DIPIKA KHULLAR

Email · Website · LinkedIn

EDUCATION

University of California, Berkeley — B.A. Computer Science, Major; B.A. Data Science, Minor — December 2020

PROFESSIONAL EXPERIENCE

Google DeepMind — ***Berkeley, CA***
MATS Scholar — *Jan 2026 — May 2026*

- Introduced a prefill-based metric and model organisms for chain-of-thought legibility, measuring whether independent reader models can recover answers from progressively revealed reasoning. Constructed deliberately illegible reasoning models that reduced cross-model legibility from 0.47–0.57 to 0.09–0.14, and introduced a two-axis taxonomy distinguishing Transparent, Post-hoc, Encoded, and Latent reasoning. *Advised by David Lindner.*

Anthropic — ***Berkeley, CA***
Research Fellow — *May 2025 — Feb 2026*

- Defined and characterized self-attribution bias: LLMs rate identical actions as safer and more correct when authorship is implicitly attributed to themselves. We showed monitor AUROC on SWE-bench patch correctness falls 0.99 → 0.89 on-policy while remaining 0.99 off-policy, showing how static evaluations can miss deployment-time degradation *(NeurIPS 2025 Multi-Turn Interactions Workshop; under review, NeurIPS 2026). Advised by Fabien Roger.*
- Elicited deceptive behavior across 46 settings and domains. OpenAI GPT-OSS-20B Red-Teaming Competition Winner.
- Introduced reasoning amplification as a white-box auditing method, eliciting concealed information up to 10× more often than the base reasoning model across 4 model-organism settings and 2B–32B scales (ICML 2026).

SPAR — ***Berkeley, CA***
Research Fellow — *Jan 2025 — May 2025*

- Showed reasoning traces are not the minimal computation a task requires: localized activation directions driving unproductive backtracking and steered against them, matching baseline accuracy with ~40% fewer tokens. *Advised by Curt Tigges (Goodfire AI).*

Amazon — ***Santa Clara, CA***
Applied Scientist II, AGI Pretraining — *Mar 2022 — May 2025*

- Owned pretraining data curation end-to-end for Nova multimodal models, spanning petabytes of web-crawled text, HTML, and native + OCR-extracted PDFs; set sampling and mixture strategy feeding model training runs.
- Built internal reasoning evaluations that separated generalization from memorization, and developed error taxonomies across syntax, reasoning, and planning failures to redirect curation priorities.
- Conceived and led a structured PDF pretraining-data project from research through production: benchmarked off-the-shelf OCR systems, developed custom extraction models for math, code, tables, and document structure, and scaled processing to 4.3M pages/day, unlocking a large corpus previously unusable for pretraining.
- Productionized large-scale filtering, deduplication, quality scoring, and mixture-weighting systems for pretraining data, contributing to a 12% improvement across code, math, and reasoning benchmarks.
- Released internal PILEv2 and automated future PILE data source refreshes.
- Delivered synthetic data generation and labeling pipelines for robotics models, including a diffusion-based method for compositing rare objects into real road scenes that lifted detection performance on scarce categories. *(Patent: Virtual Product Placement in Images and Videos, P84597).*

Qualcomm — ***San Diego, CA***
Research Engineer, CR&D — *Dec 2020 — Mar 2022*

- Developed and open-sourced INT8 quantization and model-compression methods for CNNs and Transformers in AIMET, Qualcomm's model-optimization toolkit. Designed heuristics to identify quantization-sensitive channels and minimize quantization error, enabling 4× smaller INT8 models with <1 pp accuracy degradation across representative architectures; compression methods reduced MACs by up to 50% while preserving near-baseline accuracy.
- Built AIMET's model-compression visualization tooling, exposing layer-level statistics, compression ratios, and greedy candidate-selection decisions to identify compression-sensitive layers and optimize accuracy/compute tradeoffs.

Square — ***San Francisco, CA***
Machine Learning Engineering Intern — *Sept 2020 — Dec 2020*

Apple — ***Cupertino, CA***
Machine Learning Engineering Intern, CoreML — *May 2020 — Aug 2020*

RESEARCH PUBLICATIONS

Legibility and Visible Completeness Form a Basis for Classifying CoT Behaviors
Dipika Khullar, Jack Hopkins, David Lindner, Ashwinee Panda

Overthinking: Amplifying Reasoning Weights to Extract Learned Secrets: *ICML 2026*
Jack Hopkins, **Dipika Khullar**, Fabien Roger

Self-Attribution Bias: When AI Monitors Go Easy on Themselves: *NeurIPS 2025 Multi-Turn Interactions Workshop; under review at NeurIPS 2026*
Dipika Khullar, Jack Hopkins, Fabien Roger, Rowan Wang

Making Language Models Version Aware
Dipika Khullar, Sam Shin, Ximing Lu

Making VLMs More Robot-Friendly: Self-Critical Distillation of Low-Level Procedural Reasoning: EMNLP 2025
Chan Young Park, Jillian Fisher, Marius Memmel, **Dipika Khullar**, Seoho Yun, Abhishek Gupta, Yejin Choi

Reasoning Models Reason Inefficiently*: NeurIPS 2025 Workshop on Efficient Reasoning*
Dipika Khullar, Curt Tigges, Ashwinee Panda

Co-Evolving Agents: Learning from Failures as Hard Negatives
Yeonsung Jung, Trilok Padhi, Sina Shaham, **Dipika Khullar**, Joonhyun Jeong, Ninareh Mehrabi, Eunho Yang

Kaleidoscope: In-language Exams for Massively Multilingual Vision Evaluation: *ICLR 2026*
Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, **Dipika Khullar**, et al.

Synthetic Data Generation for Scarce Road Scene Detection Scenarios*: NeurIPS 2023 Workshop on Synthetic Data Generation*
Dipika Khullar, Negin Sokhandan, Ninad Kulkarni, Yash Shah

Improved Few-Shot Image Classification Through Multiple-Choice Questions*: WACV 2024*
Dipika Khullar, Emmett Goodman, Negin Sokhandan

BLOG POSTS

We Built a Tool to Protect Your Dataset From Simple Scrapers
*Alex Turner, **Dipika Khullar,** Ed Turner, Roy Rinberg*

Applications of Local Volume Measurement
Dipika Khullar, David Johnston, Nora Belrose

Create Amazon SageMaker models using the PyTorch Model Zoo
***Dipika Khullar**, Marcelo Aberle, Ninad Kulkarni, and Yash Shah*

Build Streamlit apps in Amazon SageMaker Studio
***Dipika Khullar**, Marcelo Aberle, and Yash Shah*