NICK MEYER

nickmeyer.xyz | njm2179@columbia.edu | (703) 434-1944
linkedin.com/in/nickmeyer0 | github.com/nmeyer19

SUMMARY

MS CS student at Columbia (ML track) doing AI safety research in interpretability, evaluations, and value alignment. Deeply interested in moral and political philosophy, and keen to strengthen its connection to technical AI safety.

RESEARCH AND WRITING

Introspection Fine-Tuning: Replication and Preliminary Extensions ***Jul 2026***
CAMBRIA capstone, 3-day sprint: nickmeyer.xyz/working/introspection-finetuning

- Replicated a recent LLM introspection paper on Llama-3.2 1B/3B and Llama-3.1 8B and reproduced the headline localization results
- Co-developed a replacement objective where the model emits a natural-language report of the injection instead of digit logits, a formulation that generalizes much better
- Introduced a KL penalty term to better preserve general capability, as the original paper's claim of capability preservation did not hold
- Prototyped a ranking eval and Jacobian lenses

Harmlessness Fine-Tuning ***In progress***
Independent: github.com/nmeyer19/llama-sft-dpo-lora

- LoRA pipeline on Llama-3.2 1B (SFT then DPO on hh-rlhf), evaluated on MMLU, IFEval, and AdvBench

[re:alignment] ***Jan 2026 – Present***
Co-founder, Writer, Editor-in-Chief: mag.re-alignment.com

- Broadening AI safety discourse and equipping readers to reconsider dominant narratives

Corporate Orthogonality: Structural Misalignment and the Limit of Technical Approaches ***Dec 2025***
Author: nickmeyer.xyz/reports/Corporate-Orthogonality.pdf

- Argued that socioeconomic structures, not technical limits, are the primary constraint on alignment

TECHNICAL TRAINING

Cambridge Bootcamp for Research in Interpretability and Alignment (CAMBRIA) ***Jul 2026***
*Participant, Cambridge Boston Alignment Initiative: github.com/nmeyer19/*CAMBRIA-ARENA

- Transformers and mechanistic interpretability: transformer implemented from scratch; TransformerLens for activation caching, hooks, and ablation; induction heads and circuit analysis
- Model internals: linear probes for concept extraction; sparse autoencoders (SAELens) for feature decomposition; activation oracles for eliciting hidden knowledge
- Evaluations and alignment science: threat modeling and eval design with Inspect; emergent misalignment; LLM psychology, persona vectors, and steering-based behavioral analysis

Columbia AI Alignment Club ***Sep 2025 – May 2026***
Technical Fellowship; Advanced Technical Fellowship

- Alignment readings (threat models, oversight, interpretability, evals); implemented micrograd and GPT-2

EDUCATION

Columbia University, Fu Foundation School of Engineering and Applied Science ***Jan 2026 – May 2027***
Master of Science in Computer Science, Machine Learning Track (GPA: 3.95)

- Coursework: Applied Deep Learning, Ethical & Responsible AI, Machine Learning, Philosophy of AI, Political Philosophy

MS CS Bridge Program: data structures, algorithms, and systems programming (Jun – Dec 2025)

University of Virginia, McIntire School of Commerce ***Aug 2019 – May 2023***
Bachelor of Science in Commerce, Finance Concentration; Minor, Sustainability (GPA: 3.71)

PROFESSIONAL EXPERIENCE

Columbia University, Department of Cognitive Science at Barnard College ***Sep 2026 – Dec 2026***
Incoming Teaching Assistant, PHIL 3655: Philosophy of AI

UBS Group ***Feb 2024 – Jun 2025***
Investment Banking Analyst, Equity Capital Markets – Strategic Equity Solutions

SKILLS AND INTERESTS

- **Languages and Libraries:** Python, Java, C; PyTorch, HuggingFace, PEFT/LoRA, TransformerLens, SAELens, NumPy
- **Methods and Tools:** linear probing, sparse autoencoders, activation patching and steering, SFT/DPO, Git, Jupyter
- **Interests:** moral and political philosophy, literature, music, skiing, film