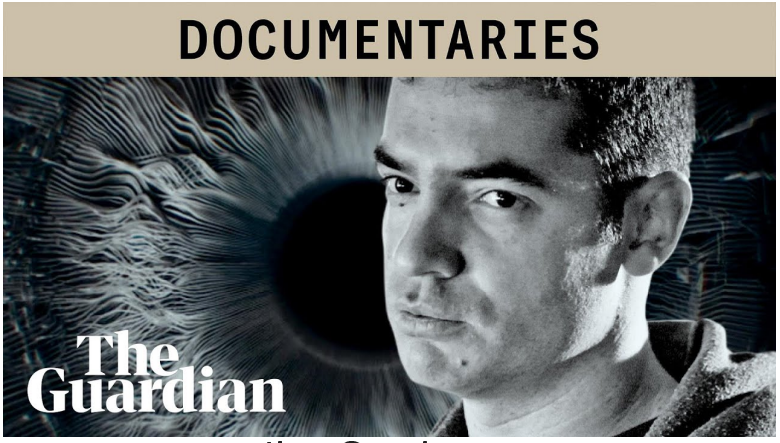
Ilya Sutskever

Pioneer of Deep Learning and AI Safety

People of GenAI

Keynote Series

October 2025



Ilya Sutskever

Pioneer of Deep Learning and AI Safety

Ilya Sutskever

Key Links

- Google Scholar: citations profile
- Twitter/X: @ilyasut
- OpenAI: openai.com
- SSI: ssi.inc

Impact

- Co-creator of foundational deep learning architectures
- Chief Scientist and co-founder of OpenAI
- Leading researcher in AI safety and alignment

	All	Since 2020
Citations	700640	533031
h-index	100	91
i10-index	149	146

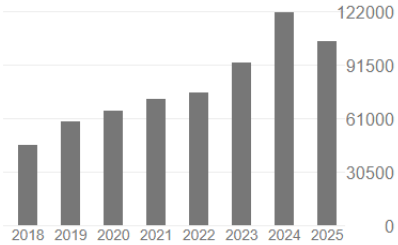


Figure: Google Scholar citations

Education and Career Timeline

Year	Milestone
2002	Moved to Canada, enrolled at University of Toronto
2005	B.Sc. in Mathematics, University of Toronto
2007	M.Sc. in Computer Science, University of Toronto
2013	Ph.D. in CS under Geoffrey Hinton, University of Toronto
2013	DNNResearch acquired by Google; joined Google Brain
2015	Left Google; co-founded OpenAI as Chief Scientist
May 2024	Departed OpenAI to pursue new project
June 2024	Co-founded Safe Superintelligence Inc. (SSI)
July 2025	Became CEO of SSI

Early Days at University of Toronto

- Student in Geoffrey Hinton's Machine Learning group
- Part of the team that revolutionized computer vision
- Worked on breakthrough deep learning techniques
- Collaborated with Alex Krizhevsky on AlexNet

"In the beginning, I was a student in the Machine Learning group of Toronto, working with Geoffrey Hinton."

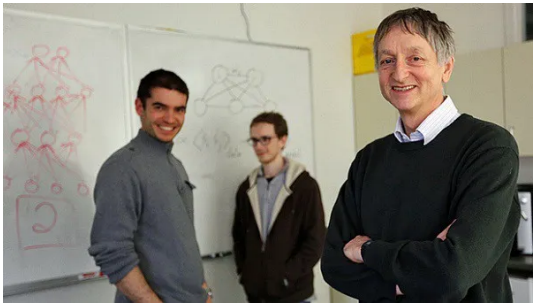


Figure: Ilya, Geoffrey Hinton, and Alex Krizhevsky (2012)

ImageNet 2012: AlexNet Revolution

The Breakthrough

- ImageNet Large Scale Visual Recognition Challenge 2012
- First deep convolutional neural network to win
- Crushed the competition by massive margin
- Sparked the deep learning revolution

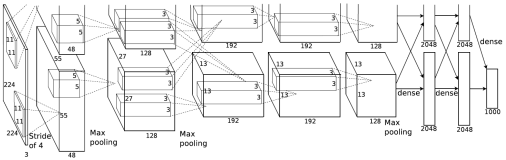


Figure: AlexNet architecture (Krizhevsky, Sutskever, and Hinton, 2012)

Model	Top-5 Error
2nd Place	26.2%
AlexNet	**15.3%**

Figure: Citation impact

Krizhevsky, et al. ImageNet Classification with Deep Convolutional Neural Networks, Curran Associates, Inc., 2012.

Ph.D. Thesis: Training Recurrent Neural Networks

Ph.D. Thesis (2013)

"Training Recurrent Neural Networks" (Sutskever, 2013)

Key Contributions:

- Developed methods to overcome difficulties in training RNNs
- Introduced new variant of Hessian-free (HF) optimizer
- Showed RNNs can learn extreme long-range temporal dependencies
- Applied to character-level language modeling
- Random initialization schemes for gradient descent with momentum

This foundational work on RNNs for NLP paved the way for modern language models.

Sequence to Sequence Learning: Foundation of LLMs

Seq2Seq Model (2014) (Sutskever, Vinyals, and Le, 2014)

- General end-to-end approach to sequence learning
- Uses LSTMs to map input sequences to fixed-dimensional vectors
- Another deep LSTM decodes the target sequence
- **Foundation idea for modern LLMs**
- Achieved state-of-the-art on English-French translation

Figure: Citation impact

Ilya later referred to LSTMs as "a horizontally displaced version of ResNet"

Recognition

NeurIPS Test of Time Award 2024

Watch: Award Talk

Sutskever, vinyals, Sequence to Sequence Learning with Neural Networks, Curran Associates, Inc., 2014.

Working at Google (2013–2015)

DNNResearch Acquisition

- Geoffrey Hinton's startup acquired by Google
- Joined Google Brain as Research Scientist
- Continued pushing boundaries of deep learning

Key Projects at Google

- Early development of **TensorFlow**
- Contributed to deep learning infrastructure
- Research on deep reinforcement learning
- Foundational work that led to **AlphaGo**

2015: Co-founded OpenAI

- Left Google to pursue more ambitious AI research
- Became Chief Scientist at OpenAI
- Vision: Build safe artificial general intelligence

OpenAI: The GPT Revolution

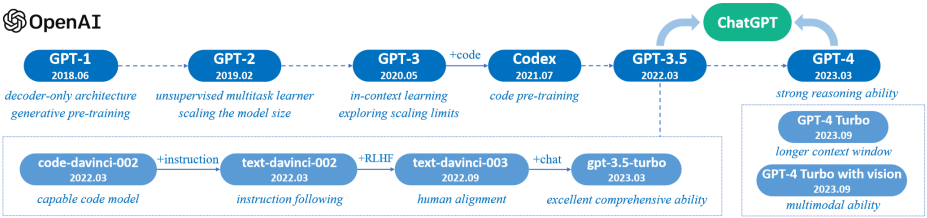


Figure: Evolution of GPT models

- **GPT-1** (Radford and Narasimhan, 2018): Generative pre-training
- **GPT-2** (Radford, Wu, et al., 2019): Unsupervised multitask learning (1.5B params)
- **GPT-3** (Brown et al., 2020): Few-shot learning (175B params)
- **InstructGPT** (Ouyang et al., 2022): RLHF alignment
- **GPT-4** (OpenAI, 2023): Multimodal capabilities
- **GPT-4o** (OpenAI, 2024): Omni-model (text, audio, vision)

Radford, et al. (2018). *Improving Language Understanding by Generative Pre-Training*.
Radford, Wu, et al. (2019). *Language Models Are Unsupervised Multitask Learners*.
Brown, et al. Language Models Are Few-Shot Learners, 2020.
Ouyang, et al. (Mar. 4, 2022). *Training Language Models to Follow Instructions with Human Feedback*.
OpenAI (2023). *Gpt-4 Technical Report*.
— (Oct. 25, 2024). *GPT-4o System Card*.

Other Projects at OpenAI

DALL-E Series: Text-to-Image Generation

- **DALL-E** (Ramesh, Pavlov, et al., 2021): Zero-shot text-to-image
- **DALL-E 2** (Ramesh, Dhariwal, et al., 2022): Using CLIP latents
- **DALL-E 3** (Betker et al., 2023): Better caption following
- **GLIDE** (Nichol et al., 2022): Photorealistic generation

Other Vision Work

- **CLIP** (Radford et al., 2021): Vision-language models
- **GPT-4V** (OpenAI, 2023): Vision capabilities

Code Generation

- **Codex** (Chen, Tworek, et al., 2021): GPT finetuned on code
- Powers GitHub Copilot
- HumanEval benchmark

Impact

As Chief Scientist, Ilya provided technical leadership and vision for all major OpenAI projects, pushing the boundaries of what's possible with large-scale AI systems.

Ramesh, Pavlov, et al. Zero-Shot Text-to-Image Generation. PMLR, 2021.
Ramesh, Dhariwal, et al. (Apr. 12, 2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*.
Betker, et al. (2023). *Improving Image Generation with Better Captions*. Pre-published.
Nichol, et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. PMLR, 2021.
Radford, et al. Learning Transferable Visual Models From Natural Language Supervision. PMLR, 2021.
OpenAI. GPT-4V(ision) System Card. 2023.
Chen, Radford, et al. Generative Pretraining From Pixels. PMLR, 2020.
Chen, Tworek, et al. (July 14, 2021). *Evaluating Large Language Models Trained on Code*.

Superalignment: Aligning Superhuman AI

The Challenge

- Future AI systems will be too complex for humans to reliably evaluate
- Need scientific breakthroughs to steer and control superhuman AI
- Humans will only be able to *weakly supervise* superhuman models

OpenAI Superalignment Team (OpenAI, 2023)

- Co-led by Ilya Sutskever and Jan Leike
- Dedicated 20% of OpenAI's compute to alignment research
- Goal: Solve alignment within four years

Key Research: Weak-to-Strong Generalization (Burns et al., 2024)

- Can weak supervisors elicit full capabilities of stronger models?
- Demonstrated promising generalization from weak to strong models
- Empirical progress on aligning superhuman AI

OpenAI (July 5, 2023). *Introducing Superalignment*.
Burns, et al. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision, PMLR, 2024.

Governance and Safety Philosophy

Vision for AI Safety

- AI systems must be built, deployed, and used safely (OpenAI, 2023a)
- Alignment research: learning from human feedback (OpenAI, 2022)
- Goal: Build aligned AI to help solve all other alignment problems

Governance of Superintelligence (OpenAI, 2023b)

- Need to think ahead about superintelligence governance
- AI systems dramatically more capable than AGI
- Coordination between organizations and nations

Recognition

Leopold Aschenbrenner's "Situational Awareness" (2024) (Aschenbrenner, 2024) is dedicated to Ilya Sutskever

Ilya's commitment to AI safety has been a driving force in making alignment research a priority in the AI community.

OpenAI (Apr. 5, 2023a). *Our Approach to AI Safety*.
— (Aug. 24, 2022). *Our Approach to Alignment Research*.
— (May 22, 2023b). *Governance of Superintelligence*.
Aschenbrenner, Situational Awareness, 2024.

Safe Superintelligence Inc. (SSI)

May 2024: Departed OpenAI

- Left OpenAI to pursue new project focused on AI safety

June 2024: Co-founded SSI

- Co-founded with Daniel Gross and Daniel Levy
- Mission: Build safe superintelligence
- Focus on safety as the primary objective
- Not distracted by management overhead or product cycles

July 2025: Became CEO of SSI

- Took on leadership role after Daniel Gross departed to Meta
- Continuing the mission of safe superintelligence

Mission

"We will pursue safe superintelligence in a straight shot, with one focus, one goal, and one product: a safe superintelligence." — SSI Mission Statement

Legacy and Impact

Technical Contributions

- Co-creator of AlexNet — sparked deep learning revolution
- Invented Seq2Seq — foundation for all modern LLMs
- Technical leadership for GPT, DALL-E, CLIP series
- Advanced RNN training methods and optimization

AI Safety Leadership

- Pioneer in AI alignment research
- Established superalignment as critical research area
- Founded SSI to focus exclusively on safe superintelligence

Influence

- One of the most cited researchers in deep learning
- Mentored next generation of AI researchers
- Shaped the direction of AI research toward safety and alignment

References

Aschenbrenner, Situational Awareness. 2024.
Burns, et al. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision, PMLR, 2024.
OpenAI (May 22, 2023a). *Governance of Superintelligence*.
— (July 5, 2023b). *Introducing Superalignment*.
— (Apr. 5, 2023c). *Our Approach to AI Safety*.
— (Aug. 24, 2022). *Our Approach to Alignment Research*.
Betker, et al. (2023). *Improving Image Generation with Better Captions*. Pre-published.
Ramesh, Dhariwal, et al. (Apr. 12, 2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*.
Chen, Radford, et al. Generative Pretraining From Pixels, PMLR, 2020.
Nichol, et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, PMLR, 2022.
OpenAI (Oct. 25, 2024). *GPT-4o System Card*.
— (2023d). *Gpt-4 Technical Report*.
OpenAI. GPT-4V(ision) System Card, 2023.
Radford, Kim, et al. Learning Transferable Visual Models From Natural Language Supervision, PMLR, 2021.
Ramesh, Pavlov, et al. Zero-Shot Text-to-Image Generation, PMLR, 2021.
Brown, et al. Language Models Are Few-Shot Learners, 2020.
Chen, Tworek, et al. (July 14, 2021). *Evaluating Large Language Models Trained on Code*.
Ouyang, et al. (Mar. 4, 2022). *Training Language Models to Follow Instructions with Human Feedback*.
Radford, et al. (2018). *Improving Language Understanding by Generative Pre-Training*.
Radford, Wu, et al. (2019). *Language Models Are Unsupervised Multitask Learners*.
Krizhevsky, et al. ImageNet Classification with Deep Convolutional Neural Networks, Curran Associates, Inc., 2012.
Sutskever, et al. Sequence to Sequence Learning with Neural Networks, Curran Associates, Inc., 2014.
Sutskever, Training Recurrent Neural Networks. 2013.

Thank You

Questions?

People of GenAI Keynote Series