

# David (Dowon) Baek

✉ dbaek@mit.edu | 🌐 david-baek | in dbaek-ai | 📺 david-baek | 🐦 dbaek\_

## EDUCATION

### Massachusetts Institute of Technology (MIT)

Cambridge, MA, USA

Ph.D. in Electrical Engineering & Computer Science (EECS), GPA: 5.0/5.0

Sep 2023 – Present

- Advisor: Max Tegmark
- Research Area: LLM Interpretability, Representation Learning, AI Safety

### 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)

## PUBLICATIONS

1. [D. Baek\\*](#), J. Engels\*, S. Kantamneni\*, M. Tegmark, “Scaling Laws of Scalable Oversight,” 2025, in preparation.
2. [D. Baek](#), M. Tegmark, “Towards Understanding Distilled Reasoning Models: A Representational Approach,” 2025, [ICLR 2025 Workshop on Building Trust in LLMs](#).
3. [D. Baek\\*](#), Z. Liu\*, R. Tyagi, M. Tegmark, “Harmonic Loss Trains Interpretable AI Models,” 2025, [arXiv](#), under review.
4. [D. Baek](#), Y. Li, M. Tegmark, “Generalization from Starvation: Hints of Universality in LLM Knowledge Graph Learning,” 2024, [arXiv](#), under review.
5. [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).
6. [D. Baek](#), Z. Liu, M. Tegmark, “GenEFT: Understanding Statics and Dynamics of Model Generalization via Effective Theory,” *Phys. Rev. E* 111, 035307 (2025).
7. S. H. Park, [D. Baek](#), I. Park, S. Hahn, “Design of Scalable Superconducting Quantum Circuits using Flip-chip Assembly,” *IEEE Transactions on Applied Superconductivity*, 33(5), pp.1-6 (2023).

## 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
- Proposed harmonic loss for training interpretable AI models (Received **1.1 million** views on Twitter)
- Studied geometrical structure of knowledge representations in Large Language Models (LLMs), with experience in fine-tuning LLMs and Sparse Autoencoders (SAEs) using PyTorch and Transformers package
- Proposed and empirically verified physics-inspired effective theory of neural network generalization

## HONORS & AWARDS (SELECTED)

- 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
- Gold Level Certificate, WorldQuant Brain Platform, 2024

## TECHNICAL SKILLS

**Mathematics:** Probability, Statistics, Stochastic Processes, Time Series Analysis, Linear Algebra, Optimization

**Programming:** Python, C/C++, Java, Matlab, Mathematica,  $\text{\LaTeX}$ , HTML, Javascript

**Libraries:** PyTorch, Tensorflow<sup>†</sup>, HuggingFace, Wandb, Numpy, Scipy, QuTiP, Vue.js/Vuetify, etc.

**Machine Learning:** Large Language Models, Diffusion Models, Computer Vision, Interpretability Techniques

## COMMUNITY SERVICE

- 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)