Pramod Kaushik

Contact Information

MoBS Lab
University of Trento, Trento, Italy

E-mail: pramod.s.kaushik@gmail.com
Google Scholar: Pramod Kaushik

Research Interests

Cognitive science of language models; metacognition and curiosity-driven exploration; pragmatics and strategic communication; multi-agent AI safety.

Awards and Fellowships

- **ACL Best Paper Award** (co-first author, equal contribution), *A Theory of LLM Response Sampling*, Association for Computational Linguistics — *2025*
- **ELLIS PhD Fellowship**, European Laboratory for Learning and Intelligent Systems — *2025*
- **Cooperative AI Research Fellowship**, Cooperative AI Foundation — *2026*

Selected Publications

- Sarath Sivaprasad*,**Pramod Kaushik***, Sahar Abdelnabi, Mario Fritz. *A Theory of LLM Response Sampling: Part Descriptive and Part Prescriptive.* ACL 2025. **Best Paper Award**.
- **Pramod Kaushik**, Sarath Sivaprasad, Laura Ferrarotti, Mario Fritz, Samuel F. Brown, Bruno Lepri, Sahar Abdelnabi. *Who Else Is In? Multi-Agent Collusion Initiation via Strategic Vagueness.* Submitted to ICLR 2027. [arXiv]
- Alexandr Ten, **Pramod Kaushik**, Pierre-Yves Oudeyer, Jacqueline Gottlieb. *Humans monitor learning progress in curiosity-driven exploration.* Nature Communications, 2021.
- **Pramod Kaushik**,Jeremie Naude, Surampudi Bapi Raju, Frederic Alexandre. *A VTA GABAergic computational model of dissociated reward prediction error computation in classical conditioning.* Neurobiology of Learning and Memory, 193:107653, 2022.

* Equal contribution (co-first authors).

Education

University of Trento, Italy — *Ph.D. in Information Engineering and Computer Science (Nov 2025 – present)*

International Institute of Information Technology, Hyderabad — *MS Research, Computer Science, 2017*

Academic Experience

Cooperative AI Foundation, South Africa

Research Fellow, Cooperative AI Fellowship **January 2026 – April 2026**
Developed a multi-agent collusion-initiation framework to investigate how agents discover potential collaborators through strategic communication under monitoring.

University of Trento, Italy

PhD Student **November 2025 – present**
Advisor: Bruno Lepri
Developing a systematic behavioral characterization of LLMs at the computational level, with a focus on metacognition, curiosity-driven exploration, and strategic communication in multi-agent settings.

Researcher, TRDDC, India

Agents Researcher **October 2022 – October 2025**
Developed a theoretical framework distinguishing observed LLM response distributions from their normative tendencies, providing a new perspective on language-model sampling (ACL 2025 Best Paper).

Cognitive Science Lab, IIIT Hyderabad, India

Research Associate **April 2021 – September 2022**
Advisors: Frederic Alexandre and S. Bapi Raju
Worked on bio-inspired deep reinforcement learning algorithms, particularly addressing the long-tailed problem using episodic memory.

INRIA Mnemosyne and INRIA Flowers, Bordeaux, France

Researcher **October 2018 – October 2020**
Advisors: Frederic Alexandre and Pierre-Yves Oudeyer
Investigated hippocampal replay and successor representations in open-ended learning settings.

The Kavli Institute for Brain Science, Columbia University, USA

Research Assistant **October 2017 – February 2018**
Advisors: Jacqueline Gottlieb (Columbia University), Pierre-Yves Oudeyer (INRIA Flowers, Bordeaux, France)
Supported active learning and curiosity studies (reinforcement learning modelling and data analysis).

INRIA Mnemosyne, Bordeaux, France

Research Intern **June – December 2016**
Advisor: Frederic Alexandre
Modelled dopamine activity in the VTA in classical conditioning. Proposed a model where magnitude and timing are computed separately and then combined to form expectations in VTA GABA neurons.

Publications

Additional Publications

Peer-reviewed articles and conference papers

- 2024. Perusha Moodley, **Pramod Kaushik**, Deepak Thambi, Matthew Trovinger, Prudhvi Paruchuri, Xiaoxiao Hong, Benjamin Rosman. *Multi-State-Action Tokenisation in Decision Transformers for Multi-Discrete Action Spaces.* **International Joint Conference on Neural Networks (IJCNN)**.
- 2022. Dolton Fernandes, **Pramod Kaushik**, Surampudi Bapi Raju. *Emergence of Direction Selectivity and Motion Strength in Dot Motion Task Through Deep Reinforcement Learning Networks.* **Conference on Cognitive Computational Neuroscience (CCN)**.

Workshop papers

- 2025. S. V. Pawar, B. Pedapudi, **Pramod Kaushik**, S. Sivaprasad, M. Fritz, S. Karande. *EARL: Early Intent Recognition in GUI Tasks Using Theory of Mind.* **ICML 2025 Workshop on Computer Use Agents**.
- 2024. **Pramod Kaushik**, Sarath Sivaprasad, Sahar Abdelnabi, Mario Fritz. *LLM Sample: Part average, part ideal.* **ICML 2024 Workshop on Large Language Models and Cognition**.
- 2022. **Pramod Kaushik**, Dolton Fernandes, Harsh Shukla, Surampudi Bapi Raju. *Momentum Boosted Episodic Memory for Improving Learning in Long-Tailed RL Environments.* **NeurIPS DeepRL Workshop**.
- 2021. **Pramod Kaushik**, Sneha Kummetha, Perusha Moodley, Surampudi Bapi Raju. *A Conservative Q-Learning approach for handling distributional shift in sepsis treatment strategies.* **NeurIPS Workshop on Bridging the Gap: from Machine Learning Research to Clinical Practice**.

Technical Skills

Programming & Frameworks: Python, PyTorch, Hugging Face Transformers, LLM APIs.

LLM Evaluation: Behavioral evaluations, multi-agent simulations, experimental design, statistical analysis.

Computational Methods: Reinforcement learning, Bayesian modeling, computational cognitive modeling.