# Zeyuan ALLEN-ZHU

Homepage: http://zeyuan.allen-zhu.com
Scholar: https://scholar.google.com/citations?user=MoJFIiQAAAAJ

## EDUCATION & WORK EXPERIENCE

| | | |
|---|---|---|
| AI Research Scientist | **FAIR AT META** | 2022–now |
| Principal Researcher | **MICROSOFT RESEARCH REDMOND** | 2017–22 |
| Postdoc | **PRINCETON UNIVERSITY + IAS**  (hosts: *Elad Hazan*, *Avi Wigderson*) | 2015–17 |
| Sc.D. in Computer Science | **MIT**  (advisors: *Jonathan Kelner*, *Silvio Micali*) | 2012–15 |
| S.M. in Computer Science | **MIT**  (thesis award) | 2010–12 |
| B.S. in Math and Physics | **TSINGHUA UNIVERSITY**  (summa cum laude) | 2006–10 |

## COMPETITION AWARDS

| | | |
|---|---|---|
| 2009 | **World's 2nd Place / Gold Medal** | *ACM/ICPC World Final* |
| 2009 | **Outstanding Prize (World's Top Prize)** | *The Mathematical Contest in Modeling (MCM)* |
| 2008 | **World's 2nd Place** | *Google Code Jam* |
| 2006 | **World Champion** | *USA Computing Olympiad (USACO)* |
| 2006 | **USA Champion** | *USA Invitational Computing Olympiad (USAICO)* |
| 2006 | **Gold Medal** | *International Olympiad in Informatics (IOI)* |
| 2005 | **Gold Medal** | *International Olympiad in Informatics (IOI)* |
| 2005 | **China Champion** | *National Olympiad in Informatics* |

## RESEARCH INTERESTS

My research focuses on *Physics of Language Models*—a scientific framework to uncover the universal laws governing how large AI models learn *knowledge*, *reason*, and *generalize*. I design controlled experiments and probe neurons to reveal the mechanisms behind their strengths and failures, aiming to provide both theoretical insight and practical guidance—on *data preparation, pre-/post-training*, and *architecture*—for building better and safer AGI beyond today's AI systems. This line of work was featured in my ICML 2024 tutorial.

Before that, I worked on the mathematics of deep learning, developing proofs on the learnability of neural networks to explain phenomena observed in practice. Our work on ensemble and knowledge distillation received the ICLR 2023 Best Paper Runner-Up, and our COLT 2023 paper provided the first formal proof of why and *how* deep networks perform deep learning (e.g., achieve superior performance over layer-wise training). This theoretical line inspired our LoRA fine-tuning method, now widely adopted across the AI community, and continues to shape the Physics of Language Models.

Earlier in my career, I also worked on optimization theory and theoretical computer science.

## RESEARCH AWARDS

| | | |
|---|---|---|
| 2025 | **Meta Special RSU Award** | *Meta* |
| 2023 | **Outstanding Paper Honorable Mention (top 9 in 4966)** | *ICLR 2023* |
| 2021 | **Microsoft Special Stock Award** | *Microsoft Research* |
| 2018 | **Microsoft Special Stock + Cash Award** | *Microsoft Research* |
| 2018 | **Best Paper Award** | *ICML 2018 eRL workshop* |
| 2016 | **Microsoft Azure Research Grant** | *Microsoft Research* |
| 2013 | **Simons Award for TCS Graduate Students** | *Simons Foundation* |
| 2009 | **Best Student Paper Runner-Up** | *ICDM 2009* |
| 2009 | **Microsoft Young Fellow** | *Microsoft Research Asia* |

# Zeyuan ALLEN-ZHU

## SELECTED OPEN SOURCE

https://github.com/facebookresearch/PhysicsLM4
https://huggingface.co/facebook/PhysicsLM4.2__LlamaCanon-8B-Nemo-1T-lr0.003

## SELECTED TALKS

| | |
|---|---|
| "Physics of Language Models" | ICML 2024 tutorial |
| "Why Does Deep Learning Perform Deep Learning" | Microsoft Research 2020 |
| "Recent Advances in Stochastic Convex and Non-Convex Optimization" | ICML 2017 tutorial |

## SELECTED PUBLICATIONS

### *[Physics of Language Models @ FAIR at Meta]*

o **"Physics of Language Models: Part 4.1, Architecture Design and the Magic of Canon Layers"**
Z. Allen-Zhu★. *(NeurIPS 2025)*

o **"Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws"**
Z. Allen-Zhu★, Yuanzhi Li. *(ICLR 2025)*

o **"Physics of Language Models: Part 3.2, Knowledge Manipulation"**
Z. Allen-Zhu★, Yuanzhi Li. *(ICLR 2025)*

o **"Physics of Language Models: Part 3.1, Knowledge Storage and Extraction"**
Z. Allen-Zhu★, Yuanzhi Li. *(ICML 2024)*

o **"Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process"**
Tian Ye, Zicheng Xu, Yuanzhi Li, Z. Allen-Zhu★. *(ICLR 2025)*

o **"Physics of Language Models: Part 2.2, How to Learn From Mistakes on Grade-School Math Problems"**
Tian Ye, Zicheng Xu, Yuanzhi Li, Z. Allen-Zhu★. *(ICLR 2025)*

o **"Physics of Language Models: Part 1, Learning Hierarchical Language Structures"**
Z. Allen-Zhu★, Yuanzhi Li. *(TMLR 2025)*

### *[Theory of Deep Learning @ Microsoft Research]*

o **"Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning"**
Z. Allen-Zhu★, Yuanzhi Li. *(ICLR 2023; best paper honorable mention)*

o **"Backward Feature Correction: How Deep Learning Performs Deep Learning"**
Z. Allen-Zhu★, Yuanzhi Li. *(COLT 2023)*

o **"Feature Purification: How Adversarial Training Performs Robust Deep Learning"**
Z. Allen-Zhu★, Yuanzhi Li. *(FOCS 2021)*

o **"LoRA: Low-Rank Adaptation of Large Language Models"**
E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu★, Y. Li, S. Wang, W. Chen. *(ICLR 2022; cited by 24k+)*

o **"A Convergence Theory for Deep Learning via Over-Parameterization"**
Z. Allen-Zhu★, Yuanzhi Li, Zhao Song. *(ICML 2019; cited by 1970)*

o **"Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers"**
Z. Allen-Zhu★, Yuanzhi Li, Yingyu Liang. *(NeurIPS 2019; cited by 1027)*

# Zeyuan ALLEN-ZHU

### *[Optimization Theory @ Princeton + IAS + MSR]*

o **"Is Q-Learning Provably Efficient?"**
Chi Jin, Z. Allen-Zhu★, S. Bubeck, M. Jordan.          *(NeurIPS 2018; cited by 1145; eRL workshop best paper)*

o **"Katyusha: The First Direct Acceleration of Stochastic Gradient Methods"**
Z. Allen-Zhu★.                                    *(STOC 2017 + JMLR 2018; cited by 781)*

o **"Katyusha X: Practical Momentum Method for Stochastic Sum-of-Nonconvex Optimization"**
Z. Allen-Zhu★.                                    *(ICML 2018)*

o **"Natasha: Faster Stochastic Non-Convex Optimization via Strongly Non-Convex Parameter"**
Z. Allen-Zhu★.                                    *(ICML 2017)*

o **"Natasha 2: Faster Non-Convex Optimization Than SGD"**
Z. Allen-Zhu★.                                    *(NeurIPS 2018)*

o **"Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent"**
Z. Allen-Zhu, Lorenzo Orecchia.                   *(ITCS 2017)*

o **"Operator Scaling via Geodesically Convex Optimization, Invariant Theory and Polynomial Identity Testing"**
Z. Allen-Zhu★, Ankit Garg, Yuanzhi Li, Rafael Oliveira, Avi Wigderson.          *(STOC 2018)*

o **"First Efficient Convergence for Streaming k-PCA: a Global, Gap-Free, and Near-Optimal Rate"**
Z. Allen-Zhu★, Yuanzhi Li.                        *(FOCS 2017)*

o **"Much Faster Algorithms for Matrix Scaling"**
Z. Allen-Zhu★, Yuanzhi Li, Rafael Oliveira, Avi Wigderson.          *(FOCS 2017)*

o **"Finding Approximate Local Minima for Nonconvex Optimization in Linear Time"**
Naman Agarwal, Z. Allen-Zhu★, Brian Bullins, Elad Hazan, Tengyu Ma.          *(STOC 2017)*