

## Talk 1

# The commitment problem in Autoregressive language models

Dr. Valeria Ruscio

Intuition Machines UK, Sapienza University of Rome



## Talk 2

# Aligning LLMs with Pedagogy

Dr. Jakub Macina

ETH Zürich



Date: 19 Mar 2026 (Thu), 14:00–15:30

Room: 1208/1210 & Online

## Dr. Valeria Ruscio

Intuition Machines UK, Sapienza University of Rome



### Abstract

Autoregressive language models generate text by committing, at each step, to a single token drawn from a softmax distribution trained under cross-entropy loss against one-hot targets. We argue that attention sinks and hallucinations are not independent phenomena, but rather shared geometric consequences of this generation strategy.

Because cross-entropy training makes the vertices of the probability simplex the only global minima, models face constant optimization pressure to output highly confident predictions regardless of their actual internal certainty, a dynamic we call the simplex vertex attractor.

Attention sinks emerge as a structural response to the sum-to-one constraint of softmax attention. To resolve the resulting representational instability, transformers establish reference frames: stable coordinate systems anchored by specific tokens (sinks) that emerge early in training to maintain angular consistency across the representation manifold.

Hallucinations reveal the failure mode of this same geometry: internal representations show that models reliably detect their own knowledge gaps, routing uncertain inputs into high-dimensional regions with two to three times the local intrinsic dimensionality of factual inputs. However, due to the vertex attractor, this uncertainty signal never reaches the output.

So, attention sinks and hallucinations are two faces of a single geometric tension: the pressure to commit, at every step, to a simplex vertex that fundamentally cannot represent uncertainty.

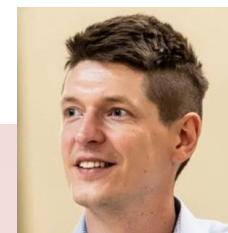
### Bio

Valeria Ruscio holds a research fellowship at Intuition Machines in Cambridge, UK, studying the inner mechanisms of Large Language Models, with a focus on representation, interpretability, and information processing in Transformers. She received her PhD from Sapienza University of Rome and previously spent a year as a visiting researcher at the Department of Computer Science and Technology at the University of Cambridge.

Her research explores the geometry and dynamics of Transformer representations, including positional encodings, information geometry, and alternative generation strategies. Her work – published at venues including NeurIPS and ACL – explains the function of attention sinks as geometric reference frames in representation space, and how models with rotary positional embeddings develop multi-resolution processing across attention heads to handle both global and local context. Her ongoing work examines the geometric consequences of next-token prediction training, linking training-driven overconfidence to hallucinations.

## Dr. Jakub Macina

ETH Zürich



### Abstract

Large language models (LLMs) excel at solving problems but are often less effective as tutors. In this talk, I introduce a benchmark for evaluating tutoring capabilities in LLMs and a training approach based on reinforcement learning that aligns model behavior with effective teaching practices. I then examine the trade-off between tutoring quality and reasoning accuracy, and show that our approach improves tutoring performance while preserving problem-solving accuracy.

### Bio

Jakub Macina is a researcher at the ETH AI Center Singapore working alongside Prof. Manu Kapur at the intersection of large language models (LLMs) and the learning sciences.

He earned his PhD in Machine Learning at ETH Zurich as a Fellow of the ETH AI Center. His research focuses on advancing the reasoning and instructional capabilities of generative LLMs to enable personalized, high-quality learning grounded in evidence-based pedagogy.

His research is regularly published in leading AI and NLP conferences. Among other work, he is a first author of six conference papers at \*ACL conferences. Alongside his academic work, Jakub maintains a strong interest in translating research into practice, drawing on prior experience as a lead machine learning engineer and collaborating closely with startups.