Ameya Prabhu

ameya.prabhu@bethgelab.org | ameya.prabhu.be
Ameya Prabhu | ameya-prabhu | @AmyPrb
Address: Tübingen AI Center, Maria-Von-Linden-Straße 1, Tübingen, Germany - 72076

Research Interest

My research agenda targets scalable post-training for automated *discovery* in general, aiming to transition LLMs from merely recalling pre-trained human knowledge to actively generating and verifying new hypotheses. I work across three main pillars:

- **Metacognitive Continual Learning:** Developing agents with stateful inference that can reflect on their own reasoning. My focus is on: (i) building systems that monitor their internal reasoning process to identify stagnation, backtrack, and self-correct; and (ii) designing memory mechanisms that learn from past mistakes without catastrophic forgetting.
- **Long-Horizon Post-Training:** I want to design better long-horizon RL trajectories as a principled, scalable approach to online data curation even in superhuman settings. My focus is on: (i) dense reward structures, (ii) intrinsic motivation by forecasting consequences of exploration – enabling on-the-fly planning guided by metacognition to enable effective exploration of massive search spaces.
- **LM-Agent driven Forecasting:** Can AI translate vague research intuitions into concrete, verifiable predictions about things we haven’t discovered yet? I aim to post-train LLMs to weigh conflicting evidence in literature among high uncertainty to develop principled ways of forecasting the most viable paths for future scientific breakthroughs (or generally larger goals).

Education

- **PhD, University of Oxford** *Oct ’20 - Feb ’25*
  Lab: Torr Vision Group Oxford, UK
  - Thesis: Scalable Continual Deep Learning With Computational Cost Considerations
    Advisor: Philip Torr Committee: Yee Whye Teh (Oxford) Cordelia Schmid (INRIA & Google)
  - Summary: I created a continual learning algorithms from both feedback and new data collected post-deployment in real-world settings. We design slow- and fast- learning system using continual training with model merging and kNN respectively, and design fast, automatic data annotation from internet searches to create a never-ending image learning system for foundational models. Parts of this project got adopted by Meta AI. Now, I am working on meta-cognition as the primary continual learning mechanism for LLM post-training.
- **MSc, International Institute of Information Technology (IIIT-H)** *Aug’18 - Aug’19*
  Lab: Center for Visual Information Technology (CVIT) Hyderabad, India
  - Thesis: Exploring Binarization and Pruning of Convolutional Neural Networks
    Advisor: Anoop Namboodiri Committee: Avinash Sharma (IIT-Jodhpur)
  - Summary: I explored simple but effective methods to reduce space and compute costs in deep models at deployment using pruning and binarization (weights and activations quantized to ±1). Asking “where to binarize" gives large performance boosts to binary models with little tradeoff in compute costs. I show that random pruning beating state-of-the-art pruning methods, and connect random pruning with expander graphs explaining why it is a good mechanism (high sparsity & connectivity). Together, this ternary system {-1, 0, 1} allows for stable performance but extreme quantization, with large speedups on custom hardware. Parts of this project got adopted by Texas Instruments.
- **BTech, International Institute of Information Technology (IIIT-H)** *Aug’14 - Aug’18*
  Field: Bachelors in Computer Science and Engineering (CGPA: 8.94/10) Hyderabad, India
  - Deans Merit List Award for Outstanding Academic Performance

Work Experience

- **Bethgelab, Tübingen AI Center [🌐]** *Mar’24 - Current*
  Postdoctoral Researcher Advisor: Matthias Bethge Tübingen, Germany
  - AI Oversight (Paper), Science of Benchmarking (Paper) and Continual Learning (Paper).
- **Intel Labs [🌐]** *Sep ’21 - June ’22*
  ML Research Intern Manager: Ozan Sener and Vladlen Koltun Munich, Germany
  - Worked on Continual Learning Algorithms at Scale for Adapting to Rapidly Changing Datastreams (Publication here).
- **SERI MATS 1.0 [🌐]** *Sep ’21 - June ’22*
  Mentee Mentor: Evan Hubinger and Ethan Perez Online, Remote
  - Worked on understanding inductive biases in DNNs (Alignment Forum) and inverse scaling of memorization (Publication here).
- **Verisk Analytics [🌐]** *Aug ’18 - Oct ’19*
  AI Resident Manager: Maneesh Singh Jersey City, USA
  - Worked on Scaling Active Learning Methods using Model-Based Active Curation (Publication here).
- **IBM Research [🌐]** *May ’18 - Sep ’18*
  ML Research Intern Manager: Anush Sankaran and Riddhiman Dasgupta Bangalore, India
  - Worked on LLMs (RNNs) for Neural Architecture Prediction. Best Intern Award. (Publication here).

Selected Media Coverage

[2024] **Computerphile (YouTube)**. Our paper “No zero-shot without exponential data" was featured on YouTube by Computerphile (2.5M subscribers), and is now the most watched video (1M+ views).

Patents

[2021] Ameya Prabhu, Charles Dognin, Maneesh Singh (2021). **Machine Learning Systems and Methods for Evaluating Sampling Bias in Deep Active Classification**. US Patent, Patent No. 20210004700. Publication Date: 07/2021.

Publications

*equal lead †equal advising

[In Submission (ICML ’26)] Nikhil Chandak*, Shashwat Goel*, Ameya Prabhu†, Moritz Hardt†, Jonas Geiping†. **Scaling Open-Ended Reasoning to Predict the Future**. *SEA Workshop, NeurIPS’25 (Best Poster Award)*.

[In Submission (Journal - JIE)] Tushita Jha, Rory Svarc, Ameya Prabhu, Matthias Bethge. **Jevons Boon: The Economics of Abundant Cognition from Cheap LLM Labour**. **Upcoming 2026**.

[ICLR ’26] Jackson Harmon, Andreas Hochlehnert, Matthias Bethge†, Ameya Prabhu†. **Mapping Post-Training Forgetting in Language Models at Scale**. *Arxiv Preprint, 2025*.

[ICLR ’26 Mikhail Terekhov*, Alexander Panfilov*, Daniil Dzenhaliou*, Caglar Gulcehre, Maksym Andriushchenko†, Ameya Prabhu†, Jonas Geiping†. **Strategic Dishonesty Can Undermine AI Safety Evaluations of Frontier LLMs**. *Arxiv Preprint, 2025*.

[ICLR ’26] Alexander Panfilov*, Evgenii Kortukov*, Kristina Nikolić, Matthias Bethge, Sebastian Lapuschkin, Wojciech Samek, Ameya Prabhu, Maksym Andriushchenko, Jonas Geiping. **Strategic Dishonesty Can Undermine AI Safety Evaluations of Frontier LLMs**. *Arxiv Preprint, 2025*.

[In Submission (TMLR ’26)] Nikhil Chandak*, Shashwat Goel*, Ameya Prabhu, Moritz Hardt†, Jonas Geiping†. **Answer matching outperforms multiple choice for LLM evaluations**. *Arxiv Preprint, 2025*.

[ICCV ’25] Daniil Zverev*, Thaddäus Wiedemer*, Ameya Prabhu, Matthias Bethge, Wieland Brendel, A. Sophia Koepke. **VGGSounder: Audio-Visual Evaluations for Foundation Models**. *ICCV, 2025*.

[COLM ’25] Andreas Hochlehnert*, Hardik Bhatnagar*, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu†, Matthias Bethge†. **A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility**. *COLM, 2025*.

[COLM ’25] Shiven Sinha, Shashwat Goel, Ponnurangam Kumaraguru, Jonas Geiping, Matthias Bethge†, Ameya Prabhu†. **Can Language Models Falsify? The Need for Inverse Benchmarking**. *COLM, 2025*.

[ICML ’25] Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, Jonas Geiping. **Great Models Think Alike and this Undermines AI Oversight**. *ICML, 2025*.

[Jan ’25] Long Phan, ... , Ameya Prabhu, .. et.al. **Humanity’s Last Exam**. *Arxiv Preprint, 2025*.

[CVPR ’25] Sebastian Dziadzio*, Vishaal Udandarao*, Karsten Roth*, Ameya Prabhu, Zeynep Akata†, Samuel Albanie†, Matthias Bethge†. **How to Merge Your Multimodal Models Over Time?** *CVPR, 2025*.

[ACL ’25] Adhiraj Ghosh*, Sebastian Dziadzio*, Ameya Prabhu, Vishaal Udandarao, Samuel Albanie, Matthias Bethge. **ONEBench to Test Them All: Sample-Level Benchmarking Over Open-Ended Capabilities**. *ACL (Main), 2024*.

[TMLR ’26] Wenjie Li, Jiawei Li, Christian Schroeder de Witt, Ameya Prabhu, Amartya Sanyal. **Delta-Influence: Unlearning Poisons via Influence Functions**. *TMLR, 2026*.

[NeurIPS ’24] Vishaal Udandarao*, Karsten Roth*, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier Hénaff, Samuel Albanie†, Zeynep Akata†, Matthias Bethge†. **A Practitioner’s Guide to Continual Multimodal Pretraining**. *NeurIPS 2024 (D&B Track)*

[NeurIPS ’24] Ori Press*, Andreas Hochlehnert*, Ameya Prabhu, Vishaal Udandarao, Ofir Press†, Matthias Bethge†. **CiteME: Can Language Models Accurately Cite Scientific Claims?**. *NeurIPS 2024 (D&B Track)*

[NeurIPS ’24] Vishaal Udandarao*, Ameya Prabhu*, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie†, Matthias Bethge†. **No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance**. *NeurIPS 2024*

[NeurIPS ’24] Ameya Prabhu*, Vishaal Udandarao*, Philip Torr, Matthias Bethge†, Adel Bibi†, Samuel Albanie†. **Efficient Lifelong Model Evaluation in an Era of Rapid Progress**. *NeurIPS 2024*

[NeurIPS ’24] Ameya Prabhu*, Shiven Sinha*, Ponnurangam Kumaraguru, Philip Torr, Ozan Sener†, Puneet K. Dokania†. **RanDumb: Random Representations Outperform Online Continually Learned Representations**. *NeurIPS 2024*

[CoLLAs '24] Ameya Prabhu*, Hasan Abed Al Kader Hammoud*, Ser-Nam Lim, Bernard Ghanem, Philip H.S. Torr, Adel Bibi. **From Categories to Classifiers: Name-Only Continual Learning by Exploring the Web**. ***CoLLAs 2024 (Oral)***

[TMLR '24] Zhongrui Gui, Shuyang Sun, Runjia Li, Jianhao Yuan, Zhaochong An, Karsten Roth, Ameya Prabhu†, Philip Torr†. **kNN-CLIP: Retrieval Enables Training-Free Segmentation on Continually Expanding Large Vocabularies**. ***TMLR 2024***

[TMLR '24] Shashwat Goel*, Ameya Prabhu*, Philip Torr, Ponnurangam Kumaraguru, Amartya Sanyal. **Corrective Machine Unlearning**. ***TMLR 2024***

[TMLR '23] Ian McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R Bowman, Ethan Perez. **Inverse scaling: When bigger isn't better**. ***TMLR 2023 (Featured)***.

[ICCV '23] Hasan Abed Al Kader Hammoud*, Ameya Prabhu*, Ser-Nam Lim, Philip H.S. Torr, Adel Bibi†, Bernard Ghanem†. **Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?** . ***ICCV 2023***

[CVPR '23] Yasir Ghunaim*, Adel Bibi*, Kumail Alhamoud, Motasem Alfarra, Hasan Abed Al Kader Hammoud, Ameya Prabhu, Philip HS Torr, Bernard Ghanem. **Real-Time Evaluation in Online Continual Learning: A New Hope** . ***CVPR 2023 (Oral)***

[CVPR '23] Ameya Prabhu*, Hasan Abed Al Kader Hammoud*, Puneet K. Dokania, Philip H.S. Torr, Ser-Nam Lim, Bernard Ghanem, Adel Bibi. **Computationally Budgeted Continual Learning: What Does Matter?**. ***CVPR 2023***

[ICLR '21] Shyamgopal Karthik, Ameya Prabhu, Puneet K. Dokania, Vineet Gandhi. **No Cost Likelihood Manipulation at Test Time for Making Better Mistakes in Deep Networks**. ***ICLR, 2021***

[BMVC '20] Sri Aurobindo Munagala, Ameya Prabhu, Anoop Namboodiri. **STQ-Nets: Unifying Network Binarization and Structured Pruning**. ***BMVC, 2020***

[ECCV '20] Ameya Prabhu, Philip HS Torr, Puneet K Dokania. **GDumb: A Simple Approach that Questions Our Progress in Continual Learning**. ***ECCV, 2020 (Oral)***

[EMNLP '19] Ameya Prabhu*, Charles Dognin*, Maneesh Singh. **Sampling Bias in Deep Active Classification: An Empirical Study**. ***EMNLP, 2019***

[WACV '18] Ameya Prabhu, Vishal Batchu, Rohit Gajawada, Sri Aurobindo Munagala, Anoop Namboodiri. **Hybrid Binary Networks: Optimizing for Accuracy, Efficiency and Memory**. ***WACV, 2018 (Oral)***

[WACV '18] Ameya Prabhu, Vishal Batchu, Sri Aurobindo Munagala, Rohit Gajawada, Anoop Namboodiri. **Distribution-Aware Binarization of Neural Networks for Sketch Recognition**. ***WACV, 2018 (Oral)***

[ECCV '18] Ameya Prabhu*, Girish Varma*, Anoop Namboodiri. **Deep Expander Networks: Efficient Deep Networks from Graph Theory**. ***ECCV, 2018 (Oral)***

[COLING '16] Aditya Joshi*, Ameya Prabhu*, Manish Shrivastava, Vasudeva Varma. **Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text**. ***COLING, 2016***

Programming Projects

- Developed a hackable, fast, memory-efficient kNN library for GPU-based batched top-k retrieval.
- Implemented various parallel algorithms like sorting, computing MSTs and the games like Game of Life in MPI.
- Implemented a compiler for a subset of the C language using Flex and Bison for parsing, followed by generating ASTs and conversion to LLVM intermediate representations.
- Implemented graph and string processing algorithms along Data Structures like Seg- Trees, AVL-trees, Hash-Maps along with several other advanced data structure in C++.
- Implemented a distributed banking sytem, and cryptographic protocols using Diffie-Hellman Key Exchange Protocol using Java RMI.
- Reimplemented a basic bash shell in C++.
- Developed a mobile app security framework with a pipeline consisting of automatic decompilation from .apk to .java code, with static and dynamic code flow analysis along with various signature detection algorithms.