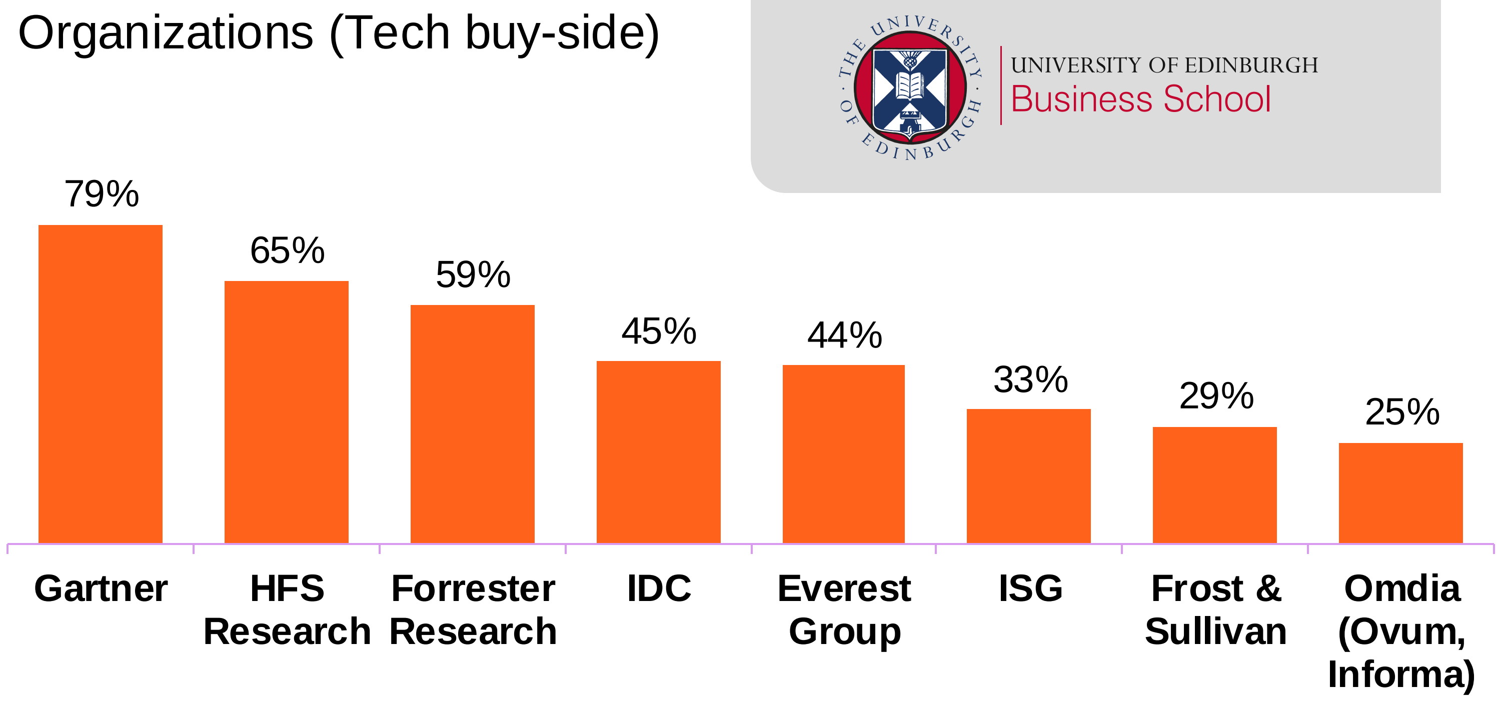
HFS
Go ahead.
Ask a question.
We're revolutionizing the way you access business research.
HFS
We help companies disrupt the status quo.
Now we've done it to our own research.

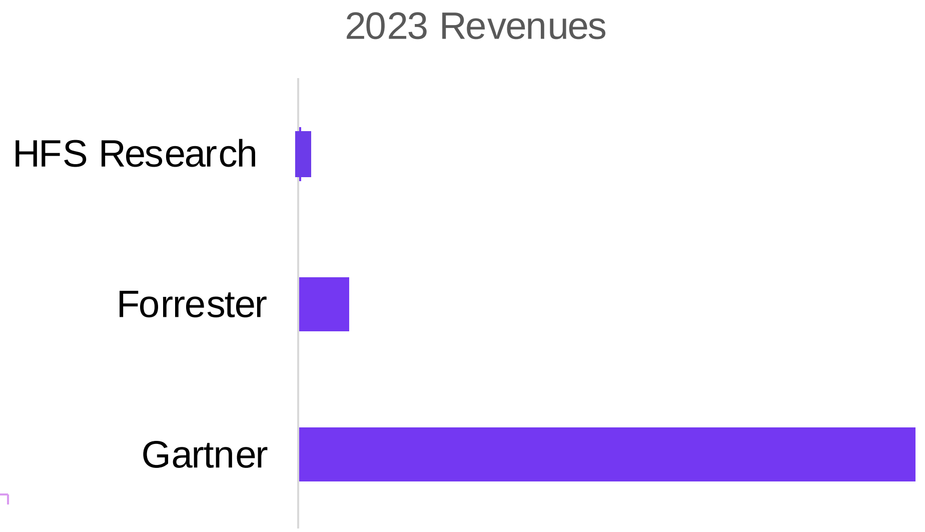
https://www.hfsresearch.com/research/trygenai/

HFS disrupts the analyst industry with our GenAI-based research platform

Analyst value survey

UNIVERSITY OF EDINBURGH
Business School





Sample: Demand-side organizations that responded to the Analyst Attitude Survey conducted by the Analyst Observatory at the University of Edinburgh.

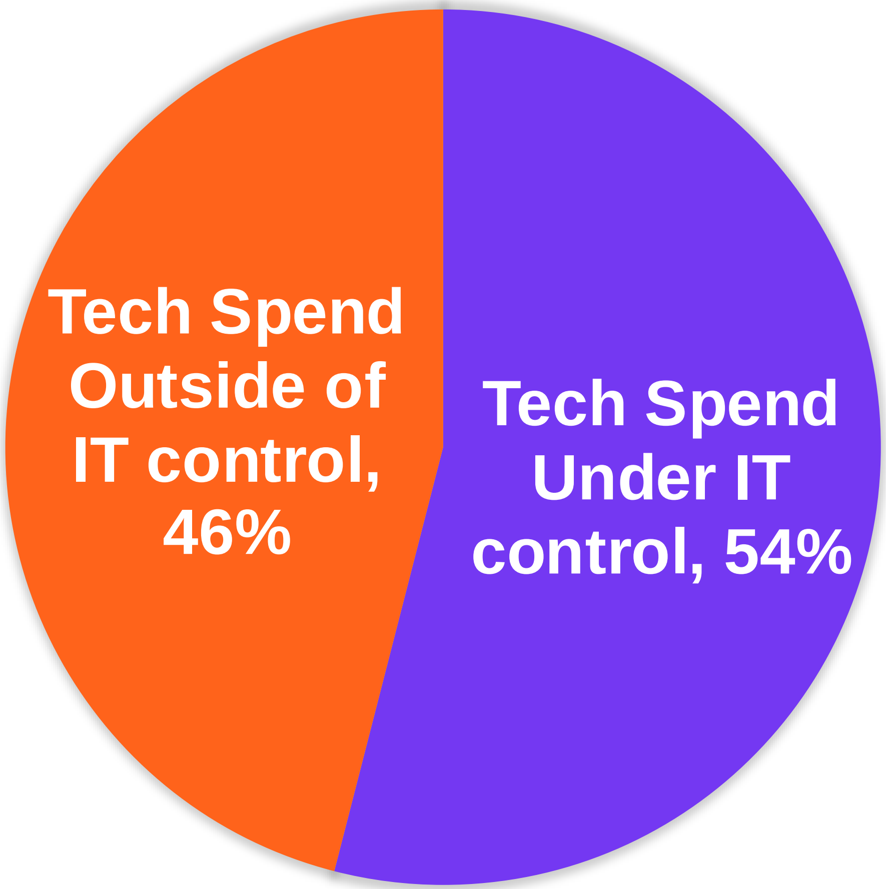
GPT-4o with Eyes

Can GPT-4o pour fuel onto the GenAI smoldering platform?

1. **Multimodal makes everything *much* more human. Text, vision and speech into the same neural network…**Old GPT was like texting a friend, GPT-4o is like calling a friend
2. **Real-time human-2-machine conversation is now possible.** converse *naturally* without first converting words to text, with real energy, emotion, and expressiveness
3. **Enhanced multilingual support and capabilities**
4. **It really does have human eyes now**
5. **It's being incorporated into Apple's iPhone and Google's Android operating systems.**
6. **Summaries are concise and relevant.** finally the end of legacy Google search and bad call transcripts?
7. **Visual interpretation and data tables are much more usable and accurate, ready to support business needs.** It accurately converts image data into a clean table format without misinterpretations
8. **Image generation capabilities are just so much sharper**.
9. **The cost of accessing its APIs is 50% cheaper**. such as Chat Completions API, Assistants API, and Batch API
10. **Coding is vastly improved.** solve many coding projects, such as multiple thousand lines of code in under 10 minutes, which previously took prompt engineering processes many hours

Winners in the Generative Era will require strong business acumen

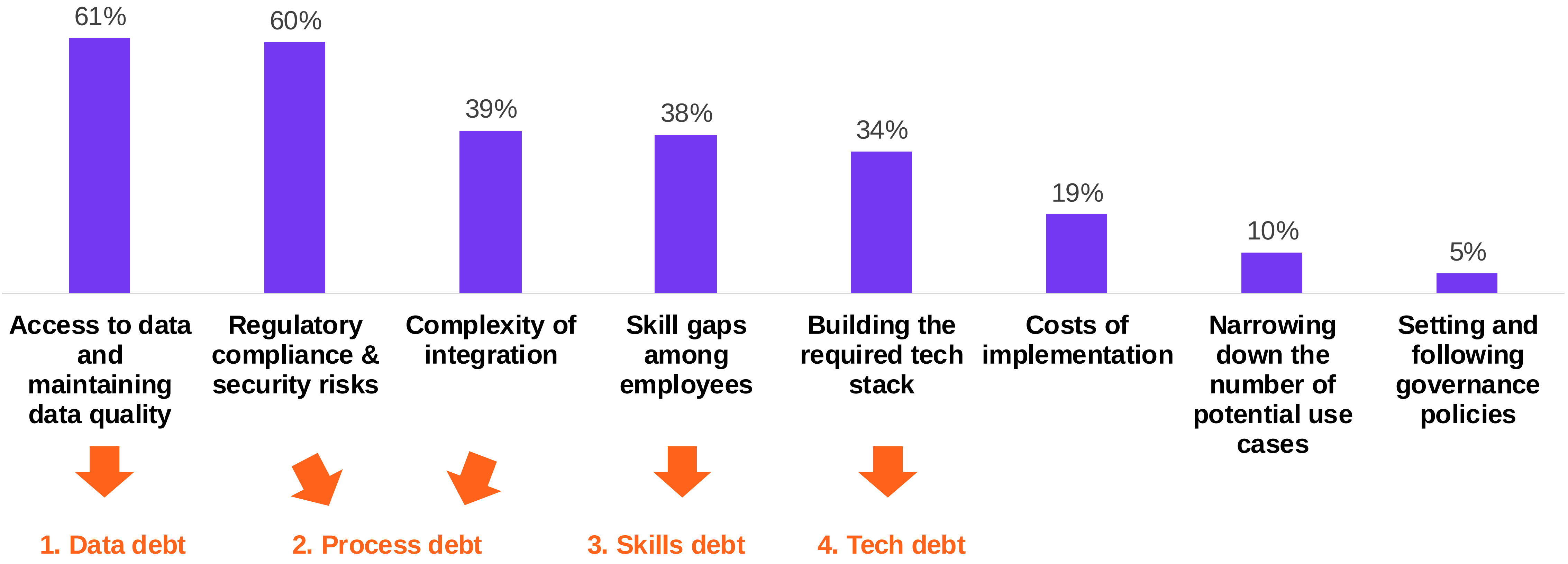
What percentage of your enterprise's technology-related spending is controlled by IT?



Sample: 551 Global 2000 enterprise executives
Source: HFS Pulse, 2023-24

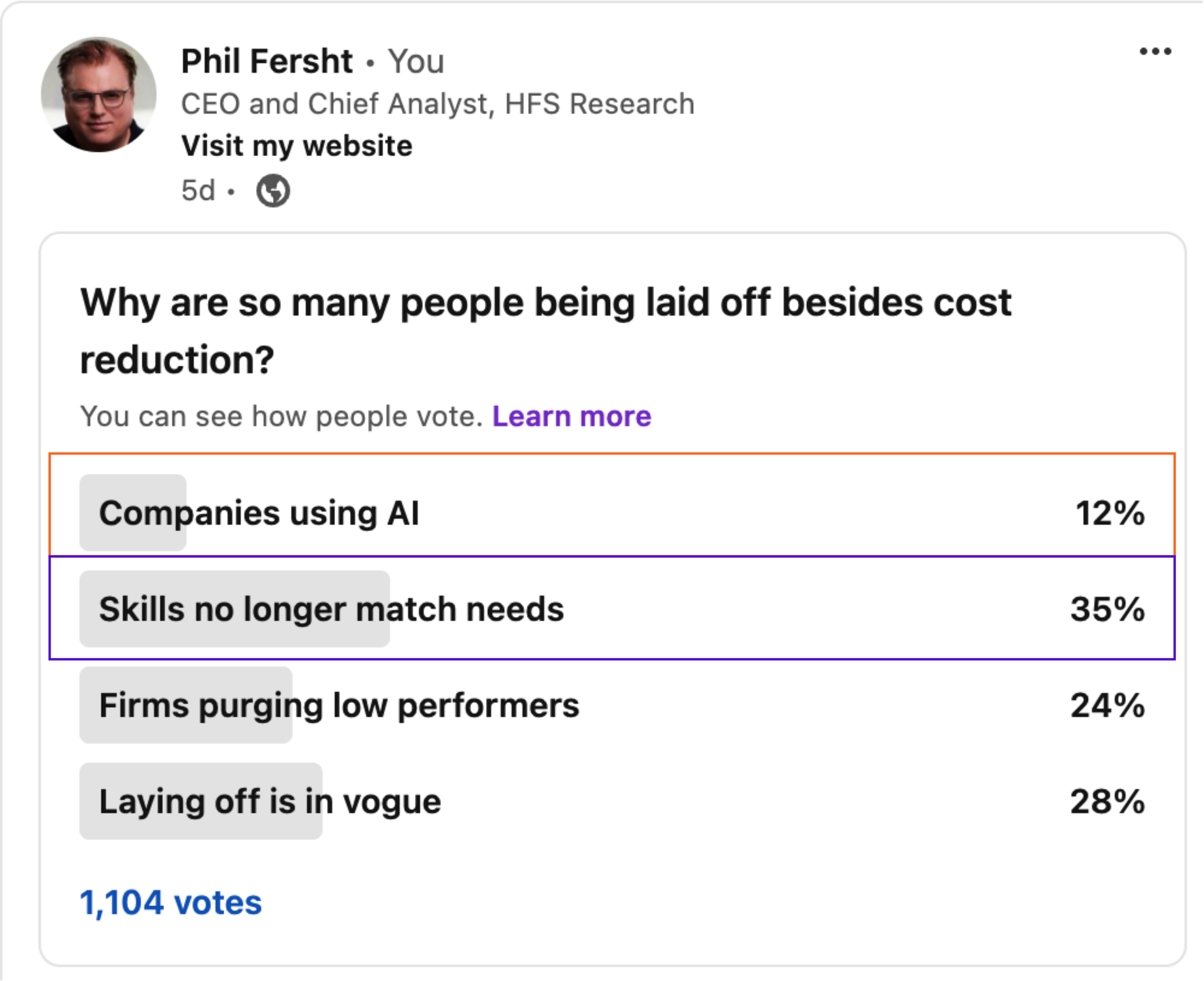
We need to pay our debts before we can become a Generative Enterprise

Which are the most significant challenges in implementing GenAI in your organization?



Sample: 550 Enterprise Leaders
Source: HFS Research, 2024

AI isn't replacing jobs… but may get replaced by someone who understands AI



Phil Fersht • You
CEO and Chief Analyst, HFS Research
Visit my website
5d •

Why are so many people being laid off besides cost reduction?

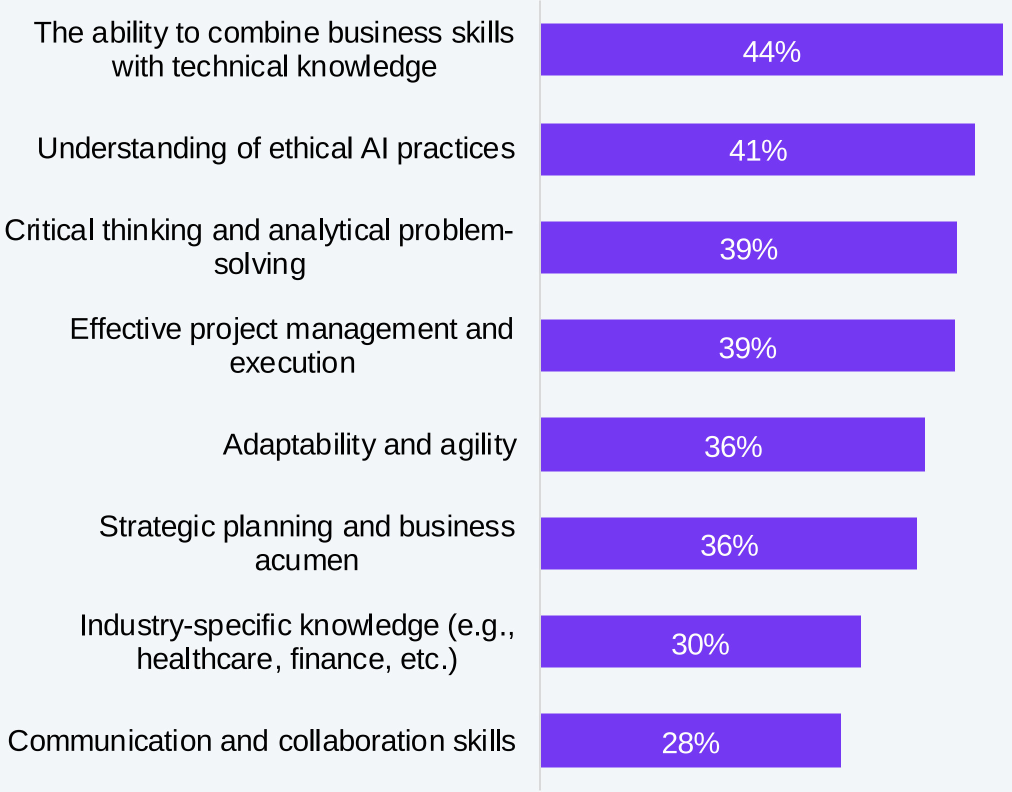
You can see how people vote. **Learn more**

Companies using AI	12%
Skills no longer match needs	35%
Firms purging low performers	24%
Laying off is in vogue	28%

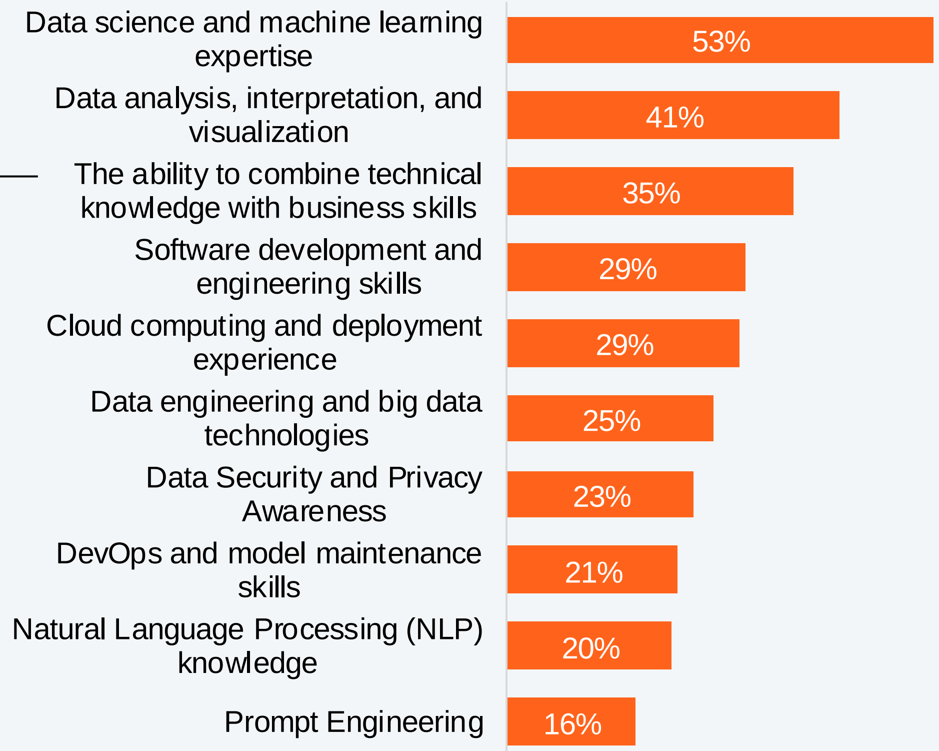
1,104 votes

The ability to combine business and technical skills will be critical to succeed in the AI-led era

Business skills needed to drive successful GenAI initiatives
N =550



Technical skills needed to drive successful GenAI initiatives
N =550

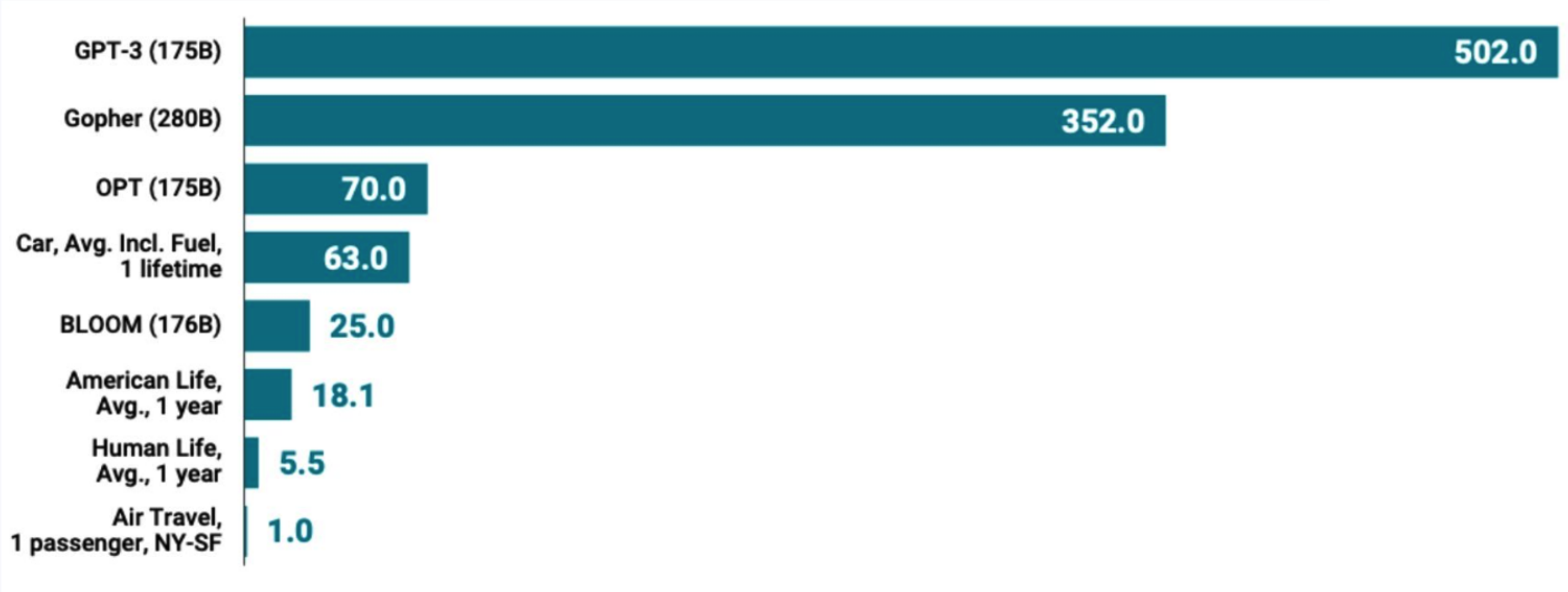


Q: What are the top three business skills you believe most crucial to drive the success of your GenAI initiatives?

Sample: 550 enterprise leaders
Source: HFS Research, 2024

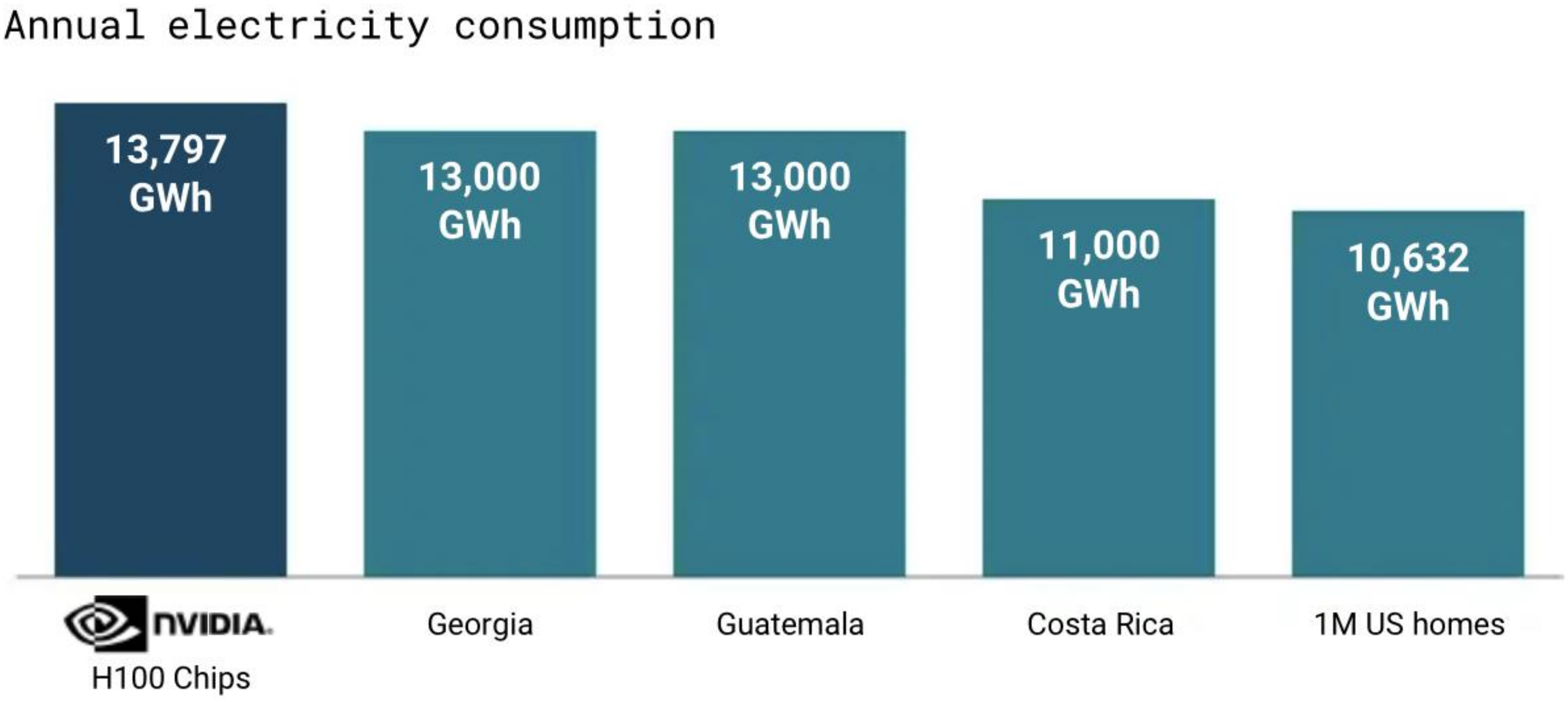
Transformer-based models produce a colossal carbon footprint

CO2 equivalent emissions (tonnes) by selected machine learning models and real-life



- Source: Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model (Luccioni et al.)

Nvidia’s high-end chips will consume the same amount of energy as a small nation in 2024



So where do we go from here?

AI is a smouldering platform, not a burning one, but the key now is to scale what we have, and to improve what we have *before* we add more.

Our industry has obsessed with shiny new tech and not real business change. This has to change

We must fix our past debts to invest in the future

It takes both enterprises and their partners to learn and work together

AI won’t replace people, but people who don’t work with AI are at risk

We are barely more than 18 months in, so let’s all take stock, there is time to adjust, learn and adapt