Harvard Business School

9-724-383
REV: JANUARY 23, 2024

DAVID YOFFIE
ORNA DAN
ELENA CORSI

AI21 Labs in 2023: Strategy for Generative AI

Machines show signs of being able to solve difficult mathematical problems and automate cognitive tasks. Of course, there are cases in which the system will fail. But this field is the next frontier, and it's the biggest thing humanity has built in the past decade. We are not far away from building a new form of intelligence.

— Professor Amnon Shashua, Co-founder and Chairman AI21 Labs

Generative artificial intelligence (GenAI), a technology that enabled computers to consume existing content and create new content, including text, images, code, and video, took the world by a storm after OpenAI released ChatGPT in November 2022. Six months later, AI21 Labs co-CEOs Professor Yoav Shoham and Ori Goshen, and Professor Amnon Shashua, chairman, debated the future of their Tel Aviv-based GenAI company. Shashua, Shoham, and Goshen had co-founded AI21 Labs in 2017 to realize the vision of true machine intelligence. In late 2020, AI21 Labs took its first step by reinventing writing and reading with the launch of Wordtune, a B2C writing application that helped users improve their prose by using GenAI software to offer alternative text suggestions. That was followed by the 2021 release of Wordtune Read, a B2C application that summarized documents. By May 2023, AI21's B2C applications had nearly 9 million users and the company was predicting an annual revenue run rate of more than $50 million.

GenAI was a nascent, evolving field that demanded constant innovation. Building and training large language models (LLMs), as well as creating, selling, and maintaining applications, was an expensive process. The presumed leader in the field, OpenAI, lost $540 million in 2022, largely *before* ChatGPT took off. Venture firms and big companies like Microsoft, Meta and Google were throwing billions of dollars at GenAI. The founders of AI21 Labs regarded their opportunities as vast, but also recognized that they couldn't do everything. What were the best opportunities to monetize LLMs? Should AI21 Labs remain focused on B2C applications like Wordtune, or were there bigger opportunities in B2B? AI21 had previously decided against launching a chatbot similar to ChatGPT. Was that the right decision? Furthermore, should AI21 Labs invest heavily in creating a platform for third parties to innovate with its LLM or should it stay focused on building proprietary applications? And what was the best way to get their technology broadly adopted: Should they make some or all of their LLMs open source, like Meta? Should they license the technology to big players, like OpenAI did with Microsoft, or should they focus on proprietary usage models? Goshen, Shoham, and Shashua

Professor David Yoffie and Research Associate Orna Dan (Israel Research Office) and Associate Director Elena Corsi (Europe Research Center) prepared this case, with the assistance of Laura Wegner and Eloi Dessain. It was reviewed and approved before publication by a company designate. Funding for the development of this case was provided by Harvard Business School and not by the company. HBS cases are developed solely as the basis for class discussion. Cases are not intended to serve as endorsements, sources of primary data, or illustrations of effective or ineffective management.

were confident that AI21 could be big; after raising another big round of funding ($155 million) during the summer of 2023 (bringing the total capital raised to $330 million with a $1.4 billion valuation) the real question was how big and where to place their bets.

AI21 Labs: Brief History

The three founders, Goshen, Shoham, and Shashua, came together with highly complementary skills. Goshen began his career as a cyber engineer in 8200, the Israeli military's Intelligence Corps unit, after which he moved to the start-up world (see **Exhibit 1** for founders' biographies). Goshen was fascinated by developments in AI when his wife, a lawyer, introduced him to Shoham, "her partner at a non-profit that taught coding in underprivileged populations." Shoham was an entrepreneurial academic and an AI expert. With a PhD in Computer Science from Yale University, he joined the Stanford University faculty in 1987 and founded several AI companies, two of which he sold to Google. When Goshen and Shoham met, the latter had just moved back to Israel and was working as a principal scientist at Google. Goshen said, "After spending some time together, we decided to launch a company. Shortly after we met with Amnon, who wanted to invest in a technology start-up. His condition to join us was that he wanted to be involved." With a PhD in Brain and Cognitive Science from MIT, Shashua was the co-founder and CEO of Mobileye (the most highly valued tech firm in Israel in 2023), a professor at the Hebrew University of Jerusalem and co-founder of several other tech ventures. Reflecting on his initial meeting with Goshen and Shoham, Shashua said, "The next frontier for me is language models. This is what drew me to get involved."

The founders' initial idea was abstract, with a few guiding principles. Goshen said:

> Our first principle was built around the observation that in 2017 everyone was doing AI and the focus was on deep learning… Deep learning was a big deal, but we thought it was going to hit a glass ceiling. We wanted to focus on how to take the deep learning statistical approach and complement it with more symbolic methods [use of symbols and deductive reasoning] and deterministic algorithms [algorithms that, given a particular input, will always produce the same output]. There must be a way to marry the two.

Another guiding principle was that they focus on natural language. Goshen continued, "Everyone else focused on computer vision or programs that would allow computers to identify objects represented in digitized images. But text is everywhere. About 80% of the data in the enterprise is text. We said, vision is a lens into the eye, but language is a lens into the mind. It's harder to crack, but if you do that, the potential is limitless." The founders' initial objective was to change how people read and wrote. Goshen continued, "We write the way Microsoft defined word processing in the 1980s and we read the way Gutenberg designed the output of the printing press in the sixteenth century - someone puts text on a piece of paper or on a digital page, and we read that. Both experiences can be rethought if you put AI not just as an afterthought or feature, but at the core of the experience." The founders' final principle was to develop technology with a commercial value. Shoham noted that, "We wanted to build the company on three pillars: applications, platform, and technology; we wanted to reinvent reading and writing... build a platform which would work for large enterprises...and build the next generation of technology around deep learning married with symbolic reasoning."

The founders picked AI21 Labs as a company name, which was a contraction of artificial intelligence in the 21st century. In January 2018, they raised $3.5 million in seed money (see **Exhibit 2** for details on early funding rounds). Goshen said, "Amnon was a prominent investor, but we also attracted two Israeli funds, Pitango Venture Capital and TPY Capital, and a fund called 8VC from San Francisco."

LLM Technology

With money in the bank, the next step was developing the technology. Large language models (LLMs) were not new. At their core, LLMs were algorithms that used deep learning techniques on massively large data sets. In deep learning, each layer of a model focused on a specific feature of a given input (see **Exhibit 3** for an example of the deep learning process).[1] For example, if a machine was trying to recognize a picture of a cat, the first layer could focus on the eyes, another one on their shape and so on. The layer was made of several nodes offering various alternatives (different eye color, etc.). Only certain nodes would activate and pass on the information to the next layer, which would examine other features. The machines would thus learn how to recognize a cat by identifying different combinations that were more likely to lead to the same result. Once trained, these networks could generate content without explicit programming for that output. In 2014, further advancements in such algorithms enabled the creation of increasingly realistic content.[2]

The breakthrough on LLMs came in 2017 when Google Brain released a paper called "Attention is All You Need."[3] The paper put forth its "transformer" neural network architecture, which became the foundation for most LLMs. LLMs were often called "generative" because they were designed to help analyze existing content and generate new content. Shoham liked to call LLMs, "autocomplete on steroids." The theory was that the more data used to train the model, the better its capabilities and at the time it was believed that the larger the number of parameters (the number of variables in the model which were trained to infer new content), the better the results. The models were also called "pre-trained," because answers were delivered based on historical data rather than searching the web in real time. ChatGPT, for example, trained its LLMs (GPT 3, 3.5 and 4) on data before September 2021.

Some LLMs were also called "foundation models," a term coined by Stanford University researchers.[4] Foundation models were so large that they could serve as the foundation for more refined models. These models were often trained on unstructured and unlabeled data. The model would learn statistically what might come next in a prompt or phrase. Using transformer neural networks, the model would recognize relationships between words and concepts. Once the model was trained, the model could "infer" a response, which might be text, translations, content summaries, classifications, sentiment analysis, or just conversational bots. Some models were also trained for images and video. To process the data, LLM algorithms broke down the information received into units, or tokens, comprised of words, phrases, or characters. By analyzing each token within the data, the models could identify patterns and relationships. Once trained, given a prompt, they generated text by predicting the words more likely to appear together.[5] Shoham explained, "LLMs are huge statistical beasts… If you give them a sentence or text, they tell you what comes next." In response to user questions or prompts, LLMs were able to create original content, including computer code and prose.

The challenges for building an LLM were numerous: both development and operating costs were high; also, depending on the data used to train the model, there was always a risk of bias. Perhaps the biggest problem was "hallucinations," where the model mysteriously provided an inaccurate (and usually highly confident) response.[6] AI21's VP of Applications Or Dagan said, "The inherent problem that causes hallucinations, in terms of technology, is that the model must decide, that is guess, the most common next word. Hallucinations appear when something from the LLM's training data seems way more reasonable to complete the sentence because it is more statistically common." To mitigate these problems, AI21 has employed a variety of techniques, including retrieval methods and explicit inclusion of citations in its applications.

Despite these challenges, Goshen explained that AI21 Labs had "to build an LLM in-house. From the onset, we decided to be cloud neutral and to run on both Amazon Web Services and Google Cloud

Platform." The founders also realized that they should build applications in addition to an LLM. Dagan said, "When OpenAI's GPT 3 LLM came out, we started playing with it and in two weeks we created our application to rewrite and paraphrase on top of GPT 3." The result was Wordtune, launched in October 2020. In 2021 the company unveiled its own LLM, Jurassic-1, used exclusively for its proprietary applications. Shoham noted, "When it came out, Jurassic-1 was slightly bigger and better than GPT 3. It may not have been innovative, but it was a good foundation."

AI21 Labs Products: Applications. In 2023, AI21's flagship application was Wordtune, (see **Exhibit 4** for a screenshot of Wordtune). The user fed in their text, and the application provided multiple suggestions for how to improve the writing. Goshen said, "Wordtune was the first application that made generative models useful. You write a sentence, ask the system to rewrite it, and you get a series of different suggestions, not small variations like others did." Dagan said, "Over time we added a few features like grammatical error correction, but the focus is still on rewrite." The program had iOS and Google integrations on desktops and mobile devices and could be used on the Wordtune site, as an app, and as a Google Chrome extension. Director of Sales Aryeh Abramowitz noted that, "We do not yet have Outlook extensions."

AI21 Labs had another standalone application: Wordtune Read, which could summarize long articles and documents. Launched in 2021, Wordtune Read was kept separate from the rephrasing application, even though it had a similar name. Dagan said, "This technology is more complicated than rephrasing as it can more easily lead to hallucinations. But we believed it was crucial that we develop it. However, even today, some customers do not know that Wordtune Read exists and when asked how they could improve our rephrasing tool, Wordtune, they ask for a summary function." To reduce the risk of hallucinations, Wordtune Read summarized text by chunks. Dagan explained that the app "shows the user where the information was taken from. In this way, the user remains in control. All our applications keep the writer in control." (See **Exhibit** 5 for a screenshot of Wordtune Read). Shoham added, "With this new application, we have gone beyond key sentence extraction to abstraction, creating a true thought companion. Unlike other reading assistants, which pull out key sentences, we built a tool that can summarize text, a capability that is significantly more complex than sentence rephrasing, as it requires a deeper understanding of the broader context."[7]

Finally, the company's latest feature release was Wordtune Spices, launched in 2023. Wordtune Spices allowed Wordtune users to develop a given text by adding historical data, examples, jokes, or expanding an argument.[8] Dagan said, "When writing on Wordtune, users can click on the Spice icon and select what they need from a list of 13 options, such as continue writing, explain, or give an example. They get a suggestion which they can ask to regenerate if they are not happy with it. Each time our text additions are short."

The company followed a quota-based, freemium pricing approach, with a free option for low use and a monthly subscription model for more frequent use (see **Exhibit 6** for AI21 application pricing). Dagan said, "When we launched Wordtune, the feature was free but we had added-on functionalities in the premium version, like shortening text. Wordtune Read was launched with a complex pricing model, based on the number of documents per month. Today both applications are on a quota structure: you must pay for more than 10 rewrites per day or to summarize more than three documents per month." The change in business model increased the conversion rate. Dagan said, "We went from a 0.5% conversion rate to a 2% conversion rate, making scaling possible."

By 2023, Wordtune had nearly 9 million registered users. Chief Marketing Officer Eran Yacobovitch said, "Our churn rate is high, but similar to that of other B2C companies. On the second year of subscription, about 30% do not renew. The issue is pricing, moving to other providers, or simply not needing our paid version anymore." Part of the challenge was that almost half of Wordtune users were

students (and the other half worked in marketing and numerous professional jobs). AI21 Labs recently hired a salesperson to promote Wordtune applications to companies. But the new sales person, Aryeh Abramowitz, admitted that, "I haven't started doing outreach yet." While Wordtune gained sales traction, the general public and most executives remained unfamiliar with the company and did not link Wordtune to AI21 Labs. In addition, some potential Wordtune users were confused on why they should use such tools. Yacobovitch said, "We need to get better at conveying that our flagship product is not a grammar checking tool, but rather an AI writing and reading assistant." In May 2023, AI21 Labs signed a deal with a branding agency to improve brand architecture and build the company's brand internationally.

Most Wordtune users were native English speakers. Dagan said, "This was a surprise for us, but it makes sense as native speakers use it to look for nuances or to up their game." Wordtune was also popular in international call centers in countries such as India and the Philippines. Dagan added, "Part of the magic with Wordtune is that users see multiple options that go beyond corrections; they get to say the same thing they wanted to say, but in a way that sounds more native, fluent, or crisp." The company also planned to merge its two applications (Wordtune and Wordtune Read) into a Wordtune suite, with one comprehensive price. Shashua asserted, "The Wordtune suite, as a product, is the future of reading and writing."

AI21 Labs LLM Models

After starting with OpenAI's LLM, AI21 Labs had progressively shifted Wordtune to its own in-house model, Jurassic. Dagan said, "Since we launched Wordtune, we got a lot of data that has helped us train our own LLM model." Jurassic-1 was released in 2021 with two models: J1-Jumbo, a 178 billion parameter model, and J1-Large, a 7 billion parameter model. The Jurassic models utilized a unique (at least at its release) 250,000-token vocabulary, which was five times larger than most existing vocabularies, and were the first to include multi-word tokens such as expressions, phrases, and named entities. As a result, their models' architecture and tokenizer, which divided text into a sequence of tokens, were more efficient when representing a given amount of text. This allowed for increased computational efficiency and significantly sped up inference.[9] A White Paper published by AI21 found that, "Our Jurassic-1 models can predict text from a broader set of domains (web, academic, legal, source code, and more) than GPT-3, achieve comparable performance in zero-shot settings, and can be superior to GPT-3…due to their ability to fit more examples into a prompt."[10] The challenge was that Jurassic, with 178 billion parameters, 3 billion more than GPT 3.5, was very expensive to train and run. The larger the model, the higher the latency, or the time it took the machine to answer to prompts. Inference costs were also higher, as the machines used more compute power to generate an answer.[11] The rule of thumb was that each LLM conversation cost single-digit cents in processing power.[12]

While OpenAI moved to even bigger models (GPT 4 was rumored to have more than 1 trillion parameters),[13] AI21 argued smaller, more efficient models could work better. Shashua explained, "You need a very large model to begin with in order to create a data flywheel. Once you have all the data, you can build a smaller, more optimized model. This is what we are doing with Wordtune. There are only about five companies that can do this, and we are one of them." The flywheel referred to AI21 capturing data from consumers using Wordtune, using that data to improve the model, then providing a better Wordtune service, which generated more data, and so on.[14] AI21 believed that LLMs should be built in different sizes, with varying levels of latency and performance.

This led AI21's developers to build their second LLM, Jurassic-2. Released in 2023, Jurassic-2 came in three different sizes: Light, Mid, and Ultra (see **Exhibit 7** for comparison of Jurassic-2 models). Ultra had the highest latency and was most capable of conducting complex generation and comprehension

tasks; Mid allowed users to easily scale language comprehension or generation tasks, including question answering, summarization, copy generation, and advanced information extraction. The Light model was the smallest and fastest, performing sentence classification, named entity recognition, short-form copy generation, sentiment analysis, and keyword extraction.[15]

Shashua noted that Jurassic-2 was optimized for AI21 Labs's business: "You cannot jump directly to the small model without first going to the big model." With the smaller model, speed and reduced latency performed up to 30% faster than Jurassic-1 models. VP of Platform Dan Padnos explained, "Smaller models are less expensive as the inference costs (those incurred by the LLM to answer queries or infer) are smaller. But smaller models have to be trained longer. The question is: do you prefer a model with less training costs, which can be very high… or one with less inference costs?"[16]

Going forward, the company's technical leadership believed that the race to develop larger and larger LLM models might be coming to an end. Shoham said, "The conventional wisdom today is that increasing the number of parameters is no longer an objective to chase. GPT 3.5 was smaller than GPT 3, even though GPT 4 is bigger because it is multi-mode [it can generate images and video in addition to text]. A recent study analyzed how many LLMs are undertrained and concluded that model size and the training dataset size should be scaled equally. Considering how big the LLMs are, there isn't enough data in the world to justify more parameters." In addition, the more parameters, the higher the computational and memory resources needed. Goshen explained, "There's an optimization between the size and efficiency. It depends on the terms stipulated by the cloud provider, but training a model with 100 billion parameters costs millions of dollars, and sometimes you need to train multiple models." Yet speed and lower costs came with trade-offs. While Jurassic-1 had many of the same capabilities as ChatGPT with GPT 3.5, Jurassic-2 was programmed for reading and writing, not coding and other conversational use cases. OpenAI's GPT 4 also had other real advantages: A big one was that GPT 4 could use conversation context. If a user asked a question, GPT 4 would remember that question, and adjust future answers accordingly. Jurassic-2 did not have that capability yet.

AI21 Studio: B2B SaaS

Creating different usage models enabled AI21 Labs to start AI Studio in 2023, a B2B platform that gave companies access to Jurassic-2 and allowed firms to build and create company-specific text-based applications. B2B apps might include chatbots and virtual assistants, and perform text simplification, write original text, and moderate content. AI21 Studio also offered five application program interfaces (APIs) for third party developers based on Jurassic-2. While AI studio was just getting started in 2023, with relatively little revenue, the team could see many potential B2B applications. For example, while individuals might like Wordtune Read for summarizing one document at a time, a company might want a function that automatically summarized all incoming information from particular sources. Padnos suggested, "The first usage is summaries. But these come in a million shapes, sizes, and flavors and we can help customize it." The second main usage was "grounded question answering," whereby Jurassic-2 would be used to provide answers developed from a specific document or set of documents. For example, this service could be used by insurance companies trying to create a chatbot for their customers, allowing them to ask specific questions related to their insurance, or by a large bank exploring automated summaries of nearly 1 million analyst reports for their brokers.

By mid-2023, 35,000 developers had registered on the Studio platform. Pricing was usage-based. Shoham said, "We have a few large brands working with us, but it's still experimentation. However, these are all $50,000 or so engagements and each present us with a six to seven figure annual opportunity." AI21 Labs had ten ongoing pilots with large corporations, some paid and others unpaid. Padnos said, "It will be an arduous process. But I'm hoping we can convert a significant fraction to

long-term clients." Yacobovitch, AI21 Labs's CMO, described Wordtune, Wordtune Read, and B2C apps as the revenue drivers in 2023, but AI Studio "is where the world is going." When reaching out to businesses, AI21 Labs pushed for standard functionalities, but it was ready to customize its products to meet customers' demands. Shoham said, "We tend to steer companies towards one of our APIs, like paraphrasing and summarization, but if we see an emerging need…we'll build one."

AI21 Labs' Business Model in 2023

In mid-2023, Goshen and Shoham managed the company as co-CEOs. Goshen said, "Both of us are deeply involved in all areas, but Yoav gravitates more towards scientific elements, and I focus more on the product and operational sides." Shashua generally came to the office one day per week. He liked to dive into all technology and strategy decisions. The company was scaling fast, with 100 people added in the prior 12 months, and nearly 200 full-time employees.

The leadership team saw a clear path to a half-billion dollars in revenues over the next several years. Between 2021 and 2022, its revenues grew from minimal revenues to a $20 million ARR and CFO Golan Mizrahi estimated $54 million in annual recurring revenues by Q4 of 2023, then doubling in 2024, and more than doubling again in 2025, getting the company to be cashflow positive. Most of the short-term revenue would come from the B2C business. The challenge was that "The investment required to train and serve LLMs is very compute intensive," according to Goshen. "There remains a considerable amount of investment in terms of R&D and marketing that may defer our time to profitability."

Similar to other LLM-focused companies, its most significant expenses were training costs. According to AI21 Labs' CTO Barak Lenz, "The biggest models run on thousands of GPUs for several months, rented from AWS or Google Cloud." The length of the training process drove up costs. For example, industry analysts speculated that OpenAI spent as much as $100 million to train GPT 4.[17] AI21 Labs' LLM training costs, on the other hand, varied with model size. Goshen said, "our smaller models cost less the $1 million to train." The company also had to pay inference costs.

For R&D, most of the costs were cloud-related and labor. AI21 Labs's R&D costs were expected to exceed revenue in 2023. But as revenue scaled and chip makers, such as Nvidia, improved and optimized performance, R&D and cloud costs were predicted to drop to roughly 20% of sales by 2025. Marketing and sales also faced similar scale economies: Acquiring premium B2C customers was costly, ranging from $180 to $250 per user. Google search was the lion's share of conversions. After ChatGPT launched, however, costs went up: more firms bid for auction search terms, such as writing assistant, paraphrasing, etc. As a result, expenditures on digital marketing and the sales team would also exceed revenues in 2023. If AI21 Labs could hit their 2025 revenue goal, however, sales and marketing would also drop to roughly 20% of the top line. Over time, training costs were expected to grow (2024 would be 4X 2023) and then hopefully drop. AI21 Labs' plan was to optimize its newer models, and then run smaller, significantly more capable models, which would reduce training and inference costs. The company's burn rate was $12 million per quarter in early 2023.

Amazon Deal

One of the biggest potential business model breakthroughs in the B2B space came in April 2023 when Amazon Web Services unveiled Amazon Bedrock, a suite of technologies offering users a platform to build generative AI-powered applications using various pretrained models, including AI21's Jurassic-2. With Bedrock, the LLMs were accessible to developers through open APIs. Clients could customize the model they selected for a given application with their own data. This ability to

train a model using the business's proprietary documents and information would facilitate the generation of output matching the company or organization's specific requirements.

As part of its partnership with Amazon Bedrock, AI21 Labs provided LLMs as a fully managed service. The client on AWS would not have to deal with the LLM infrastructure or manage anything related to workloads. Goshen said:

> We set the price and we pay for the compute. Amazon handles the billing, charges the customers, and takes a 20% cut. Amazon owns the customer, but our brand is now on AWS. We want to ensure that we can offer other added value services, so that we are not just behind the scenes. The customers are looking into these LLMs and how to productize them. They seek guidance in dealing with any problems, and while we're not a consulting shop, we are currently in a highly consultative phase."

AI21 Labs concluded a similar deal with Google Cloud. Both AWS and Google offered volume opportunities that would be otherwise hard to capture. Goshen said, "I met with several general managers from Amazon, with the intent to expand our collaboration to additional areas, not just the cloud, but also shopping and many other services that Amazon provides, because they all need LLMs." For example, an obvious problem for Amazon was how to write compelling product descriptions for its 500 million SKUs? Wordtune could offer that capability.

The Generative AI Industry in 2023

The incredible excitement for GenAI in 2023 meant that AI21 Labs, OpenAI and other players could expect intense competition. When Microsoft announced that it would invest $10 billion in OpenAI for roughly 30% of the company,[18] every venture firm and large tech company took notice. CB Insights estimated that $14.2 billion were invested in over 300 GenAI startups in the first six months of 2023 alone.[19] Many LLMs were available, differing in the training data, the architecture of their software, and the number of parameters.

All LLM competitors struggled with issues around privacy. Just as AI21 Labs used the data collected from the consumer version of Wordtune to improve the training of their models, LLMs had access to users' queries, which the model owner could read and capture to train future versions of the model. Additionally, queries stored online could potentially be hacked or made publicly accessible, putting users' private information at risk.20 This fear over security of the data led some big companies to ban ChatGPT and other LLMs.21 For AI21 Studios, however, the company guaranteed privacy for corporate solutions. Padnos said, "AI21 Labs adheres to all of the best practices in terms of data security and privacy. We can guarantee that we do not retain a company's information or data and will never be able to use it to debug our models, retrain them, or even monitor them beyond very basic instances. Furthermore, we have an offering with AWS [Bedrock] and Google Cloud Platform (GPC) to deploy our models on clients' Virtual Private Cloud, which essentially keeps the data lifecycle fully in each client's control." As part of the deal with Amazon Bedrock, customization was done on a private copy of the model, within a virtual private cloud. Inputs into the model were encrypted and could not be used to further train the general-purpose model,22 keeping the information private and proprietary.

For the industry as a whole, some experts, as well as government officials and regulating bodies around the world, voiced concerns about the potential dangers of AI, and proposed various approaches to regulate the industry. Shashua worried "Regulatory bodies could create setbacks for the industry. Europe is likely to be among the first to set such new rules."

AI21 Labs' LLM Competitors

One big challenge for everyone, including AI21 Labs, was that building an LLM was relatively easy; building a great LLM was hard. CTO Lenz noted, "There is no textbook. Companies can be far ahead today, but it is hard to sustain. The tech is changing so fast, the competitive dynamic could quickly shift. What would happen if OpenAI changed their rules and pissed off their customers?"

In Stanford University's Holistic Evaluation of Language Model rankings,[23] Jurassic-2 LLMs were often ranked in the upper end, depending on the feature analyzed.[24] However, it was still early in the technology. Billions of dollars were flowing into GenAI start-ups, in addition to billions being invested by tech giants such as Microsoft and Google.

Following the release of ChatGPT, OpenAI became AI21 Labs' most well-known competitor (see **Exhibit 8** for a list of competitors with LLMs built in-house). Launched in 2015, OpenAI was a capped-profit organization offering a range of generative AI products. Its first LLM product, GPT-1, was launched in 2018 and was followed by other updates up until GPT-3.5 in 2022 and GPT-4 in 2023.[25] The company also offered additional products, including DALL-E, which combined images and text. OpenAI's best known product was ChatGPT. Released for public use as an interactive chatbot, ChatGPT's free version was built on GPT 3.5, while its premium version ($20/month) was built on GPT 4. Both models were trained with data up to September 2021. ChatGPT's dialogue format made it possible for the chatbot to respond to prompts and answer follow-up questions in a conversational manner,[26] and it garnered millions of users around the globe in a short amount of time.[27] Goshen said, "Our Jurassic-2 competes against Open AI's LLMs offering, but ChatGPT is getting into our applications sphere, as some use it to summarize and rewrite text. But they get two or three pages of answers that they do not know if they correspond to the truth or not." Goshen also noted that while it was still very early, there were a few signs that ChatGPT may be cannibalizing some Wordtune sales.

Experts estimated that ChatGPT was expensive to train and run. Goshen said, "GPT3, which has 175 billion parameters, probably cost $50-100 million in training. GPT 4 is likely even bigger." ChatGPT was also likely to have high inference costs. Yet OpenAI was expected to generate $1 billion in revenue in 2023 and it had secured massive funding from Microsoft, totaling roughly $13 billion since 2019.[28] Shashua said, "OpenAI has so much money, that I think they'll go after bigger and bigger models." Shashua also noted that ChatGPT was a "horizontal model, where one size fits all… ChatGPT is unlikely to go deep into a vertical, like AI21, which companies can put on-prem if they are concerned about security and privacy."

In the meantime, in March 2023, Microsoft integrated GPT4 into its search engine, Bing. The latter was enabled to give direct answers to users' questions, as well as give links to current web content. Microsoft CEO, Satya Nadella, announced that ChatGPT would serve as a "co-pilot" for all of Microsoft's applications and Windows. There would be a button on every app to create new conversational and generative capabilities for Microsoft 365 and Windows 11.[29] For example, Microsoft expected that users could give "co-pilot" a prompt, and it would apply the user's data and style to build a PowerPoint deck from scratch. Due to the cost of running GPT 4, Microsoft announced that "co-pilot" would cost every Microsoft licensee an additional $30 per month.[30]

The emergence of Microsoft's commitment to ChatGPT created a "code-red" for Google.[31] Although Google had pioneered LLMs, built its own LLM called LaMDA, and developed its own chatbot called Bard, it was slow to roll out the technology. Concerns over hallucinations and LLMs' other limitations led Google to be cautious. However, when ChatGPT took off like a rocket and Microsoft embraced it, Google perceived the first real threat to its search franchise in two decades. Google probably had the most data available to train its LLMs, but its early tests of Bard were met with

mixed reviews.[32] Google announced that it planned to integrate Bard into many Google apps in 2023, including Search and Google Assistant, and open APIs for third parties.[33]

In February 2023, Meta, the owner of Facebook, also launched an LLM. The first version had 65 billion parameters and, similar to AI21 Labs, Meta also offered smaller models of 7, 13, and 33 billion parameters. The model, called LLaMA, was made open source, initially only for researchers, and later for commercial usage.[34] Similar to Google, Meta had deep expertise in GenAI, but it had been slow to bring it to market. Mark Zuckerberg was betting that even though it was late, Meta could attract developers, who often preferred to build on open source platforms.

A few startups were also gaining traction. Anthropic, which had been launched by former OpenAI employees, had raised more than $1 billion[35] and offered Claude, an AI chatbot powered by a 52 billion parameter LLM. Its models could be used by companies willing to apply generative AI to their customer service, automate workflows, provide answers, and take on various roles in a dialogue.[36] Similarly, Cohere had upwards of $400 million in funding[37] and offered LLMs enabling companies to add AI to their chat features, generate text for product descriptions, blog posts, and articles, and capture meaning of text for search, content moderation, and intent recognition.[38] Another player was Writer, focused more on using AI for marketing, such as product description.[39]

Open Source

Additional competitors were emerging, including companies that worked on developing LLMs by building on data, code, design documents and content that was open source. Google, for example, built BERT (Bidirectional Encoder Representations of Transformers) as an open source model. [40] Although BERT was no longer competitive, more open source models followed.[41] Shoham stated, "There are very few developers who can build excellent LLM models. There is so much detail and know-how required on the algorithms, architecture, engineering, and data curation sides. Open source models are often 7 to 17 billion parameters and focus on a specific task."

The potential role of open source was contentious. Google had demonstrated a decade earlier that an open source platform could be profitable, if it became a standard. It was unclear in 2008, when Google started giving Android away for free that it would ever make money. But by 2016, Google generated $31 billion in revenue and $22 billion in profits from Android.[42] Yet it was unclear in 2023 whether Meta or any LLM provider could make money. Lenz noted "the jury is still out on open source for LLMs." In a controversial memo on LLMs leaked by a Google engineer in 2023, titled "We Have No Moat," the engineer stated:

> Plainly put, [open source] are lapping us…Open source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with $100 and 13B params that we struggle with at $10 million and 540B. Indeed, In terms of engineering-hours, the pace of Improvement from [specialized] models vastly outstrips what we can do with our largest variants, and the best are already largely Indistinguishable from ChatGPT…The modern Internet runs on open source for a reason. Open source has some significant advantages that we cannot replicate.[43]

Shashua disagreed with the Google engineer. He felt that open source models in the "medium sized networks, with 7-10 billion parameters, may be competitive, but not with very large models. And you need to start with a very large model before it can be trained and optimized before going smaller."

Other AI21 Lab Competitors

Beyond GenAI firms, there were an array of tools available as applications or extensions that used machine learning to correct users' grammatical errors and improve their writing. Shoham said, "We deliberately excluded spellcheck and grammar correction from Wordtune at the beginning, even though we can do it, because this is not how we want to position ourselves. Now we offer it, but it's built into our rewrite function."

The most widely used of these applications was Grammarly, which was founded in 2009 and by 2023 had more than 70 million monthly users (see **Exhibit 9** for key players with Applications). AI21 Labs was directly attacking Grammarly's audience. By using various AI techniques, Grammarly's software was able to suggest corrections and improvements. Grammarly's algorithm analyzed individual sentencing and offered suggestions pertaining to verb tense or word choice to enhance users' text. By accepting or rejecting the application's suggestions, users helped the algorithm improve, as it could determine whether its suggestions were correct based on whether users accepted them.[44] Grammarly could be used as a desktop application on Windows or Mac, as a browser extension, on Microsoft 365 or Google Docs, and on iPhones and Android devices.[45] Goshen added, "Grammarly is estimated to generate revenues of about $500 million. It was valued at about $13 billion in 2021, but today I think that it lost some value." While Grammarly and AI21 Labs addressed the same audience, their capabilities were quite different. According to Dagan, "A common pain point for users was making a poor word choice. Grammarly won't help because the grammar may be correct, but Wordtune's ability to rewrite offers a better solution."

Other applications in the writing improvement field included Quillbot, which was founded in 2017 and had extensions for Chrome, Word, and macOS. The application offered monthly or annual plans and could correct grammar and spelling errors, as well as paraphrase text. Founded in 2015, Quillbot was purchased by venture funded, Course Hero, in 2021, for an undisclosed sum.[46] Jasper billed itself as an AI copywriting assistant that could create original content, boost ad conversions with better copy, and scale up content marketing quickly.[47]

Debating the Future

Despite the explosion of competition from tech giants and well-funded start-ups, Shashua, Shoham and Goshen were excited by the upside potential of AI21 Labs. Goshen said, "I do not think it's a fad. It's like the mobile revolution, and maybe even bigger. I speak with representatives from five to seven companies a week and I hear the same things from each CIO, CTO, or Head of AI. They got orders from their CEOs that they must do something with AI not by the end of the year, but this quarter! I don't think anyone could have predicted this halo effect."

Among the many advances the founders envisioned for AI21 Labs, one was what they referred to as "blades" or specialized language models for specific industries. Shashua envisioned developing such modules, specific to the medical, legal, financial, or other fields. He said, "Just as people have skills, you can build a language model with reasoning capability on a specific language. You can have a machine that is very good at logical reasoning, or good at summarizing papers, or creating Excel files for you, and then one which also had a specific language blade, or industry focused." Shoham added: "We've trained our LLMs on everything that has been written on the internet, scanned books, more than any individual could ever consume. It's expensive and once you have it, serving them can be costly, but when you do all that, you have something that is very substantial."

In parallel, AI21 Labs was working at going beyond LLMs. Shoham said, "LLMs are not enough. They have limitations and to overcome these limitations, you need to break the pattern. We're venturing into the area of a neurosymbolic system, which combines the power of neural nets with symbolic reasoning." AI21 Labs worked to develop a new model, built on their core technology. Goshen said,

> In a couple of years, the focus will move on to symbolic models, or AI systems that incorporate events, not just a single language, and that are more deterministic. The current LLM models have an interface with a prompt that specifies an instruction, like 'summarize this for me.' Our new model will build an intermediate plan before it generates outcomes. If you need to write something, [for example], it would ask for input to develop an outline first. If you change the outline, it will affect the generated output and vice versa. It will be much more interactive and explainable.

This new model would also be able to rely on external software. Shashua said, "For example, if the system has to do a calculation, it should not be trained so that it can do it itself, as this would require feeding the system with lots of calculations to get to the right statistics and result. Instead, the machine should understand that there is a calculation that has to be done, and then go to a software that does the calculation. We spent years building good calculators and they are more efficient."

Other players had already entered this realm. In 2023, AutoGPT launched an open source app that used OpenAI's text-generating models. AutoGPT automated multi-step projects by interacting with apps, software, and services to carry out the steps needed to complete the user's intended goal.[48] Shashua said, "Today all automation pertains to physical tasks. But we are on the verge of a transition whereby machines will be able to automate cognitive tasks… And it is going to be disruptive. If AI21 is among the small list of innovators with the capability to build these machines, we're talking about a market of trillions of dollars." Goshen added, "I think that we will see much higher levels of automation in the future, involving humans whenever necessary for their judgment."

Previously, AI21 Labs had dismissed the idea of developing a chat product. In fact, AI21 leadership believed that ChatGPT's success was unanticipated by everyone, including OpenAI. No one, not even its creators, imagined that it would become the fastest adopted app in history. Shashua explained: "A year ago we discussed this option but dismissed it, as there were many attempts to do chatbots in the past, all of them by reputable companies, and all of them were ultimately disappointing. We picked a different direction. Today some ask us if we should not reconsider it."

In some ways, the biggest challenge for Shashua, Goshen and Shoham was that there were too many opportunities for a hot nascent technology. Shoham explained, "Management attention is the scarcest resource, followed by people; money is the least of our concerns." The B2B segment, for example, seemed ready to take off, but even after the boost from the Amazon deal, revenues were just getting started. Padnos said, "I think there's going to be a lot of competition, especially in B2B, because big enterprises would not settle for one vendor."

One debate was how to balance investment between B2C and B2B. Going forward, should AI21 emphasize its current revenue driver - B2C? Remain balanced? Or go all-in on B2B? Keeping both B2C and B2B businesses brought several advantages, according to Shoham: "There are synergies among these two businesses. We have common technology and amazing data that our applications give us to help improve our model. But each has different go-to-market approaches, and we need additional engineering support. This is potentially dilutive. We will continue to assess whether the upside of this multi-pronged approach outweighs its downside." In addition, Shashua argued that the focus on both markets distinguished AI21 Labs from the competition. He said, "If you look at the landscape, there is

not only OpenAI, but many other players, like Cohere. Then you have the LLMs created by large multinationals like Google and Meta. What distinguishes AI21 Labs is that we have a flywheel that fuels our LLM."

In addition, competing on in B2B and B2C could preclude certain deals. Shoham continued, "People look at how we compete with Microsoft and Google while we offer general features within their applications and platforms. We believe we are ahead of the curve, but we will have to keep running to stay in place." One potential investor even asked whether AI21 Labs would consider selling Wordtune to focus on the B2B platform. Shoham said: "Our answer was no. But it's something we will continue to evaluate." Goshen added, "There is indeed a very limited number of companies that are really able to play in this arena, and these are mainly the big tech companies. But there is room for several players who are already operating in the field and have the same technological ability as us."

Another ongoing debate was how to think about open source. Shoham, like Shashua, was not enamored by open source: He argued that "Open source will never be competitive with proprietary models." At the same time, he wondered aloud whether the company should have considered making its first model, Jurassic-1, open source to attract a larger army of developers. (Subsequently, AI21 decommissioned Jurassic-1.) While Shashua thought it was a "dumb idea" to open source the company's large models, he pondered whether open source would be a good idea for one of the small Jurassic-2 models?

Getting on the Short List

It was fun to be in a high performing organization in a thriving market space. The future felt incredibly bright to the three founders. Yet there were always things to worry about. Shoham reflected:

> With everything moving so fast, I worry: can we deliver? Can we execute? Do we have the right people? Do we have the right vision for AI21? Whenever people speak about premier AI companies, we want to be in the conversation. We need to have technology ahead of the competition and be recognized as such.

Shashua also had big dreams for AI21 Labs:

> It was clear to me when Google published its paper on transformers, this would be the next phase of computing...I see the end game as building intelligence, and we are starting to see early signs of intelligence. But I agree with Yoav: the biggest risk is execution; if we execute, we will be on the short list.

Goshen concluded: "There is so much low hanging fruit… we just have to grab it."

Exhibit 1 AI21 Founders' Biographies

Ori Goshen, Co-Founder of AI21 Labs

Ori Goshen began his career in 8200, the Israeli military's Intelligence Corps unit, as a cyber engineer. Following his military service, he joined fring, a startup, and later founded his own company, Crowdx, which created a mobile application that was downloaded and used by millions of users worldwide. The company evolved into a wireless analytics company and was eventually acquired by Cellwize, another Israeli startup, and subsequently acquired by Qualcomm. He completed his tenure at Cellwize in 2017, after which he co-founded AI21.

Yoav Shoham, Co-Founder of AI21 Labs

Professor Yoav Shoham earned his PhD in Computer Science from Yale University. He joined Stanford University in 1987 and his online Game Theory course had been viewed by nearly 1 million people. He was the founding chair of the Artificial Intelligence Index, which tracked global artificial intelligence activity and progress, and established multiple AI companies prior to founding AI21. Shoham served as principal scientist at Google from 2015-2017.

Amnon Shashua, Co-Founder of AI21 Labs

Professor Amnon Shashua completed his PhD in Brain and Cognitive Science at MIT. He held the Sachs Chair in Computer Science at the Hebrew University of Jerusalem and published more than 160 papers pertaining to machine learning and computational vision. In 1999 he founded Mobileye, which developed autonomous driving technologies and advanced driver-assistance, including cameras, computer chips, and software. Mobileye was acquired by Intel Corporation in 2017 in the largest acquisition in Israeli history, totaling $15.3 billion.

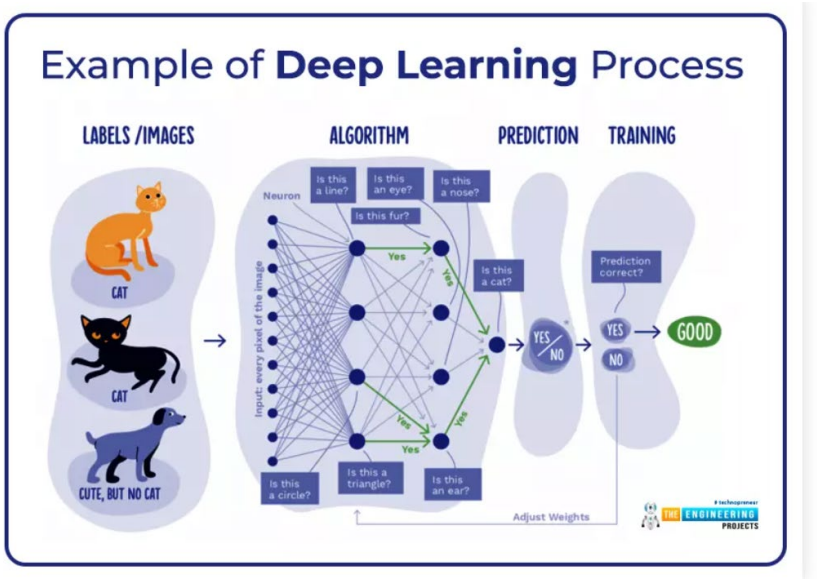
Source: AI21 Labs.

Exhibit 2 AI21 Labs Funding Rounds in $ Millions

Date	Transaction Name	Investors	Money Raised
Jul 12, 2022	Series B	Pitango VC (Israel), TPY Capital (Israel), Ahren Innovation Capital (UK), Mark Leslie (US), Walden Catalyst (US)	64
Nov 1, 2021	Venture Round	Walden Catalyst (US)	20
Jan 26, 2021	Non-Equity Assistance	Google for Startups	N/A
Nov 20, 2020	Series A	Pitango VC (Israel), TPY Capital (Israel)	34.5
Jan 1, 2019	Seed Round	Pitango VC (Israel), TPY Capital (Israel)	N/A
Jan 1, 2018	Seed Round	N.A.	3.5

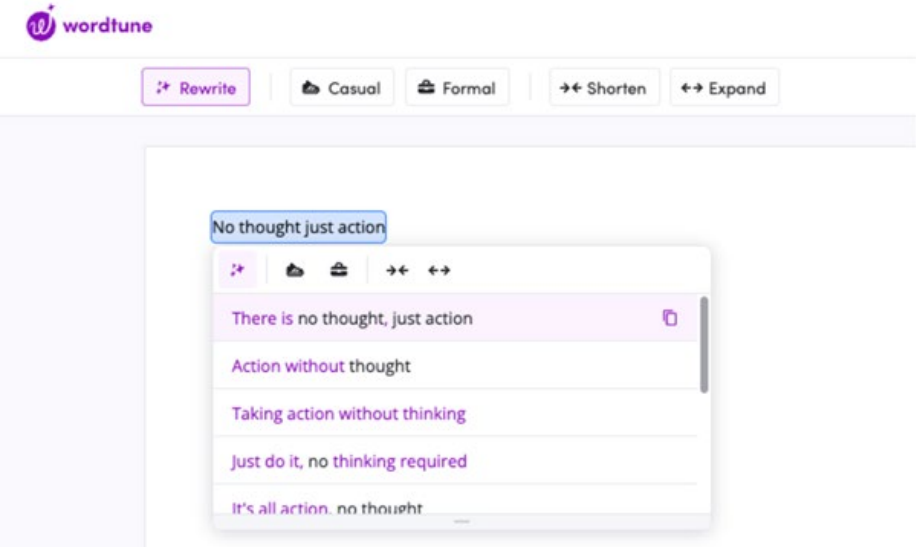
Source: Developed by case writers based on Crunchbase, "AI21 Labs," https://www.crunchbase.com/organization/ai21/company_financials

Exhibit 3 Example of Deep Learning Process



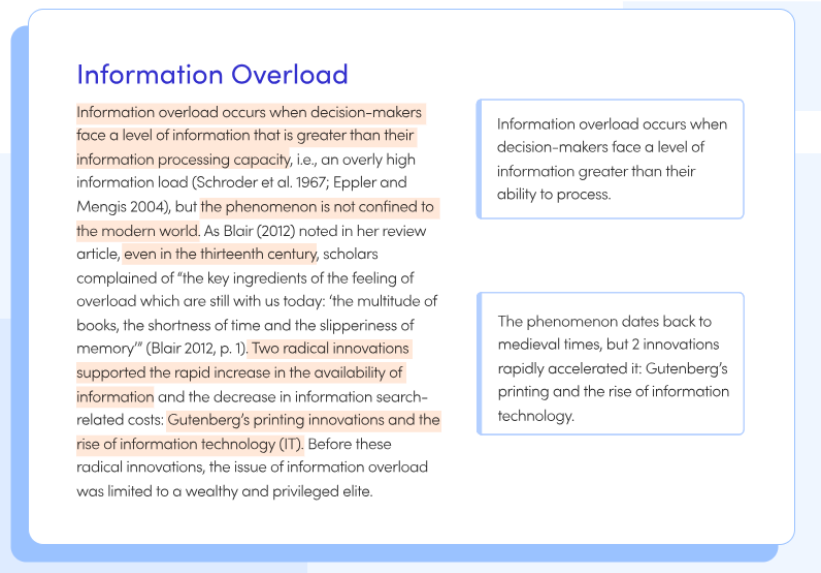
Source: David Watson, "Introduction to Deep Learning," Post on blog, The Engineering Projects, November 22, 2022, https://www.theengineeringprojects.com/2022/11/introduction-to-deep-learning.html, accessed July 2023.

Exhibit 4 Wordtune Screenshot



Source: Company's website, https://www.wordtune.com/

Exhibit 5 Wordtune Read Examples



Source: Company's website, https://www.wordtune.com/read

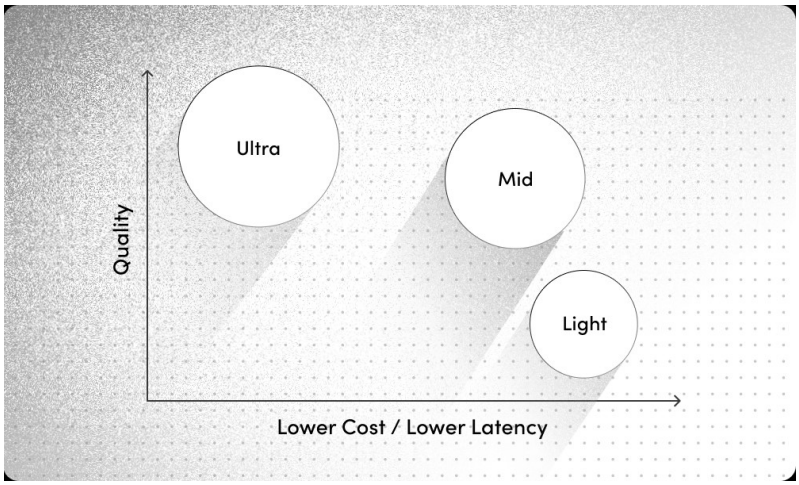
Exhibit 6 AI21 Labs Prices

Product Name	Free Service	Price	API Access	Fine-Tuning
Wordtune	Yes	Premium: $9.99/month	Yes	N/A
Wordtune Read	Yes	Premium: $9.99/month	Yes	N/A
Wordtune Spices	Yes	N/A	Yes	N/A
AI21 Studio	Yes	$29/month	Yes	N/A
Jurassic-2				
Ultra	No	$0.015/1k tokens/~750 words	Yes	$3/MB/epoch one-time fee
Mid	No	$0.01/1k tokens/~750 words	Yes	$0.5/MB/epoch one-time fee
Light	No	$0.0003/1k tokens/~750 words	Yes	$0.1/MB/epoch one-time fee
Jurassic-2 Task specific APIs				
Paraphrase	No	$0.001/API request	Yes	N/A
Summarize	No	$0.005/API request	Yes	N/A
Grammatical Error Corrections	No	$0.0005/API request	Yes	N/A
Text Improvements	No	$0.0005/API request	Yes	N/A
Text Segmentation	No	$0.001/API request	Yes	N/A
Contextual Answers	No	$0.005/API request	Yes	N/A

Source: Developed by case writers based on AI21 data.

Note: Epoch described the process of going through a training dataset one time.

Exhibit 7 Jurassic-2 Models



Source: Company's website, at https://www.ai21.com/blog/simplifying-our-jurassic-2-offering.

Exhibit 8 Key Players in the Field of LLM Development, data as of 2023

Company	Year Founded	Valuation	Funding	Active Users	Country of Origin
AI21 Labs	2017	$664 m	$128.5 m	2 million (B2C)	Israel
Writer	2020	N/A	$26 m	N/A	US
OpenAI	2015	$29 b	$11.3 bn	ChatGPT 100 million / month	US
Cohere	2019	$6 b	$414.9 m	N/A	Canada
Anthropic	2021	$4.1 b	$1.5 b	N/A	US

Source: Developed by case writers based on Crunchbase.

Exhibit 9 Key Players with Writing and Reading Applications

Company	Year Founded	Valuation	Funding	Active Users	Country of Origin
Grammarly	2009	$13 b	$400 m	+ 30 million /day	US
Quillbot	2017	N/A	$4.2 m	N/A	US
Jasper	2015	$1.5 b	$131 m	N/A	Canada

Source: Developed by case writers based on Crunchbase.

Endnotes

[1] "AI vs Machine Learning vs Deep Learning." https://www.talend.com/resources/ai-vs-machine-learning-vs-deep-learning, accessed July 25, 2023.

[2] Roose, Kevin. "How Does Chat GPT Really Work?" The New York Times, March 28, 2023 (updated April 4, 2023), https://www.nytimes.com/2023/03/28/technology/ai-chatbots-chatgpt-bing-bard-llm.html; https://bernardmarr.com/a-simple-guide-to-the-history-of-generative-ai/#:~:text=The%20Birth%20of%20Generative%20AI&text=When%20scientists%20and%20researchers%20introduced,of%20data%20based%20on%20input. Accessed July 25, 2023.

[3] Vaswani, Ashish et al. "Attention is All You Need." https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf, accessed July 17, 2023.

[4] Kerner, Sean Michael. "Large Language Model (LLM)." https://www.techtarget.com/whatis/definition/large-language-model-LLM, accessed July 17, 2023.

[5] Halpern, Sue. "What We Still Don't Know About How AI is Trained." The New Yorker, March 28, 2023. https://www.newyorker.com/news/daily-comment/what-we-still-dont-know-about-how-ai-is-trained, accessed July 25, 2023.

[6] OpenAI. "GPT-4 System Card." March 23, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf

[7] CTech. "AI21 Labs Launches AI-Based Tool That Summarizes a Document Within Seconds." Calcalist, November 16, 2021, https://www.calcalistech.com/ctech/articles/0,7340,L-3922633,00.html, accessed July 25, 2023.

[8] Chen, Brian X. "A.I. Bots Can't Report This Column. But They Can Improve It." The New York Times, February 1, 2023, https://www.nytimes.com/2023/02/01/technology/personaltech/chatgpt-ai-bots-editing.html, accessed July 25, 2023.

[9] https://www.ai21.com/blog/announcing-ai21-studio-and-jurassic-1, accessed July 25, 2023.

[10] Lieber, Opher, Or Sharir, Barak Lenz, and Yoav Shoham. "Jurassic-1: Technical Details and Evaluation." 2021, https://assets-global.website-files.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf

[11] Geifman, Yonatan. "How to Scale Generative AI Without Hurting the Bottom Line." June 7, 2023. https://www.spiceworks.com/tech/artificial-intelligence/guest-article/hidden-costs-of-generative-ai, accessed July 25, 2023.

[12] https://twitter.com/sama/status/1599671496636780546?lang=en, accessed July 25, 2023.

[13] Bastian, Matthias. "GPT-4 Has More Than A billion Parameters." https://the-decoder.com/gpt-4-has-a-trillion-parameters/#:~:text=Further%20details%20on%20GPT%2D4's,Mixture%20of%20Experts%20(MoE), accessed July 17, 2023.

[14] "Introducing the Data Flywheel." https://pages.awscloud.com/EMEA-Data-Flywheel.html?nc1=f_ls#business_momentum, accessed July 27, 2023.

[15] https://www.ai21.com/blog/simplifying-our-jurassic-2-offering#:~:text=Light%3A%20Jurassic%2D2%20Light%20is,sentiment%20analysis%2C%20and%20keyword%20extraction. Accessed July 25, 2023.

[16] "What are Models?" https://learn.microsoft.com/en-us/semantic-kernel/prompt-engineering/llm-models, accessed July 25, 2023.

[17] "Will the Estimated Training cost of GPT-4 be over $50M?"https://manifold.markets/BionicD0LPH1N/will-the-estimated-training-cost-of, July 17, 2023.

[18] https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai, accessed July 17, 2023.

[19] https://www.cbinsights.com/research/generative-ai-startups-market-map/?utm_source=CB+Insights+Newsletter&utm_campaign=3fd0923bd0-newsletter_general_Monday_2023_07_17&utm_medium=email&utm_term=0_9dc0513989-3fd0923bd0-90925801, accessed July 17, 2023.

[20] "ChatGPT and large language models: what's the risk?", National Cyber Security Centre, https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk, accessed July 25, 2023.

[21] "Companies That Have Banned ChatGPT." https://jaxon.ai/list-of-companies-that-have-banned-chatgpt/, accessed July 17, 2023.

[22] O'Donnell, Bob. "Amazon Debuts Bedrock, a New Cloud Service for AI-Generated Text and Images." TechSpot, May 4, 2023. https://www.techspot.com/news/98565-amazon-debuts-bedrock-new-cloud-service-ai-generated.html, accessed July 25, 2023.

[23] https://crfm.stanford.edu/helm/latest, accessed July 25, 2023.

[24] Spiro, James. "AI21 Labs Looking to Challenge OpenAI in Language Model Race." Calcalist, March 9, 2023, https://www.calcalistech.com/ctechnews/article/skkwflpjh, accessed July 25, 2023.

[25] Heaven, Will Douglas. "The Inside Story of How ChatGPT Was Built From the People Who Made It." MIT Technology Review, March 3, 2023. https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai, accessed July 25, 2023. https://www.makeuseof.com/gpt-models-explained-and-compared/

[26] https://openai.com/blog/chatgpt

[27] Sharma, Mukul. "How Chat GPT Overtook 'Dumb as Rock' Voice Assistants Alexa and Siri. WION (World is One News), March 16, 2023. https://www.wionews.com/technology/explained-how-chatgpt-overtook-dumb-as-rock-voice-assistants-alexa-and-siri-572514, accessed July 25, 2023.

[28] Metz, Cade and Karen Weise. "Microsoft to Invest $10 Billion in OpenAI, the Creator of ChatGPT." The New York Times, January 23, 2023. https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html, accessed July 25, 2023. https://www.businessinsider.com/how-much-money-does-chatgpt-openai-make-2023-8?utm_medium=newsletter&email=dyoffie%40hbs.edu&x=18ac8b84413b10e6022966cc2ea2e9a1c5ddd11a407066c47b495fcd3d86cdb5&utm_source=Sailthru&utm_campaign=Insider%20Today%2C%20August%2031%2C%202023&utm_term=INSIDER%20TODAY%20SEND%20LIST%20-%20ALL%20ENGAGED.

[29] https://www.microsoft.com/en-us/microsoft-365/blog/2023/03/16/introducing-microsoft-365-copilot-a-whole-new-way-to-work/, accessed July 17, 2023.

[30] Fried Ina and Ryan Heath. "Microsoft Puts a Price Tag on AI for Business." Axios, https://www.axios.com/2023/07/19/microsoft-price-tag-ai-business-office, accessed July 21, 2023.

[31] Khan, Imad. "ChatGPT" Cause 'Code Red' at Google, Report Says. https://www.cnet.com/tech/services-and-software/chatgpt-caused-code-red-at-google-report-says/, accessed July 17, 2023.

[32] James, Vincent. "Google's AI Chatbot Bard Makes Factual Error in First Demo." TheVerge, https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo, accessed July 17, 2023.

[33] Google, "What's ahead for Bard: More Global, More Visual, More Integrated,"Blog Google, https://blog.google/technology/ai/google-bard-updates-io-2023/, accessed August 2023

[34] https://ai.facebook.com/blog/large-language-model-llama-meta-ai/; https://thenewstack.io/why-open-source-developers-are-using-llama-metas-ai-model/, accessed July 25, 2023.

[35] https://www.crunchbase.com/organization/anthropic

[36] https://www.anthropic.com/product

[37] https://www.crunchbase.com/organization/cohere-82b8

[38] https://cohere.com/

[39] https://www.crunchbase.com/organization/writer

[40] Mishra, Prakhar. "Top 9 Open-Source NLP Projects." https://medium.com/mlearning-ai/top-8-open-source-nlp-projects-5b2e4138118a, accessed July 25, 2023.

[41] https://medium.com/@mparekh/meta-gets-ready-to-rumble-d5162f0defa2, Accessed July 25, 2023.

[42] Rosenblatt, Joel and Jack Clark. "Google's Android Generates $31 Billion Revenue, Oracle Says. https://www.bloomberg.com/news/articles/2016-01-21/google-s-android-generates-31-billion-revenue-oracle-says-ijor8hvt, accessed July 18, 2023.

[43] https://www.semianalysis.com/p/google-we-have-no-moat-and-neither, accessed July 18, 2023.

[44] https://www.grammarly.com/blog/how-does-grammarly-work/

[45] https://www.grammarly.com/

[46] https://www.crunchbase.com/organization/Jasper

[47] https://www.jasper.ai

[48] Wiggers, Kyle. "What is Auto-GPT and Why Does it Matter?" TechCrunch, April 22, 2023. https://techcrunch.com/2023/04/22/what-is-auto-gpt-and-why-does-it-matter/Accessed July 25, 2023.