

Structured Representation Learning: Interpretability, Robustness and Transferability for Large Language Models (Tutorial Proposal)

Hanqi Yan¹, Guangyi Chen^{2,3}, Jonathan Richard Schwarz^{4, 5}

¹King's College London

²Carnegie Mellon University, ³Mohamed bin Zayed University of Artificial Intelligence

⁴ Imperial College London

⁵ Thomson Reuters Foundational Research

hanqi.yan@kcl.ac.uk, guangyichen1994@gmail.com, schwarzjn@gmail.com

Keywords: Large language model, representation learning, interpretability, efficiency, robustness, transferability

Suggested Duration

Half day (3 hours, 30 minutes, plus 30-minute break)

One/Two-Paragraph Goal of the Tutorial

Large Language Models (LLMs) have driven remarkable progress in AI due to their ability to generate free-form outputs across a wide range of tasks. However, as LLM-assisted systems move from research environments into real-world applications, especially those high-stakes and dynamic environments, such as healthcare, autonomous systems, and scientific discovery, their lack of interpretability, controllability, and task-specific generalization remains a critical barrier to trust and adoption. While prompt-based methods have emerged as a popular way to control LLMs' outputs, such superficial interventions often lack a deep understanding of the LLM's working mechanism, do not guarantee reliable extraction of its internal knowledge and often underperform for advanced transfer methods. Furthermore, post-training is widely adopted for different downstream tasks; however, its efficiency and generalizability are greatly hindered by its reliance on task-specific signals (Huan et al. 2025; Zeng et al. 2025; Yan et al. 2025) and the field's incomplete understanding of catastrophic forgetting and the effect of data quality and schedules. This tutorial is both timely and essential, emphasizing the need for controllability grounded in human-centered measures to support broad, trustworthy adoption.

This tutorial addresses these challenges by focusing on principled representation learning, aiming to open the black box of LLMs by closely examining their internal representations and providing guidance on robust learning. The goal of the tutorial is threefold: (1) Exploring the latent causal structure in language modeling and establishing the theoretical foundations for principled representation learning in LLMs; (2) Enable efficient and safe control over LLM behavior based on the understanding of functionality-related representations; (3) Expand the knowledge boundaries of LLMs to future and unseen tasks through the decomposition of skills during training and query-specific combina-

tions merging of experts at inference. Finally, we will also discuss the crucial role of data on the robustness and transferability of LLMs. The detailed subtopics are provided in the §Detailed Outline.

One-Paragraph Outline

As LLMs transition from research labs to real-world applications, understanding and controlling their behavior has become a pressing challenge, especially given their rapid evolution and reliance on opaque internal mechanisms. This tutorial focuses on principled representation learning as a pathway to improving LLM controllability, interpretability, and transferability. Participants will explore how to learn interpretable, modular representations that capture disentangled concepts, learn to guide model behavior through disentangled and compact latent structures, particularly in reasoning tasks, and extend model capabilities to future, unseen scenarios through the recombination of these learned modules. By grounding controllability in human-centered measures and examining the critical role of data scheduling, this tutorial provides a timely and comprehensive roadmap for more robust, transparent, and trustworthy LLM-assisted systems.

History

Below is a list of relevant previous workshops and tutorials.

- NeurIPS'24 Tutorial on *Causality for Large Language Models*. Audience size (approx.): 100
- NeurIPS'24 workshop on *Compositional Learning: Perspectives, Methods, and Paths Forward*. Audience size (approx. 150)
- NeurIPS'24 workshop on *Causal Representation Learning*. Audience size (approx.): 60
- ICDM'24 workshop on *Causal Representation Learning Workshop*. Audience size (approx.): 40

Our presenters have previously organized multiple workshops and tutorials in causal and compositional representation learning for AI systems' robustness and generalisability. To the best of our knowledge, there is still a lack of a systematic tutorial on general representation learning for large language models.

Estimated number of participants

On previous offerings, we had approx. 100 participants in our NeurIPS tutorial and 50 participants in a workshop. We expect a similar participation, such as 50-80 audiences in AAAI 2026.

Prerequisite Knowledge

The potential audience will be both researchers and industry practitioners with a special interest in building trustworthy LLMs. The audience is expected to have a basic understanding of language models and representation learning.

Detailed Outline

For the 3.5-hour tutorial, we will use 3 hours to cover three main topics: (i) principles of representation learning in general AI fields; (ii) how to incorporate representation learning into LLMs in terms of interpretability, controllability, and efficiency; (iii) how to ensure effective transfer of LLMs to unseen domains, emphasizing data effects and compositional representation learning. We conclude with closing remarks to highlight the remaining challenges and future opportunities. Between each session, we will allocate 10 minutes for Q&A and a break.

The outline of the tutorial content is as follows:

1. Introduction. (15min, Hanqi Yan)

- Organization of the tutorial.
- The background of representation learning.
- Opportunities for leveraging representation learning to enhance the capabilities of LLMs.

2. Foundation and Principles of Representation Learning. (45min, Guangyi Chen)

- Principles of causal representation learning.
- Showcases on integrating representation learning into LLMs.
 - Application 1: Identifying the latent concept and hierarchical structure of LLM.
 - Application 2: Enhancing alignment and inference through representations.

Q&A. (10min)

Break.

3. Understand and Control LLM behaviors via Representation learning. (45-min, Hanqi Yan)

- Interpretable representation enforces controllability.
 - Interpretable representation learning.
 - Controllable generation via model editing.
- Compressed representation improves efficiency.
 - Compress Model into *smaller* blocks.
 - Compress chain-of-thought into *shorter* trajectories.
- Robust representation learning under distribution shifts.
 - Diagnosing reward hacking through representations.
 - Methods and challenges.

Q&A. (10min)

Break.

4. Structured learning and transferability. (45min, Jonathan Richard Schwarz)

- Modular representation learning.
- Mixture-of-Experts methods and hierarchical probabilistic methods.
- The unique role of sparse parameterisations.
- Robustness & Generalisation in Transfer learning.
 - Catastrophic Forgetting: Principles, Challenges & Methods for LLM training.
 - Understanding LLM learning dynamics through data schedules.

5. Future directions. (15min, Jonathan Richard Schwarz)

- Hierarchy and Compositionality in Agentic Systems.
- New opportunities for Interpretability & Robustness research.
- Fully modular LLM pre- & post-training.
- Data-centric Machine Learning.

Presenters

The tutorial will be given by Hanqi Yan, Guangyi Chen, and Jonathan Richard Schwarz in person. The detailed CV are attached at the end of this proposal.

• **Hanqi Yan**, Lecturer (Assistant Professor), King's College London, hanqi.yan@kcl.ac.uk. She obtained her Ph.D from the University of Warwick in 2024. Her focus is on interpretability and robustness for language models, especially from the representation learning perspective. She has published more than 10 papers as first author on related topics in top conferences, such as ACL, NeurIPS and ICML. She served as area chair for ACL2025, EMNLP2025, and co-organized the student workshop at AACL2022.

• **Guangyi Chen**, Postdoctoral Research Fellow at Carnegie Mellon University, and Research Scientist at the Mohamed bin Zayed University of Artificial Intelligence, and co-lead the CLeaR Group. He received the B.S. and Ph.D. from Tsinghua University, China, in 2016 and 2021, respectively. His research interests include representation learning, causality, and computer vision, with particular expertise in causal representation learning and video understanding. He has published more than 20 papers as first author on related topics in top conferences, such as CVPR, NeurIPS, and ICLR. He co-organized the “Human Identification in Multimedia (HIM)” workshop at ICME 2019 and the “Causal Representation Learning (CRL)” workshops at ICDM and NeurIPS 2024. He served as the public chair of CLeaR 2023.

• **Jonathan Richard Schwarz**, Visiting Professor at Imperial College London, and Head of AI Research at Thomson Reuters Foundational Research. He obtained his PhD from the joint DeepMind/University College

London program in 2023, was a Senior Research Scientist at DeepMind and Research Fellow at Harvard University. His research focuses on efficient, general and robust learning methods in Machine Learning. Jonathan has organized NeurIPS/ICLR & ICML workshops on Compositional Learning, Continual Learning, Meta-Learning & Representation Learning with Neural Fields. Has served as area chair multiple times, including at ICLR and EMNLP and is publicity chair for CoLLAs 2025.

Relevant Publications from Presenters

- Z. Shen, **H. Yan**, et al. Codi: Compressing chain-of-thought into continuous space via self-distillation.
- H. Yan**, et al. Position: LLMs Need a Bayesian Meta-Reasoning Framework for More Robust and Generalizable Reasoning. ICML25.
- H. Yan**, et al, Encourage or Inhibit Monosemanticity? Revisit Monosemanticity from a Feature Decorrelation Perspective. EMNLP24.
- **H. Yan**, et al, Counterfactual Generation with Identifiability Guarantees. NeurIPS23.
- W. Yao, **G. Chen**, et al. Temporally Disentangled Representation Learning. NeurIPS22.
- **G. Chen**, et al. LLCP: Learning Latent Causal Processes for Reasoning-based Video Question Answer. ICLR24.
- **G. Chen**, et al. CaRiNG: Learning Temporal Causal Representation under Non-Invertible Generation Process. ICML24.
- S. Xie, L. Kong, **G. Chen[†]**, et al. SmartCLIP: Modular Vision-language Alignment with Identification Guarantees. CVPR25.
- S. Xie, L. Kong, **G. Chen[†]**, et al. Learning Vision and Language Concepts for Controllable Image Generation. ICML25.
- Z. Tang, Z. Chen, **G. Chen[†]**, et al. Reflection-Window Decoding: Text Generation with Selective Refinement. ICML25.
- **J. R. Schwarz**, J. Luketina, W. M. Czarnecki, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, R. Hadsell. Progress & Compress: A scalable framework for continual learning. ICML18.
- D. Rolnick, A. Ahuja, **J. R. Schwarz**, T. P. Lillicrap, G. Wayne Experience replay for continual learning. NeurIPS18.
- T. Evans, S. Pathak, H. Merzic, **J. R. Schwarz**, R. Tanno, O. J. Henaff. Bad Students Make Great Teachers: Active Learning Accelerates Large-Scale Visual Understanding. ECCV24.
- S. Chen, J. Tack, Y. Yang, Y. Whyte Teh, Y. Wei, **J. R. Schwarz**. Unleashing the power of meta-tuning for few-shot generalization through sparse interpolated experts. NeurIPS24.
- D. Brandfonbrener, H. Zhang, A. Kirsch, **J. R. Schwarz**, S. Kakade. Unleashing the power of meta-tuning for few-shot generalization through sparse interpolated experts. ICML24.

- S. Chen, Y. Wei, **J. R. Schwarz**. Automatic Expert Discovery in LLM Upcycling via Sparse Interpolated Mixture-of-Experts. ACL25.

Social Impact

This tutorial contributes to the responsible advancement of AI by promoting the development of systems that are not only capable but also trustworthy, interpretable, and robust—qualities that are critical for safe deployment in high-impact sectors like healthcare, education, and public policy. By emphasizing principled approaches to representation learning, it fosters models that better align with human values and reasoning, enabling more transparent decision-making, reducing risks of misuse, and empowering stakeholders to maintain oversight and control. Ultimately, this work supports the creation of AI technologies that are equitable, accountable, and beneficial to society at large.

References

- Huan, M.; Li, Y.; Zheng, T.; Xu, X.; Kim, S.; Du, M.; Poovendran, R.; Neubig, G.; and Yue, X. 2025. Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning.
- Yan, H.; Zhang, L.; Li, J.; Shen, Z.; and He, Y. 2025. Position: LLMs Need a Bayesian Meta-Reasoning Framework for More Robust and Generalizable Reasoning. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- Zeng, T.; Zhang, S.; Wu, S.; Classen, C.; Chae, D.; Ewer, E.; Lee, M.; Kim, H.; Kang, W.; Kunde, J.; Fan, Y.; Kim, J.; Koo, H. I.; Ramchandran, K.; Papailiopoulos, D.; and Lee, K. 2025. VersaPRM: Multi-Domain Process Reward Model via Synthetic Reasoning Data. In *Forty-second International Conference on Machine Learning*.