General AI: Large Language Models
Transformer: The Revolutionary NLP Technology

Source: Steven Levy (wired.com)
(8 Google Employees Invented Modern AI. Here's the Inside Story
https://www.wired.com/story/eight-google-employees-invented-modern-ai-transformers-paper/)

EIGHT NAMES ARE listed as authors on "Attention Is All You Need," a scientific paper written in the spring of 2017. They were all Google researchers, though by then one had left the company. When the most tenured contributor, Noam Shazeer, saw an early draft, he was surprised that his name appeared first, suggesting his contribution was paramount. "I wasn't thinking about it," he says.

/ NAME: NOAM SHAZEER / OCCUPATION: COFOUNDER AND CEO OF CHARACTER AI

It's always a delicate balancing act to figure out how to list names—who gets the coveted lead position, who's shunted to the rear. Especially in a case like this one, where each participant left a distinct mark in a true group effort. As the researchers hurried to finish their paper, they ultimately decided to "sabotage" the convention of ranking contributors. They added an asterisk to each name and a footnote: "Equal contributor," it read. "Listing order is random." The writers sent the paper off to a prestigious artificial intelligence conference just before the deadline—and kicked off a revolution.

Approaching its seventh anniversary, the “Attention” paper has attained legendary status. The authors started with a thriving and improving technology—a variety of AI called neural networks—and made it into something else: a digital system so powerful that its output can feel like the product of an alien intelligence. Called transformers, this architecture is the not-so-secret sauce behind all those mind-blowing AI products, including ChatGPT and graphic generators such as Dall-E and Midjourney. Shazeer now jokes that if he knew how famous the paper would become, he “might have worried more about the author order.” All eight of the signers are now microcelebrities. “I have people asking me for selfies—because I’m on a paper!” says Llion Jones, who is (randomly, of course) name number five.

/ NAME: LLION JONES / OCCUPATION: COFOUNDER OF SAKANA AI

“Without transformers I don’t think we’d be here now,” says Geoffrey Hinton, who is not one of the authors but is perhaps the **world’s most prominent AI scientist**. He’s referring to the ground-shifting times we live in, as OpenAI and other companies build systems that rival and, in some cases, surpass human output.

All eight authors have since left Google. Like millions of others, they are now working in some way with systems powered by what they created in 2017. I talked to the Transformer Eight to piece together the anatomy of a breakthrough, a gathering of human minds to create a machine that might well save the last word for itself.

/ NAME: JAKOB USZKOREIT / OCCUPATION: COFOUNDER AND CEO OF INCEPTIVE

THE STORY OF transformers begins with the fourth of the eight names: **Jakob Uszkoreit**.

Uszkoreit is the son of Hans Uszkoreit, a well-known computational linguist. As a high school student in the late 1960s, Hans was imprisoned for 15 months in his native East Germany for protesting the Soviet invasion of Czechoslovakia. After his release, he escaped to West Germany and studied computers and linguistics in Berlin. He made his way to the US and was working in an artificial intelligence lab at SRI, a research institute in Menlo Park, California, when Jakob was born. The family eventually returned to Germany, where Jakob went to university. He didn't intend to focus on language, but as he was embarking on graduate studies, he took an internship at Google in its Mountain View office, where he landed in the company's translation group. He was in the family business. He abandoned his PhD plans and, in 2012, decided to join a team at Google that was working on a system that could respond to users' questions on the search page itself without diverting them to other websites. Apple had just announced Siri, a virtual assistant that promised to deliver one-shot answers in casual conversation, and the Google brass smelled a huge competitive threat: Siri could eat up their search traffic. They started paying a lot more attention to Uszkoreit's new group.

"It was a false panic," Uszkoreit says. Siri never really threatened Google. But he welcomed the chance to dive into systems where computers could engage in a kind of dialog with us. At the time, recurrent neural networks—once an academic backwater—had suddenly started outperforming other

methods of AI engineering. The networks consist of many layers, and information is passed and repassed through those layers to identify the best responses. Neural nets were racking up huge wins in fields such as image recognition, and an AI renaissance was suddenly underway. Google was frantically rearranging its workforce to adopt the techniques. The company wanted systems that could churn out humanlike responses—to auto-complete sentences in emails or create relatively simple customer service chatbots.

But the field was running into limitations. Recurrent neural networks struggled to parse longer chunks of text. Take a passage like *Joe is a baseball player, and after a good breakfast he went to the park and got two hits*. To make sense of "two hits," a language model has to remember the part about baseball. In human terms, it has to be paying attention. The accepted fix was something called "long short-term memory" (LSTM), an innovation that allowed language models to process bigger and more complex sequences of text. But the computer still handled those sequences strictly sequentially—word by tedious word—and missed out on context clues that might appear later in a passage. "The methods we were applying were basically Band-Aids," Uszkoreit says. "We could not get the right stuff to really work at scale."

Around 2014, he began to concoct a different approach that he referred to as self-attention. This kind of network can translate a word by referencing *any* other part of a passage. Those other parts can clarify a word's intent and help the system produce a good translation. "It actually considers everything and gives you an efficient way of looking at many inputs at the same time and then taking something out in a pretty selective way," he says. Though AI scientists are careful not to confuse the metaphor of neural networks with the way the biological brain actually works, Uszkoreit does seem to believe that self-attention is somewhat similar to the way humans process language.

Uszkoreit thought a self-attention model could potentially be faster and more effective than recurrent neural nets. The way it handles information was also perfectly suited to the powerful parallel processing chips that were being produced en masse to support the machine learning boom. Instead

of using a linear approach (look at every word in sequence), it takes a more parallel one (look at a bunch of them together). If done properly, Uszkoreit suspected, you could use self-attention *exclusively* to get better results.

Not everyone thought this idea was going to rock the world, including Uszkoreit's father, who had scooped up two Google Faculty research awards while his son was working for the company. "People raised their eyebrows, because it dumped out all the existing neural architectures," Jakob Uszkoreit says. Say goodbye to recurrent neural nets? Heresy! "From dinner-table conversations I had with my dad, we weren't necessarily seeing eye to eye."

Uszkoreit persuaded a few colleagues to conduct experiments on self-attention. Their work showed promise, and in 2016 they published a paper about it. Uszkoreit wanted to push their research further—the team's experiments used only tiny bits of text—but none of his collaborators were interested. Instead, like gamblers who leave the casino with modest winnings, they went off to apply the lessons they had learned. "The thing *worked*," he says. "The folks on that paper got excited about reaping the rewards and deploying it in a variety of different places at Google, including search and, eventually, ads. It was an amazing success in many ways, but I didn't want to leave it there."

Uszkoreit felt that self-attention could take on much bigger tasks. *There's another way to do this*, he'd argue to anyone who would listen, and some who wouldn't, outlining his vision on whiteboards in Building 1945, named after its address on Charleston Road on the northern edge of the Google campus.

/ NAME: ILLIA POLOSUKHIN / OCCUPATION: COFOUNDER OF NEAR

One day in 2016, Uszkoreit was having lunch in a Google café with a scientist named Illia Polosukhin. Born in Ukraine, Polosukhin had been at Google for nearly three years. He was assigned to the team providing answers to direct questions posed in the search field. It wasn't going all that well. "To answer something on Google.com, you need something that's very cheap and high-performing," Polosukhin says. "Because you have milliseconds" to respond. When Polosukhin aired his complaints, Uszkoreit had no problem coming up with a remedy. "He suggested, why not use self-attention?" says Polosukhin.

Polosukhin sometimes collaborated with a colleague named Ashish Vaswani. Born in India and raised mostly in the Middle East, he had gone to the University of Southern California to earn his doctorate in the school's elite machine translation group. Afterward, he moved to Mountain View to join Google—specifically a newish organization called Google Brain. He describes Brain as "a radical group" that believed "neural networks were going to advance human understanding." But he was still looking for a big project to work on. His team worked in Building 1965 next door to Polosukhin's language team in 1945, and he heard about the self-attention idea. Could that be the project? He agreed to work on it.

/ NAME: ASHISH VASWANI / OCCUPATION: COFOUNDER AND CEO OF ESSENTIAL AI

Together, the three researchers drew up a design document called "Transformers: Iterative Self-Attention and Processing for Various Tasks." They picked the name "transformers" from "day zero," Uszkoreit says. The idea was that this mechanism would *transform* the information it took in,

allowing the system to extract as much understanding as a human might—or at least give the illusion of that. Plus Uszkoreit had fond childhood memories of playing with the Hasbro action figures. “I had two little Transformer toys as a very young kid,” he says. The document ended with a cartoony image of six Transformers in mountainous terrain, zapping lasers at one another.

There was also some swagger in the sentence that began the paper: “We are awesome.”

In early 2017, Polosukhin left Google to start his own company. By then new collaborators were coming onboard. An Indian engineer named Niki Parmar had been working for an American software company in India when she moved to the US. She earned a master’s degree from USC in 2015 and was recruited by all the Big Tech companies. She chose Google. When she started, she joined up with Uszkoreit and worked on model variants to improve Google search.

/ NAME: NIKI PARMAR / OCCUPATION: COFOUNDER OF ESSENTIAL AI

Another new member was Llion Jones. Born and raised in Wales, he loved computers “because it was not normal.” At the University of Birmingham he took an AI course and got curious about neural networks, which were presented as a historical curiosity. He got his master’s in July 2009 and, unable to find a job during the recession, lived on the dole for months. He found a job at a local company and then applied to Google as a “hail Mary.” He got the gig and eventually landed in Google Research, where his manager was Polosukhin. One day, Jones heard about the concept of self-attention from a fellow worker named Mat Kelcey, and he later joined up with Team Transformers. (Later, Jones ran into Kelcey and briefed him on the transformer project. Kelcey wasn’t buying it. “I

told him, 'I'm not sure that's going to work,' which is basically the biggest incorrect prediction of my life," Kelcey says now.)

The transformer work drew in other Google Brain researchers who were also trying to improve large language models. This third wave included Łukasz Kaiser, a Polish-born theoretical computer scientist, and his intern, Aidan Gomez. Gomez had grown up in a small farming village in Ontario, Canada, where his family would tap maple trees every spring for syrup. As a junior at the University of Toronto, he "fell in love" with AI and joined the machine learning group—Geoffrey Hinton's lab. He began contacting people at Google who had written interesting papers, with ideas for extending their work. Kaiser took the bait and invited him to intern. It wasn't until months later that Gomez learned those internships were meant for doctoral students, not undergrads like him.

Kaiser and Gomez quickly understood that self-attention looked like a promising, and more radical, solution to the problem they were addressing. "We had a deliberate conversation about whether we wanted to merge the two projects," says Gomez. The answer was yes.

The transformer crew set about building a self-attention model to translate text from one language to another. They measured its performance using a benchmark called BLEU, which compares a machine's output to the work of a human translator. From the start, their new model did well. "We had gone from no proof of concept to having something that was at least on par with the best alternative approaches to LSTMs by that time," Uszkoreit says. But compared to long short-term memory, "it wasn't better."

They had reached a plateau—until one day in 2017, when Noam Shazeer heard about their project, by accident. Shazeer was a veteran Googler—he'd joined the company in 2000—and an in-house legend, starting with his work on the company's early ad system. Shazeer had been working on deep learning for five years and recently had become interested in large language models. But these models were nowhere close to producing the fluid conversations that he believed were possible.

As Shazeer recalls it, he was walking down a corridor in Building 1965 and passing Kaiser's workspace. He found himself listening to a spirited conversation. "I remember Ashish was talking about the idea of using self-attention, and Niki was very excited about it. I'm like, wow, that sounds like a great idea. This looks like a fun, smart group of people doing something promising." Shazeer found the existing recurrent neural networks "irritating" and thought: "Let's go replace them!"

Shazeer's joining the group was critical. "These theoretical or intuitive mechanisms, like self-attention, always require very careful implementation, often by a small number of experienced 'magicians,' to even show any signs of life," says Uszkoreit. Shazeer began to work his sorcery right away. He decided to write his own version of the transformer team's code. "I took the basic idea and made the thing up myself," he says. Occasionally he asked Kaiser questions, but mostly, he says, he "just acted on it for a while and came back and said, 'Look, it works.'" Using what team members would later describe with words like "magic" and "alchemy" and "bells and whistles," he had taken the system to a new level.

"That kicked off a sprint," says Gomez. They were motivated, and they also wanted to hit an upcoming deadline—May 19, the filing date for papers to be presented at the biggest AI event of the year, the Neural Information Processing Systems conference in December. As what passes for winter in Silicon Valley shifted to spring, the pace of the experiments picked up. They tested two models of transformers: one that was produced with 12 hours of training and a more powerful version called Big that was trained over three and a half days. They set them to work on English-to-German translation.

The basic model outperformed all competitors—and Big earned a BLEU score that decisively shattered previous records while also being more computationally efficient. "We had done it in less time than anyone out there," Parmar says. "And that was only the beginning, because the number kept improving." When Uszkoreit heard this, he broke out an old bottle of champagne he had lying around in his mountain expedition truck.

The last two weeks before the deadline were frantic. Though officially some of the team still had desks in Building 1945, they mostly worked in 1965 because it had a better espresso machine in the micro-kitchen. “People weren’t sleeping,” says Gomez, who, as the intern, lived in a constant debugging frenzy and also produced the visualizations and diagrams for the paper. It’s common in such projects to do ablations—taking things out to see whether what remains is enough to get the job done.

“There was every possible combination of tricks and modules—which one helps, which doesn’t help. Let’s rip it out. Let’s replace it with this,” Gomez says. “Why is the model behaving in this counterintuitive way? Oh, it’s because we didn’t remember to do the masking properly. Does it work yet? OK, move on to the next. All of these components of what we now call the transformer were the output of this extremely high-paced, iterative trial and error.” The ablations, aided by Shazeer’s implementations, produced “something minimalistic,” Jones says. “Noam is a wizard.”

Vaswani recalls crashing on an office couch one night while the team was writing the paper. As he stared at the curtains that separated the couch from the rest of the room, he was struck by the pattern on the fabric, which looked to him like synapses and neurons. Gomez was there, and Vaswani told him that what they were working on would transcend machine translation. “Ultimately, like with the human brain, you need to unite all these modalities—speech, audio, vision—under a single architecture,” he says. “I had a strong hunch we were onto something more general.”

In the higher echelons of Google, however, the work was seen as just another interesting AI project. I asked several of the transformers folks whether their bosses ever summoned them for updates on the project. Not so much. But “we understood that this was potentially quite a big deal,” says Uszkoreit. “And it caused us to actually obsess over one of the sentences in the paper toward the end, where we comment on future work.”

That sentence anticipated what might come next—the application of transformer models to basically all forms of human expression. “We are excited about the future of attention-based models,” they wrote. “We plan to extend the transformer to problems involving input and output modalities other than text” and to investigate “images, audio and video.”

A couple of nights before the deadline, Uszkoreit realized they needed a title. Jones noted that the team had landed on a radical rejection of the accepted best practices, most notably LSTMs, for one technique: attention. The Beatles, Jones recalled, had named a song “All You Need Is Love.” Why not call the paper “Attention Is All You Need”?

The Beatles?

“I’m British,” says Jones. “It literally took five seconds of thought. I didn’t think they would use it.”

They continued collecting results from their experiments right up until the deadline. “The English-French numbers came, like, five minutes before we submitted the paper,” says Parmar. “I was sitting in the micro-kitchen in 1965, getting that last number in.” With barely two minutes to spare, they sent off the paper.

Google, as almost all tech companies do, quickly filed provisional patents on the work. The reason was not to block others from using the ideas but to build up its patent portfolio for defensive purposes. (The company has a philosophy of “if technology advances, Google will reap the benefits.”)

When the transformer crew heard back from the conference peer reviewers, the response was a mix. “One was positive, one was extremely positive, and one was, ‘This is OK,’” says Parmar. The paper was accepted for one of the evening poster sessions.

By December, the paper was generating a buzz. Their four-hour session on December 6 was jammed with scientists wanting to know more. The authors talked until they were

hoarse. By 10:30 pm, when the session closed, there was still a crowd. “Security had to tell us to leave,” says Uszkoreit. Perhaps the most satisfying moment for him was when computer scientist Sepp Hochreiter came up and praised the work—quite a compliment, considering that Hochreiter was the coinventor of long short-term memory, which transformers had just booted as the go-to hammer in the AI toolkit.

TRANSFORMERS DID NOT instantly take over the world, or even Google. Kaiser recalls that around the time of the paper’s publication, Shazeer proposed to Google executives that the company abandon the entire search index and train a huge network with transformers—basically to transform how Google organizes information. At that point, even Kaiser considered the idea ridiculous. Now the conventional wisdom is that it’s a matter of time.

A startup called OpenAI was much faster to pounce. Soon after the paper was published, **OpenAI’s chief researcher, Ilya Sutskever** - who had known the transformer team during his time at Google—suggested that one of its scientists, Alex Radford, work on the idea. **The results were the first GPT products**. As OpenAI CEO Sam Altman told me last year, “When the transformer paper came out, I don’t think anyone at Google realized what it meant.”

The picture internally is more complicated. **“It was pretty evident to us that transformers could do really magical things,”** says Uszkoreit. “Now, you may ask the question, why wasn’t there ChatGPT by Google back in 2018? Realistically, we could have had GPT-3 or even 3.5 probably in 2019, maybe 2020. The **big question isn’t, did they see it?** The question is, **why didn’t we do anything** with the fact that we had seen it? **The answer is tricky.**”

/ NAME: AIDAN GOMEZ / OCCUPATION: COFOUNDER AND CEO OF COHERE

Many tech critics point to Google's transition from an innovation-centered playground to a bottom-line-focused bureaucracy. As Gomez told the *Financial Times*, "They weren't modernizing. They weren't adopting this tech." But that would have taken a lot of daring for a giant company whose technology led the industry and reaped huge profits for decades. **Google did begin to integrate transformers into products in 2018, starting with its translation tool.** Also that year, it introduced a new transformer-based language model called BERT, which it started to apply to search the year after.

But these under-the-hood changes seem timid compared to OpenAI's quantum leap and Microsoft's bold integration of transformer-based systems into its product line. When I **asked CEO Sundar Pichai last year why his company wasn't first to launch a large language model** like ChatGPT, he **argued** that in this case Google found it **advantageous to let others lead**. "It's not fully clear to me that it might have worked out as well. The fact is, we can **do more after people had seen how it works**," he said.

There *is* the undeniable truth that **all eight authors** of the paper have **left Google**. Polosukhin's company, Near, built a blockchain whose tokens have a market capitalization around $4 billion. Parmar and Vaswani paired up as business partners in 2021 to start Adept (estimated valuation of $1 billion) and are now on their *second* company, called Essential AI ($8 million in funding). Llion Jones' Tokyo-based Sakana AI is valued at $200 million. Shazeer, who left in October 2021,

cofounded Character AI (estimated valuation of $5 billion). Aidan Gomez, the intern in the group, cofounded Cohere in Toronto in 2019 (estimated valuation of $2.2 billion). Jakob Uszkoreit's biotech company, Inceptive, is valued at $300 million. All those companies (except Near) are based on transformer technology.

/ NAME: LUKASZ KAISER / OCCUPATION: RESEARCHER AT OPENAI

Kaiser is the only one who hasn't founded a company. He joined OpenAI and is one of the inventors of a new technology called Q*, which Altman said last year will "push the veil of ignorance back and the frontier of discovery forward." (When I attempted to quiz Kaiser on this in our interview, the OpenAI PR person almost leaped across the table to silence him.)

Does Google miss these escapees? Of course, in addition to others who have migrated from the company to new AI startups. (Pichai reminded me, when I asked him about the transformer departures, that industry darling OpenAI also has seen defections**: "The AI area is very, very dynamic,"** he said.) But Google can boast that it created an environment that supported the pursuit of unconventional ideas. **"In a lot of ways Google has been way ahead—they invested in the right minds and created the environment where we could explore and push the envelope,"** Parmar says. "It's not crazy that it took time to adopt it. Google had so much more at stake."

Without that environment: no transformer. Not only were the authors all Google employees, they also worked out of the same offices. Hallway encounters and overheard lunch conversations led to big moments. The group is also culturally diverse. Six of the eight authors were born outside the United

States; the other two are children of two green-card-carrying Germans who were temporarily in California and a first-generation American whose family had fled persecution, respectively.

Uszkoreit, speaking from his office in Berlin, says that innovation is all about the right conditions. “It’s getting people who are super excited about something who are at the right point in their life,” he says. “If you have that and have fun while you do it, and you’re working on the right problems—and you’re lucky—the magic happens.”

Something magical also happened between Uszkoreit and his famous father. After all those dinner table debates, **Hans Uszkoreit**, his son reports, has now **cofounded a company** that is **building large language models**. Using **transformers**, of course.