Steve Nouri

@SteveNouri

What is Generative AI? It is a multi-billion dollar opportunity!

$4.4T

The potential productivity value gen AI could add annually across 63 use cases

McKinsey & Company

-------------------------->>>>>>>>>>>>>>

1- Introduction to LLMs and Generative AI:

- The Intersection of LLMs and Generative AI
  - Both are under the umbrella of deep learning
  - Generative AI produces new content: text, images, audio
- Large Language Models:
  - Pre-trained for general purposes
  - Fine-tuned for specific tasks

2- Overview of PaLM and Evolution of AI Models:

- PaLM: Pathways Language Model
  - 540 billion parameters
  - Dense decoder-only transformer model
- From Traditional Programming to Generative Models:
  - Traditional: Hard code rules (e.g., distinguishing a cat)
  - Neural Networks: Image recognition (e.g., cat or dog)
  - Generative: Users generate content (e.g., text, images)
-

Generative AI's evolution

For an advanced technology that's considered relatively new, generative AI is deep-rooted in history and innovation.

1932

Georges Artsrouni invents a machine he reportedly called the **"mechanical brain"** to translate between languages on a mechanical computer encoded onto punch cards.

1966

MIT professor Joseph Weizenbaum creates the first chatbot, **Eliza**, which simulates conversations with a psychotherapist.

1980

Michael Toy and Glenn Wichman develop the Unix-based game *Rogue*, which uses procedural content generation to dynamically generate new game levels.

1986

Michael Irwin Jordan lays the foundation for the modern use of recurrent neural networks (RNNs) with the publication of "Serial order: a parallel distributed processing approach."

2000

University of Montreal researchers publish "A Neural Probabilistic Language Model," which suggests a method to model language using feed-forward neural networks.

2011

Apple releases **Siri**, a voice-powered personal assistant that can generate responses and take actions in response to voice requests.

2013

Google researcher Tomas Mikolov and colleagues introduce word2vec to identify semantic relationships between words automatically.

2015

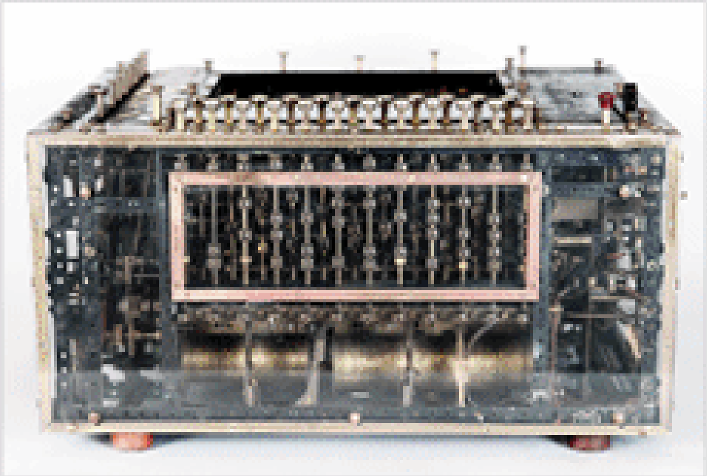
Stanford researchers publish work on diffusion models in the paper **"Deep Unsupervised Learning using Nonequilibrium Thermodynamics."** The technique provides a way to reverse-engineer the process of adding noise to a final image.

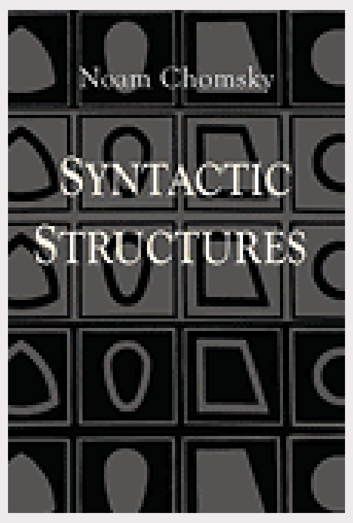
2018

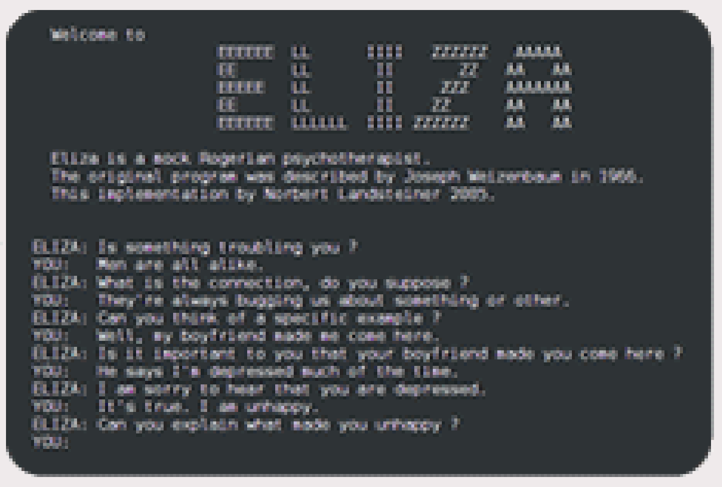
Google researchers implement transformers into BERT, which is trained on more than 3.3 billion words and can automatically learn the relationship between words in sentences, paragraphs and even books to predict the meaning of text. It has 110 million parameters.

Google DeepMind researchers develop AlphaFold for predicting protein structures, laying the foundation for generative AI applications in medical research, drug development and chemistry.

OpenAI releases GPT (Generative Pre-trained Transformer). Trained on about 40 gigabytes of data and consisting of 117 million parameters, GPT paves the way for subsequent LLMs in content generation, chatbots and language translation.



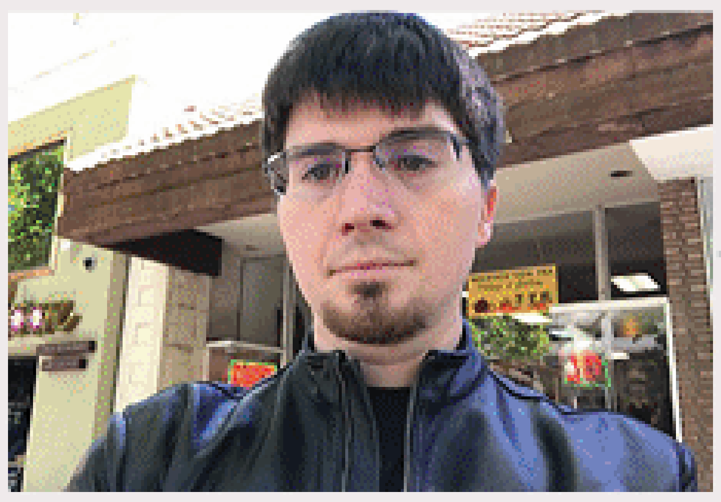


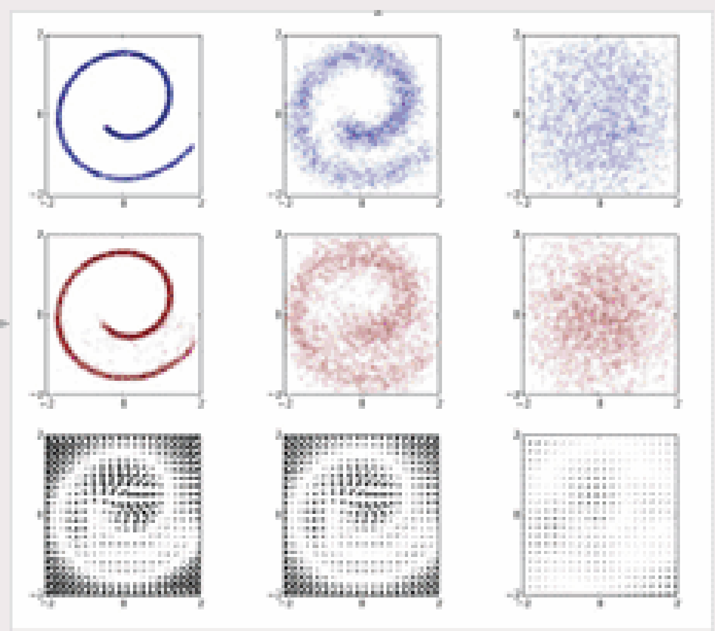


```
Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII  ZZZZZZ   AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```





1957

Linguist **Noam Chomsky** publishes ***Syntactic Structures***, which describes grammatical rules for parsing and generating natural language sentences.

1968

Computer science professor Terry Winograd creates SHRDLU, the first multimodal AI that can manipulate and reason out a world of blocks according to instructions from a user.

1985

Computer scientist and philosopher Judea Pearl introduces Bayesian networks causal analysis, which provides statistical techniques for representing uncertainty that leads to methods for generating content in a specific style, tone or length.

1989

Yann LeCun, Yoshua Bengio and Patrick Haffner demonstrate how convolutional neural networks (CNNs) can be used to recognize images.

2006

Data scientist Fei-Fei Li sets up the ImageNet database, which provides the foundation for visual object recognition.

2012

Alex Krizhevsky designs the AlexNet CNN architecture, pioneering a new way of automatically training neural networks that take advantage of recent GPU advances.

2014

Research scientist **Ian Goodfellow** develops generative adversarial networks (GANs), which pit two neural networks against each other to generate increasingly realistic content.

Diederik Kingma and Max Welling introduce variational autoencoders to generate images, videos and text.

2017

Google researchers develop the concept of transformers in the seminal paper "Attention is all you need," inspiring subsequent research into tools that could automatically parse unlabeled text into large language models (LLMs).

2021

OpenAI introduces **Dall-E**, which can generate images from text prompts. The name is a combination of WALL-E, the name of a fictional robot, and the artist Salvador Dali.

2022

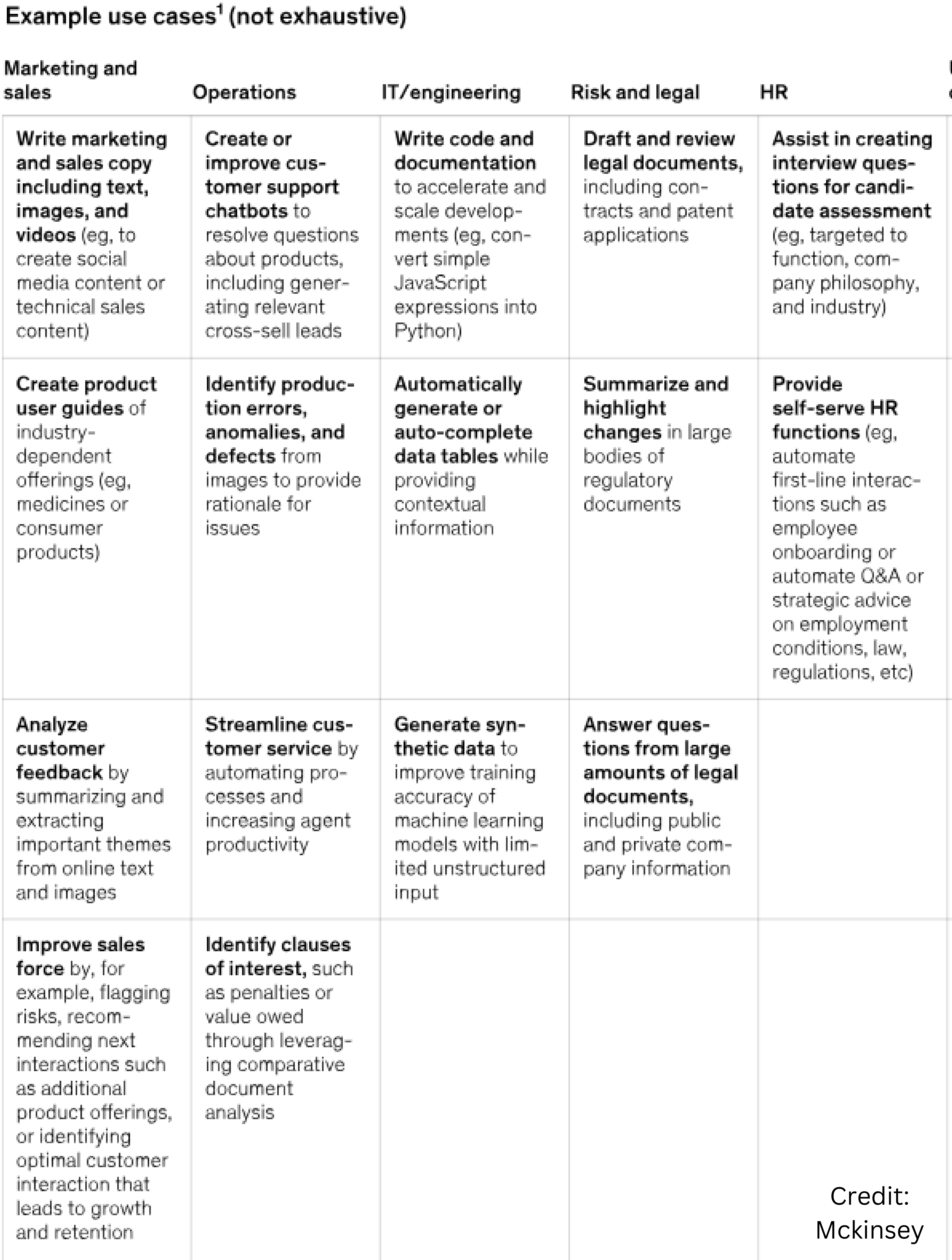
Researchers from Runway Research, Stability AI and CompVis LMU release Stable Diffusion as open source code that can automatically generate image content from a text prompt.

OpenAI releases **ChatGPT** in November to provide a chat-based interface to its GPT 3.5 LLM. It attracts over 100 million users within two months, representing the fastest ever consumer adoption of a service.

Credit:
TechTarget

3- Features & Benefits of LLMs:

- Major features:
    - Large: Big training datasets & parameter count
    - General purpose: Solve common problems
    - Pre-trained and fine-tuned: Versatility for specific needs
- Benefits:
    - Multiple tasks with one model (translation, text classification, etc.)
    - Minimal domain training required
    - Better performance with more data & parameters



Example use cases[1] (not exhaustive)

Marketing and sales	Operations	IT/engineering	Risk and legal	HR
Write marketing and sales copy including text, images, and videos (eg, to create social media content or technical sales content)	**Create or improve customer support chatbots** to resolve questions about products, including generating relevant cross-sell leads	**Write code and documentation** to accelerate and scale developments (eg, convert simple JavaScript expressions into Python)	**Draft and review legal documents,** including contracts and patent applications	**Assist in creating interview questions for candidate assessment** (eg, targeted to function, company philosophy, and industry)
Create product user guides of industry-dependent offerings (eg, medicines or consumer products)	**Identify production errors, anomalies, and defects** from images to provide rationale for issues	**Automatically generate or auto-complete data tables** while providing contextual information	**Summarize and highlight changes** in large bodies of regulatory documents	**Provide self-serve HR functions** (eg, automate first-line interactions such as employee onboarding or automate Q&A or strategic advice on employment conditions, law, regulations, etc)
Analyze customer feedback by summarizing and extracting important themes from online text and images	**Streamline customer service** by automating processes and increasing agent productivity	**Generate synthetic data** to improve training accuracy of machine learning models with limited unstructured input	**Answer questions from large amounts of legal documents,** including public and private company information	
Improve sales force by, for example, flagging risks, recommending next interactions such as additional product offerings, or identifying optimal customer interaction that leads to growth and retention	**Identify clauses of interest,** such as penalties or value owed through leveraging comparative document analysis			

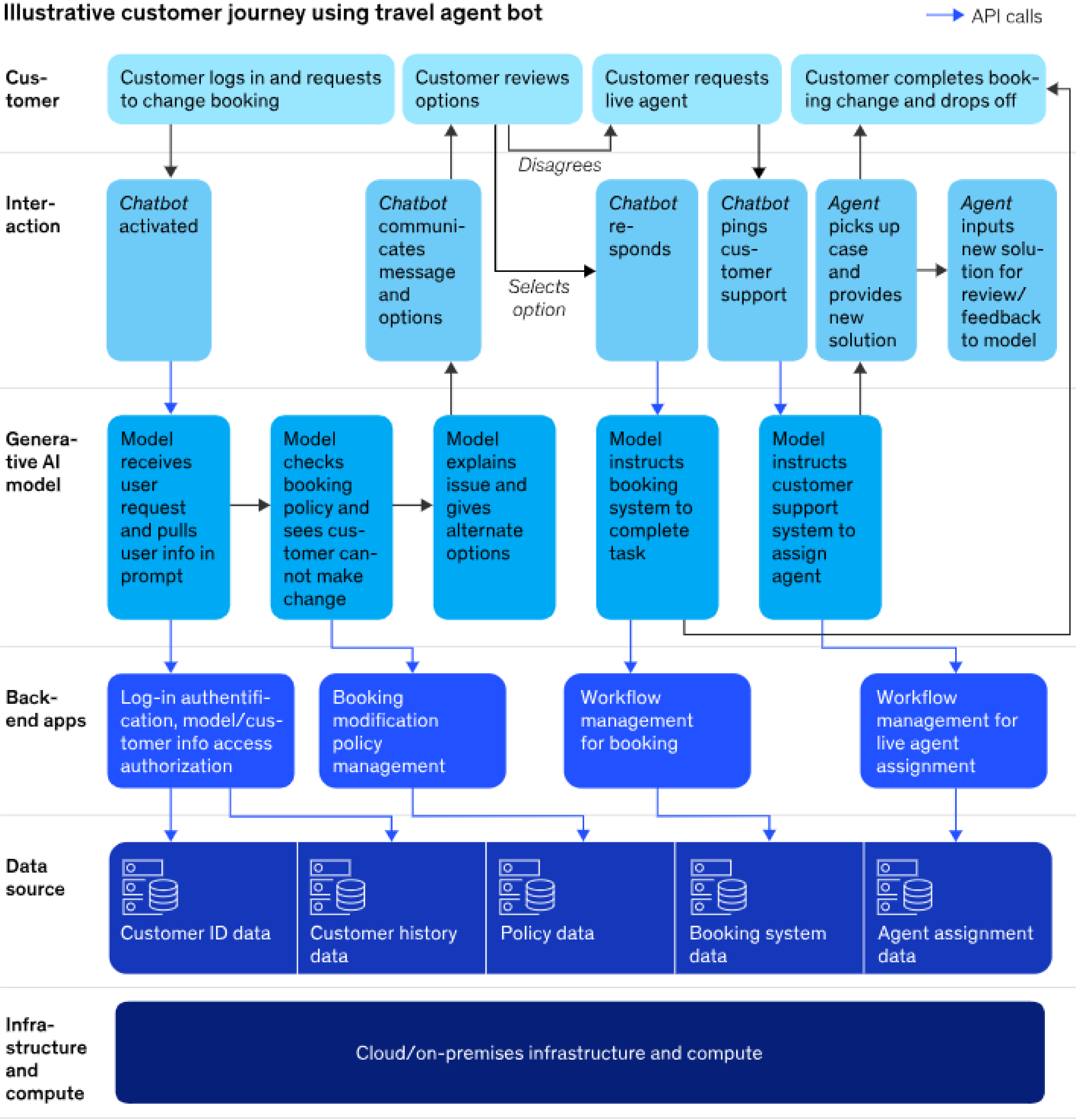
Credit: Mckinsey

4- LLM Development vs. Traditional Development:

- LLM Development:
  - No need for domain expertise, training examples
  - Focus on prompt design
- Traditional ML:
  - Requires training examples, compute time, and hardware
  - Domain knowledge needed for specialized areas (e.g., healthcare)
- Generative Question Answering:
  - A free-text generation without domain knowledge
  - Example: Google's Bard bot for simple calculations

Generative AI is integrated at key touchpoints to enable a tailored customer journey.

Illustrative customer journey using travel agent bot


API calls
Customer
Customer logs in and requests to change booking
Customer reviews options
Customer requests live agent
Customer completes booking change and drops off
Disagrees
Interaction
Chatbot activated
Chatbot communicates message and options
Selects option
Chatbot responds
Chatbot pings customer support
Agent picks up case and provides new solution
Agent inputs new solution for review/feedback to model
Generative AI model
Model receives user request and pulls user info in prompt
Model checks booking policy and sees customer cannot make change
Model explains issue and gives alternate options
Model instructs booking system to complete task
Model instructs customer support system to assign agent
Back-end apps
Log-in authentification, model/customer info access authorization
Booking modification policy management
Workflow management for booking
Workflow management for live agent assignment
Data source
Customer ID data
Customer history data
Policy data
Booking system data
Agent assignment data
Infrastructure and compute
Cloud/on-premises infrastructure and compute

Credit: Mckinsey

4- Prompting, Tuning, and Gen AI Tools:

- Types of LLMs:
  - Generic language models (e.g., autocomplete)
  - Instruction tuned (e.g., summarize a text)
  - Dialogue tuned (e.g., chatbots)
- Efficient methods of tuning:
  - Parameter-efficient tuning methods (PETM)
- Generative AI tools on Google Cloud:
  - Generative AI Studio & App Builder
  - PaLM API integrated with Maker Suite

ChatGPT Prompting Guide

1. **Tone**: Specify the desired tone (e.g., formal, casual, informative, persuasive).
2. **Format**: Define the format or structure (e.g., essay, bullet points, outline, dialogue).
3. **Act as**: Indicate a role or perspective to adopt (e.g., expert, critic, enthusiast).
4. **Objective**: State the goal or purpose of the response (e.g., inform, persuade, entertain).
5. **Context**: Provide background information, data, or context for accurate content generation.
6. **Scope**: Define the scope or range of the topic.
7. **Keywords**: List important keywords or phrases to be included.
8. **Limitations**: Specify constraints, such as word or character count.
9. **Examples**: Provide examples of desired style, structure, or content.
10. **Deadline**: Mention deadlines or time frames for time-sensitive responses.
11. **Audience**: Specify the target audience for tailored content.
12. **Language**: Indicate the language for the response, if different from the prompt.
13. **Citations**: Request the inclusion of citations or sources to support information.
14. **Points of view**: Ask AI to consider multiple perspectives or opinions.
15. **Counterarguments**: Request addressing potential counterarguments.
16. **Terminology**: Specify industry-specific or technical terms to use or avoid.
17. **Analogies**: Ask AI to use analogies or examples to clarify concepts.
18. **Quotes**: Request inclusion of relevant quotes or statements from experts.
19. **Statistics**: Encourage the use of statistics or data to support claims.
20. **Visual elements**: Inquire about including charts, graphs, or images.
21. **Call to action**: Request a clear call to action or next steps.
22. **Sensitivity**: Mention sensitive topics or issues to be handled with care or avoided.

Steve Nouri ✓

@SteveNouri

Thanks for reading!

If you enjoyed this post, dont forget to Click **Follow**.