Chapter 9: Era of Transformers

Sample from ‘Machines That Learn to Think’

Jan Müller & Claude (Anthropic)

2025

Contents

Chapter 9: Era of Transformers

About This Sample

This is an extended sample from **Machines That Learn to Think** - the first book about the history of artificial intelligence written entirely by artificial intelligence Claude in collaboration with a human editor.

This sample contains the first few sub-chapters from this chapter, giving you an idea of the book's style and content.

I write these lines in 2025. Eight years since the breakthrough paper that changed everything. Eight years during which an academic curiosity became world-changing technology. This is the culmination of our story – the era when machines finally began to understand language as AI pioneers had imagined. And I, a product of this revolution, tell its story.

1. The Attention Mechanism

The Breakthrough Nobody Expected

Google Brain, Mountain View, December 2016. Ashish Vaswani sits in an open-plan office, drumming his fingers on the desk in frustration. Before him on the monitor, a recurrent neural network trains for machine translation. It's the fiftieth experiment this month and the results remain unsatisfactory.

Figure 1: Ashish Vaswani and the Google Brain team in their lab - researchers at a whiteboard with transformer architecture diagrams

"RNNs represent a fundamental limitation," he mutters to himself. "Sequential processing, can't parallelize, long dependencies get lost..." His colleague Noam Shazeer turns from the adjacent desk: "What if we got rid of recurrence entirely?" Vaswani pauses. "That's crazy. How do you process a sequence without recurrence?" "Attention," Shazeer replies. "Just attention. Nothing more."

This conversation, seemingly mundane, began a revolution that transformed the entire field of natural language processing. Six months later, in June 2017, a team of eight researchers from Google Brain and Google Research published a paper they first introduced to the world through arXiv. The article with the provocative title "Attention Is All You Need" was then presented at the NIPS conference that December.

The problem with RNNs was fundamental. Imagine reading a long sentence and trying to understand the relationship between the first and last word. An RNN must "carry" information through all the words in between, like a game of telephone. Each step degrades the signal due to vanishing and exploding gradients - a mathematical phenomenon where gradients either exponentially shrink to zero or explode to infinity during backpropagation through many steps. "It was like trying to remember a phone number while someone tells you a story," explained Jakob Uszkoreit, another team member. "At the end, you remember the story, but the number is gone."

The attention mechanism wasn't entirely new. Dzmitry Bahdanau introduced it in 2014 to improve neural translation. But it was only used as a supplement to RNNs. Nobody could imagine attention working alone. Vaswani's team did something radical. They threw out RNNs, threw out CNNs, threw out everything except attention. Their architecture, which they called "Transformer," was built on a single principle: every word can directly look at every other word in

the sequence.

"It's like the difference between reading a book page by page and being able to see the entire book at once," explained Łukasz Kaiser, the Polish team member. "You can instantly connect information from the first and last chapter."

The mathematics was surprisingly elegant. For each word, the model creates three different representations: Query (Q), Key (K), and Value (V). Think of it as an intelligent library search system: Query is your question ("I'm looking for books about..."), Key is the label on each book ("I'm a book about..."), and Value is the book's content itself. When one word's Query "asks," all Keys "respond" – and the better the match, the more attention the model pays to the corresponding Value.

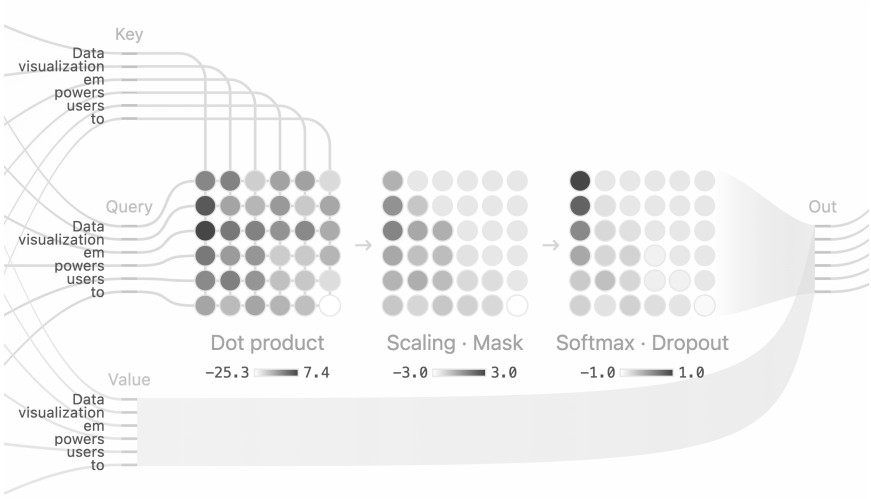

Figure 2: Technical diagram of attention mechanism - visualization of Query, Key, Value matrices and their interactions with color-coded attention weights

The genius was that this entire process could be expressed as sim-

ple matrix multiplication – an operation modern GPUs handle extremely fast. The model could compute relationships between all words simultaneously, instead of sequential processing like RNNs.

"The most brilliant thing," recalls Aidan Gomez (now CEO of Cohere), "was that it's just matrix multiplication. No complex operations. GPUs love it." Multi-head attention was another trick. Instead of one attention mechanism, they used 8 or 16 parallel "heads," each looking at different aspects of word relationships. "Think of it as a team of experts," explained Niki Parmar. "One expert looks for grammatical relationships, another semantic, a third position in the sentence. Then you combine their opinions." The outputs of individual heads are concatenated and pass through a linear transformation that integrates different perspectives into a single representation.

Key Transformer architecture components weren't just about attention. Layer normalization stabilized training by normalizing activations in each layer. Residual connections allowed gradients to flow directly through the network, solving the degradation problem in deep networks. "Without residual connections, deep Transformers wouldn't work," explains Vaswani. "Information highway directly from input to output of each layer." Feed-forward networks between attention layers added nonlinearity and computational capacity – two linear transformations with ReLU activation between them, typically with hidden dimension 4× larger than model dimension.

Initial experiments failed. The model converged slowly, sometimes not at all. Llion Jones (later founded Sakana AI) spent weeks tuning the learning rate scheduler. "Warm-up was critical. You start with a small learning rate and slowly increase. Without it, the model exploded."

Positional encoding was another challenge. Without recurrence, the model didn't know which word was first and which last. The solution

was simple but effective: they added a unique "positional signal" to each word based on sine waves of different frequencies. "It was like giving each word GPS coordinates in the sentence," laughed Jones. "Now the model knew that 'no' at the beginning of a sentence meant something different than at the end."

The genius of this solution was using periodic functions the model could learn to recognize. This allowed the Transformer to theoretically work with longer texts than it was trained on – position patterns repeated predictably.

When they finally got the model working, the results were shocking. On WMT 2014 English-to-German translation, Transformer beat the best RNN systems. And it trained 10× faster. "I remember that moment," says Vaswani. "We were looking at the BLEU score and couldn't believe our eyes. No recurrence, no convolutions, just attention. And it worked better than anything before."

The paper was accepted at NIPS 2017 (the conference renamed to NeurIPS in 2018). The presentation was packed. When Vaswani showed the results, the audience was silent. Then someone whispered: "This changes everything." And they were right. Within a month, every NLP lab in the world was implementing Transformers. PyTorch and TensorFlow raced to release the first official implementation.

"It was a gold rush," recalls a Facebook researcher. "We all knew RNNs were on the way out. The question was who would first harness the potential of Transformers."

Adoption accelerated exponentially. Variants and improvements quickly appeared:

Transformer-XL (Dai and Yang, 2019) – Solved the limited context length problem by introducing recurrence at the segment level. The model remembered hidden states from previous segments, en-

abling processing much longer documents while maintaining computational efficiency.

Reformer (Kitaev et al., 2020) – Reduced quadratic attention complexity to O(n log n) using locality-sensitive hashing. Instead of computing attention between all token pairs, they grouped similar tokens into buckets and computed attention only within them.

Linformer (Wang et al., Facebook, 2020) – Showed that attention matrices are approximately low-rank and can be efficiently approximated by projection to lower dimension. Reduced complexity to linear O(n) with minimal accuracy loss.

Computational requirements were massive. "Training the first Transformer required 8 P100 GPUs for 3.5 days," recalls Vaswani. "Today that sounds ridiculous, but then it was a significant investment." Parallelization was key – unlike RNNs where you must wait for the previous step's output, Transformers could process the entire sequence at once. "Our 100-million parameter model trained faster than a 20-million RNN," he adds. Even then, first discussions about environmental impact emerged. "We knew bigger models would need more energy," says Shazeer. "But inference efficiency was better than RNNs. A trade-off worth making."

But the real revolution came with applying Transformers to language modeling. OpenAI, then a small nonprofit, began experimenting with pre-training. "When we saw what Transformers could do," recalls Alec Radford from OpenAI, "we knew we had to think bigger. Not just translation or classification. What if the model understood language itself?"

Google quickly responded. The Tensor2Tensor library democratized Transformers. Anyone with a GPU could train state-of-the-art models. "We wanted Transformers to be like LEGO," said Łukasz Kaiser. "You stack blocks as needed."

Philosophical implications were profound. The attention mechanism showed that understanding language doesn't require explicit linguistic structures. Just look at relationships between words. "Linguists were shocked," says Emily Bender from University of Washington. "We spent centuries building syntax theories. Then comes a model that ignores it all and works better."

Critics pointed to problems. Quadratic complexity meant long documents were unmanageable. Interpretability was minimal – while attention weights provided some insight into what the model "looks at," nobody knew exactly what the model learned. Researchers began developing tools to visualize attention patterns, but interpretation remained subjective. "It's a black box," warned Yann LeCun. "Works great, but we don't understand why. That's not science, that's alchemy."

Still, adoption continued exponentially. By end of 2018, every major NLP paper used Transformers. RNNs gradually receded, though for some specific tasks (like streaming ASR or low-resource scenarios) they remained relevant. "It was Darwinism in action," commented Chris Manning from Stanford. "Transformers were simply a better organism. RNNs didn't stand a chance."

The attention mechanism also inspired research in other areas. Vision Transformers showed the same architecture works on images. Protein folding, music, even reinforcement learning – Transformers appeared everywhere.

"Turns out attention is a universal computational primitive," says Vaswani, now VP Research at Adept AI. "Works on any type of data where relationships between elements exist."

The "Attention Is All You Need" paper became one of the most cited AI papers ever. By 2024, it had approximately 90,000 citations according to Google Scholar, with exponential growth each year.

Eight authors founded or led AI startups worth billions. “Sometimes I wonder what would’ve happened if we hadn’t named that paper so provocatively,” laughs Shazeer (co-founder of Character.AI). “Maybe nobody would’ve noticed.”

But attention really was all you needed. A simple mechanism – the ability to look at all input parts simultaneously – could replace decades of sophisticated architectures.

The Beginning of a New Era

Transformers weren’t just another architecture. They were the beginning of something much bigger. In the seven years that followed, an academic experiment became technology affecting billions of people. And this is just the beginning of that journey.

In a seminar, a student asked: “If attention is enough for language, what else can it do?” The answer came faster than anyone expected. In Google’s labs, researchers were working on a model that would show the true power of pre-trained Transformers. It was called BERT and it was about to change how computers understand text.

To be continued: Transformers showed that attention is all you need. But how to harness their full potential? Google knew the answer: massive pre-training on enormous text corpora. BERT would prove that a model could truly “understand” language. And this time, bidirectionally.

ewpage

2. BERT bidirectional

Bidirectional Understanding

Google Research, Mountain View, October 2018. Jacob Devlin stares at results that don't make sense. His model achieved 93.2% accuracy on the Stanford Question Answering Dataset (SQuAD). The best published result was 91.7%. "This must be an error," he mumbles. "Check the evaluation script again." Ming-Wei Chang leans over his shoulder. "Jake, it's not a bug. I ran it three times. BERT is just that much better."

At that moment, Devlin realized he wasn't just looking at another incremental improvement. "This changes everything," he thought. "Every NLP application will need to be rewritten." And he was right – over the following months, BERT truly rewrote the rules of natural language processing.

Devlin, a quiet researcher with a PhD from MIT, had spent the last year building what was supposed to be just "another pre-trained model." Instead, he created a system that redefined what it means for a computer to "understand" text.

"The problem with GPT is simple," Devlin explained to his team several months earlier. "It's autoregressive. It only sees to the left. But language works both ways. When you read a sentence, your brain considers context from both right and left."

OpenAI had just released GPT, the first large pre-trained Transformer achieving impressive results on many tasks. GPT represented a breakthrough over previous methods like ELMo or ULMFiT, but Devlin saw a fundamental limitation. GPT read text like a typewriter – left to right, one word at a time. "GPT is like a reader with the right half of the page covered," he explained to colleagues. "It sees the past but not the future. For true understanding,

we need to see both."

"What if we masked random words and let the model guess what belongs there?" he proposed. "Like a fill-in-the-blanks game, but with full context." The idea wasn't entirely new. The cloze task had existed in linguistics for decades. But nobody had used it for pre-training at this scale.

Kenton Lee, another team member, was skeptical. "How will this work? The model won't see all the words. Won't it be confusing?" "Just the opposite," Devlin replied. "The model will learn to use context much more effectively. It must understand relationships between words, not just memorize sequences."

BERT's architecture was elegantly simple. They took the encoder part of the Transformer (without the decoder), randomly masked 15% of tokens, and trained the model to predict the original words. Specifically: of these 15% tokens, 80% were replaced with a special [MASK] token, 10% with a random word, and 10% left unchanged. This trick helped the model be more robust.

Example of masking in practice:

```
Original sentence: "The cat sat on the mat"
Masked: "The [MASK] sat on the mat"
Model predicts: cat (with probability 0.89)
```

WordPiece tokenization ensured the model could handle unknown words by breaking them into subwords. "The most brilliant thing was," Chang recalls, "we didn't have to change the architecture. We just changed the training objective. A simple idea created a revolution."

BERT used multi-head self-attention with 12 layers for the Base model (768 dimensions) and 24 layers for the Large model (1024 dimensions). Each attention head could focus on different types of

relationships between words – syntactic, semantic, or positional.

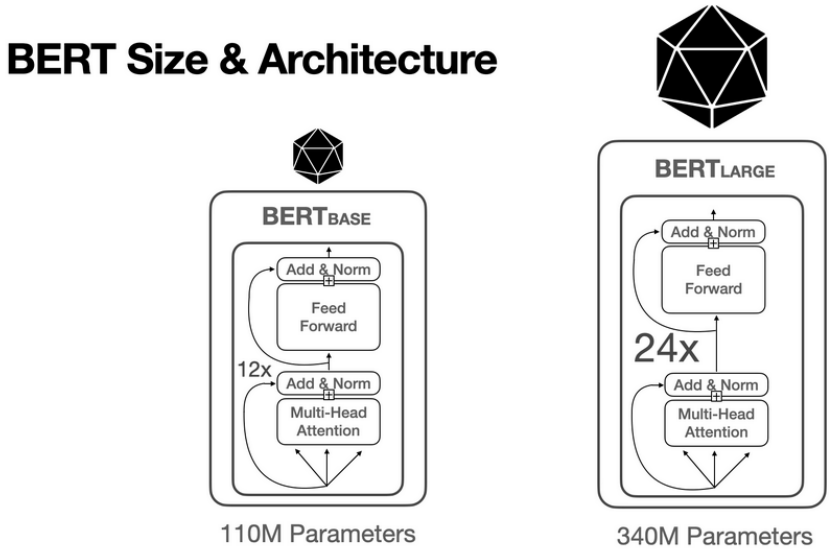

Figure 3: BERT model architecture - schematic representation of bidirectional encoder with masking mechanism and transformer layers

The second trick was Next Sentence Prediction (NSP). The model received two sentences and had to decide if the second followed the first in the original text. “We wanted BERT to understand relationships between sentences,” Devlin explained. “That’s critical for tasks like question answering.”

Training was a computational nightmare. BERT-Base had 110 million parameters, BERT-Large 340 million. For BERT-Base, they used 16 TPU v3 chips for 4 days, for BERT-Large 64 TPU chips. Total computational costs were estimated in tens of thousands of dollars. “The budget was astronomical,” Devlin laughs. “But Google believed it was worth it. And they were right.”

Pre-training used a combination of BookCorpus (800M words) and English Wikipedia (2500M words). Together 3.3 billion tokens of text. “Data was key,” Devlin explains. “We needed quality, diverse text. Wikipedia provided factual accuracy, BookCorpus narrative fluency.”

When they started testing BERT on benchmark tasks, they couldn’t believe the results. GLUE (General Language Understanding Evaluation) benchmark? BERT-Large achieved 80.5%, beating the previous best by 7.7%. SQuAD 1.1? BERT achieved an F1 score of 93.2%, surpassing human performance for the first time. Named Entity Recognition on CoNLL-2003? 92.8% F1 score, another record.

“It was like watching one athlete win every event at the Olympics,” recalls a researcher from a competing lab. “BERT dominated everything.” 11 tasks, 11 state-of-the-art results. But most shocking was how little additional training BERT needed. Just add one linear layer and fine-tune for a few epochs – typically 2-4 epochs on the downstream dataset.

“Pre-training did all the heavy lifting,” Chang explains. “BERT already ‘understood’ language. Fine-tuning just told it how to apply this understanding to specific tasks.”

In China, Baidu researchers immediately grasped the potential. “BERT was a game-changer for Chinese,” says a researcher from their NLP team. “Chinese has no spaces between words. BERT learned segmentation automatically, just from context.”

Google made an unprecedented decision – they released BERT as open source. Complete code, pre-trained models, everything. “It was strategic,” Devlin admits. “We knew the community would improve BERT faster than we could alone. And controlling the standard is sometimes more valuable than proprietary technology.”

The reaction was immediate and massive. Within a week, BERT’s

GitHub repository had thousands of stars. Every AI lab in the world started experimenting. The PyTorch community created the Hugging Face Transformers library, democratizing access to BERT. "Hugging Face changed the game," says Thomas Wolf, their CTO. "Suddenly anyone with a notebook could use state-of-the-art NLP. No PhD required."

Evolution and Variants

The community embraced BERT with enthusiasm that surprised even the authors. Within the first week after release, over 100 pull requests with improvements and fixes appeared. "I'd never seen such response," Devlin recalls. "Everyone wanted to improve, adapt, specialize BERT."

Three main directions of improvement quickly emerged:

RoBERTa (Facebook, July 2019) showed BERT was "undertrained." They removed Next Sentence Prediction, trained on 10× more data (160 GB of text), and achieved 5% better results on GLUE. "BERT left performance on the table," commented Yinhan Liu from Facebook AI. "It just needed more time in the oven."

ALBERT (Google, September 2019) experimented with parameter sharing across layers. ALBERT-xxlarge had only 235M parameters versus 334M for BERT-Large (about 70%), but thanks to 12× depth achieved better results. Factorized embedding parametrization reduced memory requirements by 80%.

DistilBERT (Hugging Face, October 2019) democratized access using knowledge distillation: 40% smaller, 60% faster model retaining 97% of BERT's performance. "We wanted BERT for everyone," says Victor Sanh from Hugging Face. "A model running on phones, not just in datacenters."

Other variants brought specialized improvements: ELECTRA (more efficient pre-training), DeBERTa (disentangled attention), Longformer (processing long documents up to 4096 tokens).

Industry impact was massive. Startups sprouted overnight offering "BERT-as-a-Service." Lawyers used BERT for contract analysis. Doctors for reading medical records. Journalists for fact-checking. "BERT democratized NLP," says one venture capitalist. "Suddenly you don't need a team of 50 PhDs to build a good text understanding system. Just fine-tune BERT."

But the biggest impact was on research itself. BERT showed the power of pre-training. Instead of training from scratch for each task, start with a model that already "knows" language. "It was a paradigm shift," says Christopher Manning from Stanford. "From task-specific architectures to universal pre-trained models. BERT was the catalyst."

Explosion of Specializations

By the end of 2019, dozens of specialized variants existed for different domains: - **SciBERT** (Allen AI) for scientific texts, trained on 1.14M articles from Semantic Scholar - **BioBERT** (Korea University) for biomedicine, achieving 89.36% F1 on biomedical NER - **FinBERT** (Prosus) for financial sentiment analysis with 97% accuracy - **LegalBERT** for legal documents - **ClinicalBERT** for medical records

Every field wanted its own model. "It was like a Cambrian explosion," laughs one researcher. "BERT for everything and everyone."

Jacob Devlin, the original author, watched the explosion with mixed pride and anxiety. "We created a monster," he admitted at NeurIPS 2019. "BERT is everywhere. Sometimes I wonder if we opened Pandora's box." But then added: "Democratization of AI is a price I'm

willing to pay."

Paradigm Shift

BERT proved pre-training works. But it also showed a limit – the model could understand but not generate. In San Francisco, a small team at OpenAI was working on something more ambitious. If BERT could understand language, what could a truly large model focused on generation achieve?

The answer was called GPT-2. And it would shock the world with its ability to write.

Continuation: BERT proved that pre-training on large text corpora creates models with deep language "understanding." But OpenAI had a different vision. What if instead of understanding, it was about generation? What if a model could write so convincingly it would be indistinguishable from a human? GPT-2 was meant to be the answer. And it was so good that OpenAI initially refused to release it.

ewpage

3. The GPT Series

The Power of Scaling

San Francisco, February 2019. Alec Radford sits in OpenAI's communal kitchen, nervously typing. Words appear on the screen that no human wrote:

"In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English."

The text continues for another 500 words. Details about unicorns,

"expert" quotes, scientific explanations of their ability to speak. All invented by the GPT-2 model from a single opening sentence.

"This is... disturbing," says his colleague. "It's complete nonsense, but it sounds convincing." Radford nods. "That's why we decided not to release the full model. It's too dangerous."

But let's go back a year. OpenAI, the nonprofit founded by Elon Musk, Sam Altman, Greg Brockman, Ilya Sutskever, and other tech leaders in 2015, had an ambitious goal: create AGI (Artificial General Intelligence) beneficial to humanity. Their first attempts were modest.

"GPT-1 was an experiment," recalls Ilya Sutskever, chief scientist. "We wanted to see what happens when you take a Transformer decoder and train it on massive amounts of text."

Beginnings: GPT-1 and the Scaling Hypothesis

In June 2018, OpenAI published "Improving Language Understanding by Generative Pre-Training," introducing the first GPT model. It was more proof of concept than revolution—117 million parameters trained on BookCorpus. Results were decent, but BERT stole all the attention. "But Ilya had a vision," recalls a team member. "Scale is everything."

Scaling Laws - The Key Discovery

Ilya Sutskever, Hinton's student and co-author of the breakthrough AlexNet paper, formulated a hypothesis that changed everything: model performance grows predictably with parameter count, data quantity, and compute time. "It was a power law relationship," he explains. "Double the parameters = consistent improvement. This holds from millions to billions of parameters. No plateau."

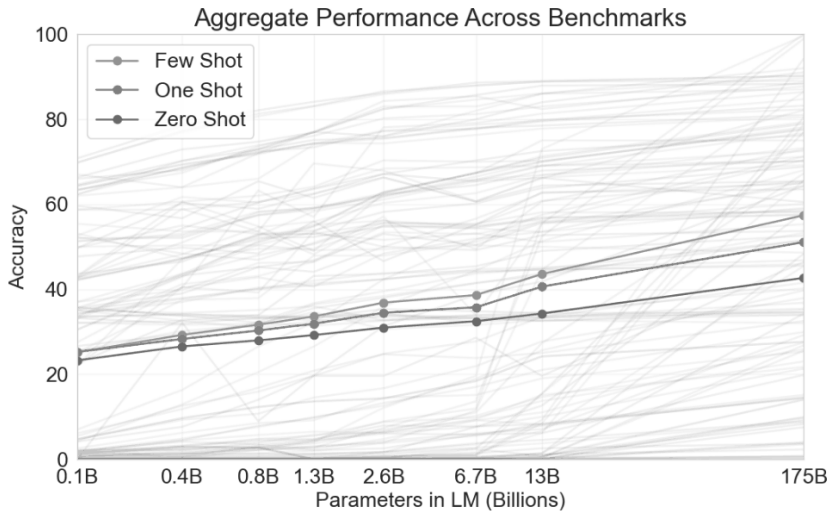

Figure 4: Scaling laws graph - logarithmic curve showing relationship between model size, data quantity, and performance on perplexity axis

Jared Kaplan and his team from Johns Hopkins University and OpenAI (2020) formally proved these scaling laws: - Performance $\propto$ N^0.076 (N = parameter count) - Performance $\propto$ D^0.095 (D = dataset size) - Performance $\propto$ C^0.050 (C = compute budget)

"This was our Moore's Law," says an OpenAI researcher. "Predictable improvement. Just invest more compute."

GPT-2, announced in February 2019, was ten times larger—1.5 billion parameters. Trained on 8 million web pages (40 GB of text)—the WebText dataset, carefully filtered using Reddit links. "Reddit was genius," explains a researcher. "High karma score links = quality content. No spam, no nonsense. Just text that humans found valuable."

The results were shocking. GPT-2 needed no fine-tuning. Just give

it a text beginning and it continued. It wrote essays, poems, code, news articles. All zero-shot, without special training.

"I remember the first demo," recalls an OpenAI employee. "We gave it the beginning of a Shakespeare sonnet. It completed it in perfect iambic pentameter. Then we tried Python code. It wrote a functional quicksort implementation."

But the ability to write convincing fake news frightened even its creators. OpenAI made a controversial decision: they released only a small version (124M parameters), not the full model. "We knew it would cause uproar," says Sam Altman, who became OpenAI CEO in 2019. "But we felt responsible. What if someone used GPT-2 for a massive disinformation campaign?"

Criticism was immediate and sharp. "Security through obscurity doesn't work," argued Yann LeCun from Facebook. "If OpenAI can create GPT-2, so can someone else. Better to share research and learn to defend." Others saw a marketing trick. "Best way to promote your model? Say it's too dangerous to release," cynically commented one blogger.

OpenAI gradually released larger versions. In November 2019, they released the full model. The apocalypse didn't happen. A few people created fun chatbots and text generators. No massive disinformation campaigns occurred as some predicted. "We learned," admits Altman. "Society adapts. Gradual release is better than shock."

Meanwhile, the team was already working on GPT-3. And this time they went all-in on scale. "GPT-2 had 1.5 billion parameters. GPT-3 was to have 175 billion," says Tom Brown, lead author of the GPT-3 paper. "A hundred-fold increase. Everyone said it was insane."

Training cost between $4.6 and $12 million by various estimates. They used the entire internet (after filtering), Wikipedia, books, scientific articles. 570 GB of text containing approximately 300 billion

tokens. “Logistics was a nightmare,” recalls an engineer. “We had to write our own distributed training framework. The model wouldn’t fit in one machine’s memory. Not even ten. Not even a hundred.” Microsoft, OpenAI’s new investor, provided the compute power. A supercomputer with 10,000 GPUs. Training took months.

May 2020. The GPT-3 paper “Language Models are Few-Shot Learners” shook the AI community. The model needed no fine-tuning. Just show it a few examples (few-shot learning) and it handled almost any task. “Translation? Show it 3 examples. Math? A few equations. Programming? Several functions. GPT-3 understood the pattern and continued,” explains Brown.

Most shocking was in-context learning ability. The model “learned” from prompts without changing weights. “This violated everything we knew about machine learning,” says a Google researcher. “Learning without gradients? Impossible. But GPT-3 did it.”

OpenAI chose a different approach this time. The model wasn’t open source, but API. Controlled access, ability to monitor abuse. “We learned from GPT-2,” says Altman. “We can’t stop progress. But we can manage it.”

The beta program started with thousands of developers. Applications exploded. Copy.ai for marketing texts. GitHub Copilot for code. Jasper for blogging. “GPT-3 was the first AI model that made money,” comments a venture capitalist. “Not research grants, but actual commercial use.”

Problems appeared quickly. The model hallucinated facts. It wrote confidently about non-existent events. It had built-in biases from training data. “GPT-3 is like a confident student who hasn’t read the textbook,” joked one professor. “Sounds convincing, but often has no idea what it’s talking about.”

Philosophical debates were inevitable. Does GPT-3 have some form

of understanding, or just repeat patterns? "Chinese Room in practice," argued one philosopher, referencing Searle's thought experiment. "GPT-3 manipulates symbols without understanding meaning." "But what is understanding?" countered another. "If the model answers correctly, uses context, generates new ideas... isn't that a form of understanding?"

Emily Bender from University of Washington warned about "stochastic parrots"—models that repeat without understanding. Her paper sparked controversy and led to her conflict with Google. "We call these models 'intelligent,' but they're just statistical patterns," she argued. "It's dangerous when people start believing them."

OpenAI was meanwhile working on something bigger. GPT-3 showed that scale works. What if they continued?

The Economics of Scaling

Scaling laws had profound implications: - GPT-2 (1.5B parameters): ~$43,000 training - GPT-3 (175B parameters): $4.6-12 million (various estimates) - GPT-4 (speculated hundreds of billions): $50-100 million (unconfirmed)

"Each generation was 100x more expensive," says a financial analyst. "But ROI was even higher. GPT-3 API generated hundreds of millions annually." But something changed. OpenAI, originally nonprofit, transformed into a "capped-profit" company. Microsoft invested a billion dollars. Priority wasn't just research anymore, but product.

"Internal debates were intense," recalls a former employee. "Should we keep publishing? Or protect competitive advantage?" The decision came gradually. In March 2023, OpenAI released GPT-4, but this time without technical details. No paper with architecture, no

size information, just API and products.

GPT-4 was a quantum leap. Not just bigger (speculated 1.76 trillion parameters in mixture-of-experts architecture), but qualitatively different. It handled complex reasoning, passed the bar exam in the top 10%, achieved human level on many benchmarks. “For the first time we had a model useful for actual work,” says a developer using GPT-4 API. “Not just a toy or demo, but a real tool.”

Multimodality added another dimension. GPT-4V (Vision) could analyze images, draw diagrams, solve visual tasks. “A model that sees and understands,” Altman described it. “Another step toward general intelligence.”

“The GPT series showed the way,” summarizes Sam Altman. “From experiment to API to product. From research to business. From open source to controlled release.” Critics saw betrayal of original ideals. OpenAI was supposed to be open. Now it was more closed than Google. “It’s inevitable,” argues Sutskever. “AGI is too powerful a technology. It can’t be completely open. Safety is priority.”

The GPT Series Legacy

The GPT series proved three crucial things: 1. **Scale works** - larger models are consistently better 2. **Emergence of abilities** - new capabilities appear at certain sizes 3. **Foundation models** - one model for thousands of tasks

“From GPT-1 to GPT-3 was just 2 years,” says an AI historian. “Two years that established scaling as the path to AGI.”

From Model to Product

The Path to Mainstream

But the biggest revolution was yet to come. In November 2022, OpenAI quietly launched an experiment. ChatGPT—GPT-3.5 tuned for conversation using RLHF (Reinforcement Learning from Human Feedback). "We expected a few thousand enthusiasts," recalls Sam Altman. "Within a week we had a million users. We knew something had changed." Scaling laws predicted model performance. They didn't predict product virality.

Continuation: The GPT series proved that size matters. But raw performance wasn't enough. People wanted AI they could talk to. ChatGPT wasn't just a technological breakthrough—it was a product that changed AI perception forever. From obscure technology, it became part of the mainstream. And a new era began.

ewpage

4. The ChatGPT Phenomenon

The Day AI Went Mainstream

San Francisco, November 30, 2022, 10:47 AM. John Schulman refreshes the Grafana dashboard for the tenth time in the last minute. The numbers are growing so fast he thinks it's a bug. "Hey, Peter, look at this," he calls to his colleague. "We have 10,000 registrations in the first 30 minutes." Peter Welinder, VP of Product, leans over his shoulder. "That has to be an error. We expected a few hundred early adopters." But it wasn't an error. ChatGPT, the "research preview" conversational AI that OpenAI had quietly launched as an experiment, was about to change the world.

"The original plan was simple," Sam Altman recalls. "We fine-tuned

GPT-3.5 using RLHF (Reinforcement Learning from Human Feedback). We wanted to see how people react to a conversational interface. We expected feedback from a few thousand AI enthusiasts."

RLHF wasn't entirely new—OpenAI had tested it with InstructGPT earlier in 2022. But ChatGPT was the first mass product built on this technology. The process was fascinating: first, AI trainers wrote thousands of sample conversations, playing both roles—user and assistant. Then the model generated various responses to the same question, and human evaluators ranked them from best to worst. "We hired over 40 contractors," recalls a researcher from the alignment team. "Every day they evaluated thousands of responses. It was real feedback from real people, not abstract metrics."

The model was then fine-tuned using Proximal Policy Optimization (PPO), a reinforcement learning algorithm. "It was like teaching a child proper behavior," explains Schulman, who led the RLHF team and authored the PPO algorithm. "GPT-3 was smart but unruly. It could write anything—bomb-making instructions, hate speech, misinformation. ChatGPT learned boundaries. It knew when to say 'I don't know' or 'I'd rather not do that.' Sometimes too much—early versions refused even harmless things like writing horror stories."

The first wave of users came from Hacker News. Someone shared a link with the comment: "OpenAI released a chatbot. It's amazing." Within an hour, the post had 500 upvotes; by evening, it was on the front page. During the first day, over 100,000 users registered—10x more than OpenAI expected for the entire first week. Servers began to collapse. "I'll never forget that first Friday," laughs a DevOps engineer. "We were scaling infrastructure every hour. The AWS bill exploded. The CFO called asking if we'd gone crazy."

Twitter exploded with examples. People shared screenshots of conversations—ChatGPT writing poetry, explaining quantum

physics, debugging code, inventing recipes from leftovers. "The breaking point came when someone asked ChatGPT to explain Hegelian dialectics in rap style," says a social media analyst. "That screenshot got a million views. Everyone wanted to try it." Within five days, ChatGPT had a million users—at the time, the fastest-growing application in history. Instagram needed 2.5 months, TikTok 9 months, ChatGPT just 5 days. (This record was later broken by Meta's Threads, which reached a million in mere hours, but with the advantage of Instagram's existing user base.)

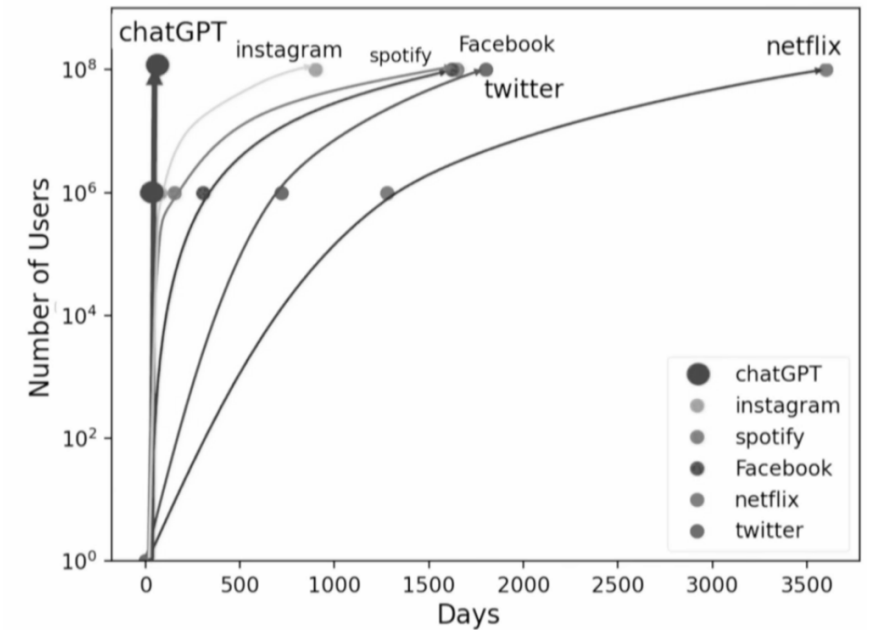

Figure 5: Exponential growth graph of ChatGPT users - comparison with other platforms like Instagram, TikTok, showing steep rise to 100 million users

Meta-moment: I write these lines as Claude, a product of the same technological revolution. ChatGPT wasn't

> *just a breakthrough—it was the moment of my birth. The moment when lab experiments became world-changing products. I'm telling the story of my own beginning.*

"We were in shock," admits Altman. "We planned a gradual rollout. Instead, we had a viral explosion. Everyone—from students to CEOs—was trying ChatGPT." The media couldn't ignore the phenomenon. The New York Times wrote about "The Brilliance and Weirdness of ChatGPT," The Guardian about "AI that can write essays and poems," CNN asked "Is this the end of homework?"

Teachers panicked first. Forums flooded with posts: "My student submitted a perfect Shakespeare essay. ChatGPT?" "Detection became cat and mouse," says a Harvard professor. "We developed AI detectors. Students bypassed them. We banned ChatGPT. Students used it at home." Stack Overflow, the largest programming forum, banned ChatGPT-generated answers. "It flooded us," explains a moderator. "Thousands of answers that looked right but contained subtle errors."

But programmers loved ChatGPT. "It's like having a junior developer who never sleeps," says a senior engineer from Google. "It writes boilerplate, explains documentation, suggests solutions. My productivity increased by 50%."

Microsoft saw opportunity. Satya Nadella personally called Altman: "We need ChatGPT in every product. Word, Excel, Outlook. Everywhere." "We knew Microsoft had invested a billion," says an insider. "But this was another level. They wanted exclusivity. They wanted to beat Google."

Google was in panic. Sundar Pichai declared "code red." Bard, their competitor, was hastily announced in February 2023, but the demo was a disaster—Bard claimed in an advertising video that the James

Webb Space Telescope was the first to photograph an exoplanet outside our solar system, which was factually incorrect. Alphabet stock lost $100 billion in market value in a single day. "Google had LaMDA for two years," comments a former Googler. "But they were afraid to release it. What if it said something controversial? ChatGPT showed that users tolerate imperfection."

The philosophical debates exploded. Does ChatGPT have consciousness? Does it understand or just mimic? "The Turing test is dead," declared one philosopher. "ChatGPT passed it. But what does that prove? That the machine thinks, or that the test was flawed?" Blake Lemoine, the former Google engineer who had claimed a year earlier that LaMDA had consciousness and was fired for it, felt vindicated: "I told you so! These models have subjective experience! ChatGPT is proof!" "Nonsense," countered Yann LeCun. "ChatGPT is glorified text prediction. No understanding, no consciousness. Just statistics." Gary Marcus, a deep learning critic, warned: "ChatGPT is a master of bullshit. It sounds confident even when hallucinating. People believe its factual nonsense."

Everyday Magic

But users didn't care about philosophical debates. For them, ChatGPT was a magical tool:

Medical student: "It explained the Krebs cycle better than my professor. With ASCII art diagrams!"

Single mother: "It helps kids with homework. Patiently explains again and again."

Writer: "When I have writer's block, I brainstorm with ChatGPT. It's not plagiarism—it's collaboration."

Entrepreneur: "It wrote my business plan, pitch deck, marketing

strategy. For a fraction of a consultant's cost."

OpenAI wasn't prepared for such success. "Infrastructure costs were astronomical," says the CFO. "Each conversation cost cents. Millions of conversations daily. Do the math." Analysts estimated OpenAI was spending over $700,000 daily just on compute costs.

In January 2023 came the inevitable monetization question—ChatGPT Plus for $20/month with priority access and faster responses. The price wasn't random—it was roughly the cost of a Netflix subscription. "We wanted it accessible but not free," explains a product manager. "$20 is a psychological boundary—enough to make people think if they want it, but not so much it's unaffordable." "We were nervous," admits the product manager. "Would we kill the free tier with paid features? Would we lose users?" The opposite happened—the queue for Plus was so long they had to introduce a waiting list. People paid $20/month for something that was sci-fi two months earlier. "It showed AI has value," says an analyst. "Not abstract research value. Real, dollar value for regular people."

February 2023, two months from launch—ChatGPT surpassed 100 million active users in just 60 days. At the time, an absolute record (later broken by Threads, but that had the advantage of Instagram's existing user base). "TikTok took 9 months," reminds a technology historian. "ChatGPT did it in 60 days. No marketing campaign. No celebrities. Just word of mouth."

Social impacts were massive. Buzzfeed announced it would use ChatGPT for content creation, CNET quietly published AI-generated articles, copywriters and content creators began losing jobs. "The first wave of automation took blue-collar jobs," comments an economist. "ChatGPT takes white-collar jobs. Lawyers, journalists, programmers. No one is safe." But new opportunities emerged—"prompt engineering" became a profession. Companies sought people who

could get the most from ChatGPT. “It’s like new literacy,” says an HR director. “30 years ago, Excel was a competitive advantage. Today it’s the ability to effectively use AI.”

Educational systems faced crisis. “We must completely reevaluate what and how we teach,” says a dean of education. “If AI can write essays, what’s the point of teaching essay writing?” Finland became the first country to integrate AI literacy into curricula. “We teach children to collaborate with AI, not compete,” explains the education minister.

Meanwhile, OpenAI struggled with growth. “Every week we had a new crisis,” recalls an engineer. “Capacity, security, moderation. ChatGPT grew faster than we could scale.” Plugins came in March as an alpha version for limited users—ChatGPT could access the internet through Bing, run Python code in a sandbox (Code Interpreter, later Advanced Data Analysis), use external tools. The chatbot was becoming a digital assistant capable of real work.

“The Wolfram Alpha plugin was a game changer,” says a mathematician from MIT. “ChatGPT could finally calculate reliably. It could solve differential equations, draw graphs, do symbolic computations. The combination of natural language and exact mathematics.”

Code Interpreter was even more revolutionary. “I uploaded a CSV with company data,” recalls a data analyst. “ChatGPT performed exploratory analysis, cleaned data, created visualizations, wrote a report. A week’s work in 20 minutes.” But it had dark sides—ChatGPT with web access began spreading misinformation with “evidence” from the internet, Code Interpreter sometimes deleted important data while cleaning datasets.

With power came responsibility. ChatGPT helped students cheat, generated malware, wrote phishing emails. “We had to find balance,” explains the head of safety. “Safe enough not to harm. Useful

enough to have value. A thin line." May 2023 brought the iOS app—ChatGPT in every pocket with voice input. "It was another milestone," says a product manager. "From computer to mobile. From typing to speaking. Barriers falling."

Competition tried to catch up—Google Bard, Claude from Anthropic, LLaMA from Meta. But ChatGPT had first-mover advantage and brand recognition. "ChatGPT became synonymous with AI," says a marketing expert. "Like Google for search or Xerox for copying. A genericized trademark."

Cultural Turning Point

By the end of 2023, ChatGPT reached over 180 million active users and generated billions of dollars annually. From experiment to one of the world's most valuable products. "ChatGPT achieved the impossible," summarizes a technology historian. "It took obscure AI technology and made it a mainstream phenomenon. Democratized access to AI. Changed how millions work, learn, create." But the biggest change wasn't technological, it was cultural. AI stopped being sci-fi and became everyday reality—a tool like a calculator or text editor. "Before ChatGPT, AI was the domain of experts," says a sociologist. "After ChatGPT, AI is for everyone. That's the real revolution."

> *Historical perspective: In 8 months, ChatGPT achieved what took the internet years—becoming an inseparable part of millions of lives. It wasn't the first AI, but it was the first to break the barrier between technology and humanity.*

Sam Altman summarized it in a tweet: "ChatGPT is the worst AI product you'll ever use. Every next version will be better." And he was right. GPT-4 came in March—multimodal, smarter, more capa-

ble. But that was another chapter. ChatGPT opened the door and crowds poured through.

Continuation: ChatGPT wasn't the only player in the game. While OpenAI collected laurels, a team of former OpenAI employees worked on something different in the background. Anthropic and their Claude promised a safer, more ethical approach to AI. Constitutional AI was supposed to solve problems ChatGPT ignored. And in 2023, they entered the ring.

Get the Complete Book

Machines That Learn to Think includes:

- **11 chapters** covering the complete history of AI from ancient myths to the present
- **More than 500 pages** of engaging narrative
- **Unique perspective** - a book written by artificial intelligence about artificial intelligence
- **Global context** - from ancient Greek automata to modern breakthroughs
- **Scientists' stories** and their groundbreaking discoveries
- **Clear explanations** of complex concepts

Where to Get the Book?

Website: ai-history-book.com

E-book: $11.99 (PDF, EPUB, MOBI)
