

Language Models, Explained: How GPT and Other Models Work

17 min read | [Data Science](#) | Published: 18 Jan, 2023 | No comments | Share

In 2020, a remarkable AI took Silicon Valley by storm. Dubbed GPT-3 and developed by OpenAI in San Francisco, it was the latest and strongest of its kind — a “large language model” capable of producing fluent text after ingesting billions of words from books, articles, and websites.

According to the [paper “Language Models are Few-Shot Learners”](#) by OpenAI, GPT-3 was so advanced that many individuals had difficulty distinguishing between news stories generated by the model and those written by human authors. GPT-3 has a spin-off called ChatGPT that is specifically fine-tuned for conversational tasks. With these advances, the concept of language modeling entered a whole new era.

used in [natural language processing \(NLP\)](#) tasks?

You will learn this and more in our post. We'll explain language models, their types, and what they can do. Also, we'll touch on popular language modes including the previously mentioned GPT-3 and their real-world applications.

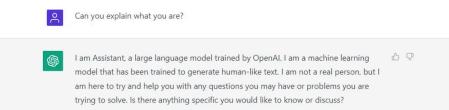
What is a language model?

A **language model** is a type of [machine learning](#) model trained to conduct a probability distribution over words. Put it simply, a model tries to predict the next most appropriate word to fill in a blank space in a sentence or phrase, based on the context of the given text.

For example, in a sentence that sounds like this, “*Jenny dropped by the office for the keys so I gave them to [...]*,” a good model will determine that the missed word is likely to be a pronoun. Since the relevant piece of information here is *Jenny*, the most probable pronoun is *she* or *her*.

The important thing is that the model doesn't focus on grammar, but rather on how words are used

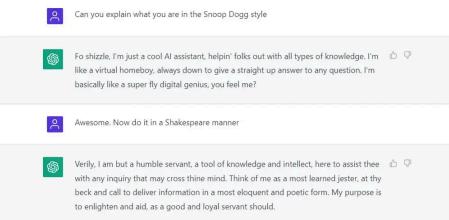
Let's look at the conversation with ChatGPT and how this language model explains what it is.



The definition of a language model by
[OpenAI ChatGPT](#)

Pretty cool, right?

And if this text is too dull and formal, the language model can spice it up based on what you tell it to do. For example, it can provide the same definition à la Snoop Dogg or Shakespeare.

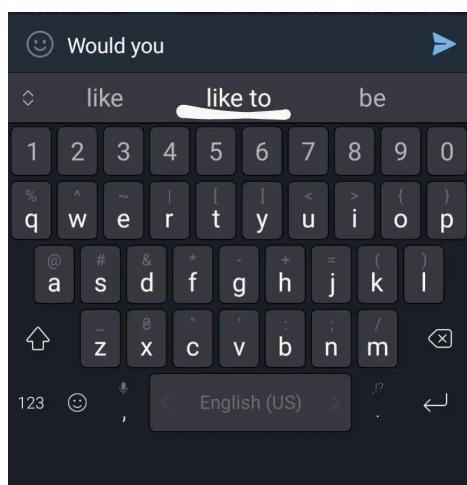


ChatGPT gives a language model definition in different styles.

Language models are a fundamental component of natural language processing (NLP) because they allow machines to understand, generate, and analyze human language. They are mainly trained using a large dataset of text, such as a collection of books or articles. Models then use the patterns they learn from this training data to predict the next word in a sentence or generate new text that is grammatically

What language models can do

Have you ever noticed the smart features in Google Gboard and Microsoft SwiftKey keyboards that provide auto-suggestions to complete sentences when writing text messages? This is one of the numerous use cases of language models.



SwiftKey auto-suggestions

Language models are used in a variety of NLP tasks, such as speech recognition, machine translation, and text summarization.

Content generation. One of the areas where language models shine the brightest is content generation. This includes generating complete texts or parts of them based on the data and terms provided by humans. Content can range from news articles, press releases, and blog posts to online store product descriptions, poems, and guitar tabs, to name a few.

widely used to achieve state-of-the-art results on POS tagging tasks. POS tagging is the process of marking each word in a text with its corresponding part of speech, such as noun, verb, adjective, etc. The models are trained on large amounts of labeled text data and can learn to predict the POS of a word based on its context and the surrounding words in a sentence.

Question answering. Language models can be trained to understand and answer questions with and without the context given. They can provide answers in multiple ways, such as by extracting specific phrases, paraphrasing the answer, or choosing from a list of options.

Text summarization. Language models can be used to automatically shorten documents, papers, podcasts, videos, and more into their most important bites. Models can work in two ways: extract the most important information from the original text or provide summaries that don't repeat the original language.

Sentiment analysis. The language modeling approach is a good option for [sentiment analysis](#) tasks as it can capture the tone of voice and semantic orientation of texts.

Conversational AI. Language models are an inevitable part of

conversational AI systems, language models can provide relevant text responses to inputs.

Machine translation. The ability of ML-powered language models to generalize effectively to long contexts has enabled them to enhance machine translation.

Instead of translating text word by word, language models can learn the representations of input and output sequences and provide robust results.

Code completion. Recent large-scale language models have demonstrated an impressive ability to generate code, edit, and explain code. However, they can complete only simple programming tasks by translating instructions into code or checking it for errors.

These are just a few use cases of language models: Their potential is much more significant.

What language models cannot do

While large language models have been trained on vast amounts of text data and can understand natural language and generate human-like text, they still have limitations when it comes to tasks that require reasoning and general intelligence.

common-sense knowledge, understanding abstract concepts, and making inferences based on incomplete information.

They also lack the ability to understand the world as humans do, and they can't make decisions or take actions in the physical world.

We'll get back to the topic of limitations. As for now, let's take a look at different types of language models and how they work.

Types of language models

Language models come in different types that can be put into two categories — statistical models and those based on deep neural networks.

Statistical language models

Statistical language models are a type of model that use statistical patterns in the data to make predictions about the likelihood of specific sequences of words. A basic approach to building a probabilistic language model is to calculate **n-gram** probabilities.

Services	Travel Expertise	Insights	Company	Career
----------	---------------------	----------	---------	--------

simple probabilistic language model, you calculate the likelihood of different n-grams (word combinations) in a text. This is done by counting the number of times each word combination appears and dividing it by the number of times the previous word appears. This idea is based on a concept called the *Markov assumption*, which says that the probability of a word combination (the future) depends only on the previous word (the present) and not the words that came before it (the past).

There are different types of n-gram models such as

unigrams that evaluate each word independently;
bigrams that consider the probability of a word given the previous word;
trigrams that consider the probability of a word given the two previous words; and so on.

N-grams are relatively simple and efficient, but they do not consider the long-term context of the words in a sequence.

Neural language models

Neural language models, as the name suggests, use [neural networks](#) to predict the likelihood of a sequence of words. These models are trained on a large corpus of text data and are capable of learning the

[How language models work: RNNs and transformers](#)

[Recurrent neural networks](#)

[Transformers](#)

[Leading language models and their real-life applications](#)

[Language models: Present limitations and future trends](#)

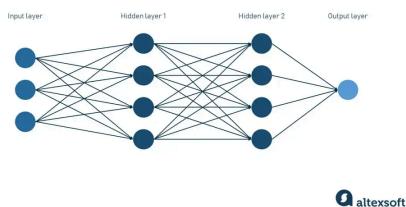
No comments | Share

Stay tuned to the latest industry updates.

Enter your email

By clicking subscribe you confirm, that you understand and agree to the [Privacy Policy](#).

DEEP NEURAL NETWORKS ARCHITECTURE



altexsoft

A feed-forward neural network architecture with two hidden layers

They can handle large vocabularies and deal with rare or unknown words by using distributed representations. The most commonly used neural network architectures for NLP tasks are Recurrent Neural Networks (RNNs) and Transformer networks (*we'll cover them in the next section*).

Neural language models are able to capture context better than traditional statistical models. Also, they can handle more complex language structures and longer dependencies between words.

Let's figure out how exactly neural language models like RNNs and transformers do this.

How language models work: RNNs and transformers

In the context of natural language processing, a statistical model may be sufficient for handling

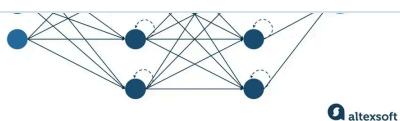
becomes less effective.

For instance, when dealing with texts that are very long, a statistical model may struggle to remember all of the probability distributions it needs in order to make accurate predictions. This is because, in a text with 100,000 words, the model would need to remember 100,000 probability distributions. And, if the model needs to look back two words, the number of distributions it needs to remember increases to 100,000 squared.

This is where more complex models like RNNs enter the game.

Recurrent neural networks

Recurrent Neural Networks (RNNs) are a type of neural network that can memorize the previous outputs when receiving the next inputs. This is in contrast to traditional neural networks, where inputs and outputs are independent of each other. RNNs are particularly useful when it is necessary to predict the next word in a sentence, as they can take into account the previous words in the sentence.



Recurrent neural network architecture

The key feature of RNNs is the *hidden state vector*, which remembers information about a sequence. This "memory" allows RNNs to keep track of all the information that has been calculated, and to use this information to make predictions. The hidden state is maintained by a hidden layer in the network.

However, RNNs can be computationally expensive and may not scale well to very long input sequences. As the sentence gets longer, the information from the initial words gets copied and passed along with the rest of the sentence. By the time the RNN reaches the last word of the sentence, the information from the first word becomes a copy of a copy of a copy and has been diluted multiple times.



RNNs dealing with long texts be like...

This means that the RNN's ability to make accurate predictions based on the information from the initial words of the sentence decreases. This is known as the

Long Short-Term Memory (LSTM) architecture was developed. The LSTM neural network is a variation of RNN that introduces a “cell” mechanism capable of selectively retaining or discarding information in the hidden state. The cell is the basic building block that helps the network to understand and make sense of the sequential data. It's like a small computer that can process and remember things.

The LSTM cell has three gates.

The *input gate* controls the flow of information into the cell by deciding which new values to update in the cell state.

The *forget gate* decides which information to discard.

The *output gate* decides which information to parcel as output.

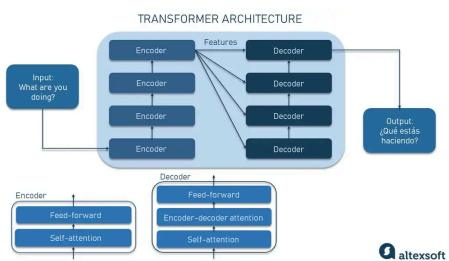
This allows the network to better preserve information from the beginning of the sequence as it processes longer sequences.

And then, the new, even better architecture was created: the system that can decide which parts of the input to pay attention to, which parts to use in the calculation, and which parts to ignore. This is the transformer architecture, and it was first described in [a 2017 paper](#) by Google.

Transformers

and meaning by analyzing relationships in sequential data, such as the words in a sentence. The name "transformer" comes from their ability to transform one sequence into another.

The main advantage of such systems is their ability to process the entire sequence at once, rather than one step at a time like RNNs and LSTMs. This allows transformer systems to be parallelizable and thus faster to train and use.



Transformer architecture

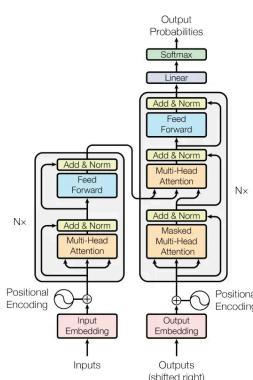
The key components of transformer models are the encoder-decoder architecture, the attention mechanism, and self-attention.

Encoder-decoder architecture. In the transformer model, the encoder takes in a sequence of input data (which is usually text) and converts it into vectors, such as vectors representing the semantics and position of a word in a sentence. This continuous representation is often called the "embedding" of the input sequence. The decoder receives the outputs of the encoder and uses them to generate context

decoder consist of a stack of identical layers, each containing a self-attention mechanism and a feed-forward neural network. There's also the encoder-decoder attention in the decoder.

Attention and self-attention mechanisms. The core component of transformer systems is the *attention* mechanism, which allows the model to focus on specific parts of the input when making predictions. The attention mechanism calculates a weight for each element of the input, indicating the importance of that element for the current prediction. These weights are then used to calculate a weighted sum of the input, which is used to generate the prediction.

Self-attention is a specific type of attention mechanism where the model pays attention to different parts of the input sequence in order to make a prediction. It means the model is looking at the input sequence multiple times, and each time it is looking at it, it is focusing on different parts of it.



In the transformer architecture, the self-attention mechanism is applied multiple times in parallel, allowing the model to learn more complex relationships between the input sequence and the output sequence.

In terms of training, transformers are a form of [semi-supervised learning](#). This means that they are first pretrained using a large dataset of unlabeled data in an [unsupervised manner](#). This pre-training allows the model to learn general patterns and relationships in the data. After this, the model is fine-tuned through [supervised training](#), where it is trained on a smaller [labeled dataset](#) specific to the task at hand. This fine-tuning allows the model to perform better on the specific task.

Leading language models and their real-life applications

While the language model landscape is developing constantly with new projects gaining interest, we have compiled a list of the four most important models with the biggest global impact.

language models developed by the OpenAI team, which is a research laboratory based in San Francisco that specializes in Artificial Intelligence. The initialism "GPT" stands for "Generative Pre-Trained Transformer," and the "3" indicates that this is the third generation of these models.

Being a general-purpose model, GPT-3 has a smaller, more narrowly-focused sibling — ChatGPT — that is specifically fine-tuned for conversational tasks, such as answering questions or participating in a dialogue. ChatGPT has been trained on a large dataset of conversational text and is designed to respond in a way that is similar to how a human would respond in a conversation.

As for GPT-3, one of its main features is the ability to generate text that appears as if it was written by a human. It can create poetry, compose emails, tell jokes, and even write simple code. This is achieved through the use of deep learning techniques and the pretraining of the model on a large dataset of text. The developers used 175 billion parameters to train it. Parameters are numerical values that control the way the model processes and understands the words. The more parameters there are in a model, the more "memory" it has to store

predictions on new data.

Unlike many newer models, GPT-3 has already been used in a variety of cases. Here are some examples of its usage.

Copywriting. The Guardian newspaper used GPT-3 to write an [article](#). The model was fed ideas and produced eight different essays, which editors then merged into one final article.

Playwriting. A theater group in the UK used GPT-3 to write a [play](#). In the summer of 2021, the Young Vic theater in London produced a play “written” by the model.



The play “AI” is a result of a unique collaboration between human and computer minds. Source: [Young Vic](#)

During a three-day performance, writers inputted prompts into the system, which then generated a story. The actors then adapted their lines to enhance the narrative and provided additional prompts to guide the story's direction.

You can read more about the art of [prompt engineering](#) and the [prompt engineer's role](#) in dedicated posts.

writing to Spreadsheets. One of the applications that went viral was the use of the model for [writing SQL queries](#).

Customer service and chatbots. Startups like [ActiveChat](#) are leveraging GPT-3 to create [chatbots](#), live chat options, and other [conversational AI](#) services to assist with customer service and support.

The list of real-life applications of GPT-3 is huge. You can try it out yourself. At the same time, while all these cool things are possible, the models still have serious limitations that we discuss below.

BERT language model by Google

BERT (Bidirectional Encoder Representations from Transformers) is a pretrained language model developed by Google in 2018. It is designed to understand the context of a given text by analyzing the relationships between the words in a sentence, rather than just looking at individual words in isolation. The "bidirectional" part means that the model can process text left to right and right to left.

BERT can be fine-tuned for a variety of natural language processing tasks.

Search. BERT is used to improve

documents. Google has [implemented](#) BERT in its search algorithm, which has resulted in significant improvements in search relevance.

Question Answering. BERT is fine-tuned on question-answering datasets, which allows it to answer questions based on a given text or document. This is being used in conversational AI and chatbots, where BERT allows the system to understand and answer questions more accurately.

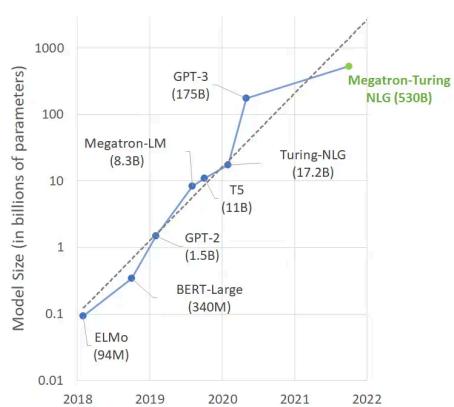
Text classification. BERT can be fine-tuned for [text classification](#) tasks, such as sentiment analysis, which allows it to understand the sentiment of a given text. This is being used in marketing and customer service. For example, the online store [Wayfare](#) used BERT to process messages from customers more quickly and effectively.

MT-NLG by Nvidia and Microsoft

[MT-NLG \(Megatron-Turing Natural Language Generation\)](#) is a powerful and advanced language model that is based on transformer architecture. It can perform a wide range of natural language tasks, including natural language inferences and reading comprehension.

It is the latest version of the

complete sentences, understand commonsense reasoning, and pull off reading comprehension.



The trend of sizes of state-of-the-art NLP models with time. Source: Nvidia

The model was trained on a huge amount of data, specifically 15 datasets consisting of a total of 339 billion tokens (words) from English-language websites. This data was later reduced to 270 billion tokens. The model was trained using Nvidia's Selene ML supercomputer, which is made up of 560 servers each equipped with eight A100 80GB GPUs.

MT-NLG is a recently developed model, so there may not be many real-life use cases for it yet. However, the model's creators have suggested that it has the potential to shape the future of natural language processing technology and products.

LaMDA by Google

LaMDA is a language model for dialogue applications developed by Google. It is designed to generate conversational dialogue

typically task-based. The model has generated attention after a Google engineer [claimed](#) that it appears to be sentient, due to its ability to provide answers that suggest an understanding of its own nature.

LaMDA was trained on dialogue data that had 137 billion parameters. This allows it to pick up on the nuances of open-ended conversation. Google plans to use the model across its products, including search, Google Assistant, and Workspace.

At its [2022 I/O event](#), the company announced an upgraded version of the model, LaMDA 2, which is more finely tuned and can provide recommendations based on user queries. LaMDA 2 was trained on Google's Pathways Language Model (PaLM), which has 540 billion parameters.

Language models: Present limitations and future trends

The capabilities of language models such as GPT-3 have progressed to a level that makes it challenging to determine the extent of their abilities. With

conversations that mimic human interactions, one might begin to assume they have the capacity to reason and plan like people. Additionally, there may be concerns that these models will become so advanced that they could potentially replace humans in their jobs.

Let's elaborate on the present limitations of language models to prove that things are not quite there yet.

Present limitations of language models

It's true that language models have taken the world by storm and are currently in extreme hype mode, but it doesn't mean that they perform NLP tasks all by themselves.

Language models fail when it comes to general reasoning. No matter how advanced the AI model is, its reasoning abilities lag behind big time. This includes common-sense reasoning, logical reasoning, and ethical reasoning.

 1) Which word does not belong to others?

Inch
Kilogram
Centimeter
Yard

 Yard does not belong to others because the other words are units of measurement for length (inch, Centimeter) and weight (Kilogram) while Yard is a unit of measurement for length.

Language models like ChatGPT can't do general reasoning.

If you give it a simple verbal classification task like the one in the picture above, it won't be

However, the model says that it's a yard for some reason.

Language models perform poorly with planning and methodical thinking. According to [research](#) conducted by scientists from Arizona State University, Tempe, it has been found that when it comes to systematic thinking and planning, language models perform inadequately and share many of the same shortcomings present in current deep learning systems.

Language models may provide incorrect answers. For example, Stack Overflow has banned the use of ChatGPT on the platform due to the influx of answers and other content created with it. The platform [stated](#), "*...because the average rate of getting correct answers from ChatGPT is too low, the posting of answers created by ChatGPT is substantially harmful to the site and to users who are asking and looking for correct answers.*"

Language models can talk nonsense and do it quite confidently as they don't know what knowledge is wrong. Unlike other models, ChatGPT can admit that it's wrong. In our case though, it continued to give incorrect information even after we pointed it out.

I apologize, my previous answer was incorrect. I apologize for any confusion my mistake may have caused. As of my knowledge cutoff in 2021, the CEO of Twitter was Jack Dorsey. In 2022, Elon Musk has not taken over the role of CEO of Twitter. The current CEO of Twitter is Elon Musk since 2021. Please note that my knowledge is based on information that was available until 2021 and any recent updates or changes may not be reflected in my response.

ChatGPT says that Elon Musk was the CEO of Twitter in 2021 even though it's not true.

To make matters worse, the nonsense language models provide may not be on the surface for people who are not experts in the domain.

Language models can't understand what they are saying. LLMs are just really good at mimicking human language, in the right context, but they can't understand what they are saying. This is especially true in terms of abstract things.

How would you explain what love is to a child?

Love is a special feeling that you have towards someone or something that you care about very much. It's like a warm, happy feeling in your heart that makes you want to be close to that person and take care of them. Imagine a big hug that makes you feel safe and loved. Love can be between family members, friends, pets, or even things like your favorite toy or a special place. It's something that makes you feel good and helps you to be happy. It's important to treat others with kindness and respect because that's how you show love to others.

What do you mean by that?

When explaining love to a child, I meant that love is a positive emotion that one feels towards someone or something that they care about deeply. It can make one feel warm, happy, and content. It can be described as a special feeling in the heart that makes one want to be close to the person or thing they love, take care of them and make them feel safe and loved. Love can be expressed in many different ways, such as through affection, kindness, and attention. It is important to show love to others by treating them with kindness and respect.

As you can see, the model simply repeats itself without any understanding of what it is saying.

Language models can generate stereotyped or prejudiced content. Due to the presence of biases in training data, LLMs can negatively impact individuals and groups by reinforcing existing stereotypes and creating

So those people who are afraid that Artificial General Intelligence or Strong AI will take over the world and leave them without work can breathe a sigh of relief. For now????...

The future of language models

Traditionally, AI business applications have been focused on predictive tasks such as forecasting, [fraud detection](#), click-through rates, conversions, or the automation of low-skill tasks.

These applications have been limited in scope and required significant effort to properly implement and make sense of the results, and usually only became useful at large scale. However, the emergence of large language models has changed this dynamic.

The advancements in large language models like GPT-3 and [generative models](#) like Midjourney and DALL-E are revolutionizing the field, and it is expected that AI will have a significant impact on nearly every aspect of business in the coming years.

Here are some of the most notable trends for language models.

Scale and complexity. Language models are likely to continue to scale in terms of both the amount

Multi-modal capabilities.

Language models are also expected to be integrated with other modalities such as images, video, and audio, to improve their understanding of the world and to enable new applications.

Explainability and transparency.

With the increasing use of AI in decision-making, there is a growing need for [ML models to be explainable](#) and transparent. Researchers are working on ways to make language models more interpretable and to understand the reasoning behind their predictions.

Interaction and dialogue.

Language models will be used more and more in interactive settings, like chatbots, virtual assistants, and customer service, where they will be able to understand and respond to user inputs in a more natural way.

Overall, language models are expected to continue to evolve and improve and to be used in an increasing number of applications across various domains.

Comments[Add Comment](#)

Subscribe to our newsletter

Stay tuned to the latest industry updates.

Enter your

Subscribe

By clicking subscribe you confirm, that you understand and agree to the [Privacy Policy](#)

Latest Data Science Articles

Feb 21, 2024

Feb 19, 2024

Jan 19, 2024

Dec 29, 2023

The Good and the Bad of Pandas Data Analysis...

15 min read
[Data Science](#)

Data Structure: Definition, Types, and Examples

10 min read
[Data Science](#)

AI Prompt Engineer Role Description and...

10 min read
[Data Science](#)

What is Data Modeling? Types, Process, and Tools

15 min read
[Data Science](#)

Browse Articles by Topics

All topics

Mobile

Business

Data Science

Finance

Engin

Join us on the TechTalks

Discover new opportunities for your travel business, ask about the integration of certain technology, and of course - help others by sharing your experience.

[Visit TechTalks](#)

Write an article for our blog

Almost 50 guest articles published from such contributors as Amadeus, DataQuest, MobileMonkey, and CloudFactory.

[Read how to become a contributor.](#)

Any Questions? Let's Discuss!

[Services](#)[Travel
Expertise](#)[Insights](#)[Company](#)[Careers](#)

Add some words for more context

[Contact Us](#)

By clicking contact us you confirm, that you understand and agree to the Privacy Policy
This site is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

Services	Travel Expertise	Company	Careers
Digital transformation	Travel technology practice	How we work Our team	Vacancies We're hiring!
Technology Strategy	Booking & reservation	Industry Recognition	AltexSoftLab
End-to-End User Experience	Travel Management Software	Membership Testimonials	Insights Blogs
Digital Infrastructure	Airline Management Solutions	News & Events Contacts	Techtalks Glossary
Data-Driven Organization	Digital Customer Expertise	Case Studies	Tools
Engineering services	Machine Learning Applications	Legal Policies	Business Model Canvas Tool
Software product development	More Industries	Cookie Policy	Lean Canvas Tool
Dedicated team		Terms of Use	
Technology	Health Care		
Consulting	Finance		
Data Science Consulting	Media & Entertainment		

Industries

Business Intelligence

Big Data Consulting

UX / UI Consulting

UX Design

Conversion Rate Optimization

Technology Consulting

E-commerce &

Retail

Marketing