Chang-Han Chen

Contact Information

- ☎ Tel: 608-886-6081
- ✉ E-mail: changhanc@berkeley.edu
- 🐾 Website: https://chang-han-chen.github.io

Education

- PhD in theoretical physics at the **University of California, Berkeley**, Class of 2028.
- BSc in Physics and Mathematics (GPA: 4.9/5) at **Massachusetts Institute of Technology**, Class of 2022. Best senior thesis award.

Research Interest

I aim to tackle the most puzzling, mysterious aspects of our universe.

- **Central dogma of quantum gravity: GR+QM=hologram?**
  Instead of betting on one particular theories such as string theory or loop quantum gravity, I believe in a more general lore: *any* non-perturbative quantum theory derived from classical gravity will be holographic. Proving the "central dogma" is the goal of my PhD.

Recently, I am convinced in the urgency of AI safety research and have started some **LLM interpretability/alignment** projects.

Selected AI projects

- **Can off-policy probes detect genuine deception?** — 2025 Mid-May (Expected)
  - Investigated training lie detectors on off-policy data sources to address the scarcity of realistic lying samples generated by current models, exploring whether detectors can be effectively trained on transcripts from other LLMs or humans.
  - As a by-product, we developed an LLM-assisted automated pipeline generating approximately 200 model organisms (large enough for hypothesis testing but not training a probe), creating a valuable dataset that enables statistically rigorous alignment research.

- **LLM debate with look-ahead tree search**, independent — 2025 February
  - Investigated AI debate for scalable oversight by enhancing their debate strategies. The best-of-N strategy is shown to saturate at $N = 4$ on most datasets.
  - Expanded this LLM debate rerpo to include look-ahead tree search as a debate strategy.
  - Unfortunately, the result was null; despite being more mathematically correct, finding minimax optimal solutions in the debate tree does not seem to increase the faithfulness on MATH dataset. I might try it on more datasets with the future, more capable models.

Selected physics projects

- **A clock is just a way to tell the time** with G. Penington — 2024 Spring
  - Resolved prevalent misconceptions in the field as a junior graduate student and pioneered practical applications of abstract mathematical tools to cosmological modeling.
  - Our novel approach and insights have become fairly influential to the field!

- **Firewalls at exponentially late times** with A. Blommaert and Y. Nomura — 2024 Spring
  - Tackled an ambitious theoretical challenge by developing toy models and rigorous partial progress, establishing a mathematical framework for future research

- **Geometry and entanglement in AdS/CFT and beyond** — 2022 Spring
  - Conducted independently and received the best senior thesis award in Theoretical Physics from MIT.

Honors and Awards

- *Invitation talk* at the seminar "Information, Geometry and Physics Seminar," Caltech — 2024
- *Invitation talk* at the workshop "New ideas for old black holes," UC Berkeley — 2024
- *Morse/Orloff Research Award* for my best senior thesis in theoretical physics, MIT — 2022
- *Gold medal* in International Physics Olympiad (IPhO). — 2018
- *Silver medal* (first place) in Asian Physics Olympiad (APhO). — 2018

References

Prof. **Hong Liu**	Prof. **Geoff Penington**	Prof. **Yasunori Nomura**
MIT	Berkeley	Berkeley
E-mail: hong_liu@mit.edu	E-mail: geoffp@berkeley.edu	E-mail: ynomura@berkeley.edu