David Baek

dbaek.org | dbaek-ai | david-baek | dbaek__ | Google Scholar

Summary

MIT Ph.D. student focusing on building reliable and safe AI systems, with extensive experience in **LLM post-training** and **Reinforcement Learning** for both academic research and industrial applications. My research has been published in top-tier conferences and workshops, including NeurIPS and ICLR. I am passionate about applying quantitative skills to solve challenging problems.

Education

Massachusetts Institute of Technology (MIT) — Cambridge, MA, USA
Ph.D. in Electrical Engineering & Computer Science (EECS), GPA: 5.0/5.0 — Sep 2023 – May 2027

- Advisor: Max Tegmark
- Research Area: Responsible AI, Large Language Models, AI Alignment

Seoul National University (SNU) — Seoul, Korea
B.S. in Physics and Computer Science, Summa Cum Laude, GPA: 4.23/4.3 — Mar 2017 – Aug 2023

- Presidential Award (Ranked **1st** among graduating cohort in College of Natural Sciences)
- Includes two years on leave for compulsory military service (2020–21, Job: Cyber Security Specialist)

Selected Publications

1. D. Baek*, J. Engels*, ..., "Scaling Laws for Scalable Oversight," **NeurIPS 2025 (Spotlight, Top 3%)**.
2. D. Baek, M. Tegmark, "Towards Understanding Distilled Reasoning Models: A Representational Approach," **ICLR 2025 BuildingTrust Workshop**.
3. D. Baek, Z. Liu, M. Tegmark, "GenEFT: Understanding Statics and Dynamics of Model Generalization via Effective Theory," **Physical Review E 111, 035307 (2025)**.
4. D. Baek*, Y. Li*, E. Michaud*, J. Engels, X. Sun, M. Tegmark, "The Geometry of Concepts: Sparse Autoencoder Feature Structure," **Entropy 27(4), 344 (2025)**.
5. J. Zhang, A. Estornell, D. Baek, B. Li, X. Xu, "Any-Depth Alignment: Unlocking Innate Safety Alignment of LLMs to Any-Depth," under review (ICLR 2026).

Experience

Tegmark AI Safety Group — Dec 2023 - Present
Graduate Research Assistant (Advisor: Prof. Max Tegmark) — Cambridge, MA, USA

- Studied various weak-to-strong oversight protocols and theory of hierarchical oversight
- Studied geometrical structure of representations in LLMs, with extensive experience in fine-tuning/training LLMs
- Proposed and empirically verified physics-inspired effective theory of neural network generalization

ByteDance — Jun 2025 - Aug 2025
Machine Learning Research/Engineer Intern (Mentor: Jie Mei, Andrew Estornell) — Bellevue, WA, USA

- Research on Multi-agent Reinforcement Learning and improving the safety and robustness of LLMs
- Proposed a novel product moderation paradigm in TikTok Shop and trained a 7B model, saving $660k/year in total

Honors & Awards (Selected)

- Gold Level Certificate, WorldQuant Brain Platform, 2024
- Silver Medal, University Physics Competition, 2018
- Finalist, Samsung Collegiate Programming Cup (SCPC), 2018
- Silver Medal, Korean Mathematical Olympiad (High School Division), 2016
- Silver Medal, International Junior Science Olympiad (IJSO), 2014

Technical Skills

Mathematics: Probability, Statistics, Stochastic Processes, Time Series Analysis, Linear Algebra, Optimization
Programming: Python, C/C++, Java, Matlab, Mathematica, LaTeX, HTML, Javascript
Libraries: PyTorch, Tensorflow†, HuggingFace, Wandb, Numpy, Scipy, QuTiP, Vue.js/Vuetify, etc.
Machine Learning: Large Language Models, Diffusion Models, Computer Vision, Interpretability Techniques

Community Service/Other Information

- Chair of Publicity & Communications Committee @ Ashdown House (MIT Graduate Housing) — Nov 2023 - Present
- Vice President of Publicity @ MIT EECS Graduate Student Association — Jan 2024 - Jan 2025
- Undergraduate Student Research Mentoring: Riya Tyagi (Spring 2025), Duru Ozer (Spring 2025)