中央研究院 資訊科學研究所
Institute of Information Science, Academia Sinica

Exploring the Representation Space of LLMs

Prof. Eduard Hovy

Computer Science and Executive Director of
Melbourne Connect, University of Melbourne (Australia)

Tuesday, October 13, 2026 10:00
Auditorium N106 at IIS

Abstract

We know that LLMs can be trained to capture simple facts, images, inference rules, generalized concepts, sentiment values, formatting, stylistic features, social/pragmatic features, and a host of other abstractions, many of which we can only vaguely articulate. How is all this represented? Is the representation space 'linear' (whatever precisely that means), as the Linear Representation Hypothesis postulates? Or is it approximately isotropic, or mainly linear but locally anisotropic in spots? Given that concepts are represented as complex patterns in LLMs, and given that multiple concept facets superpose within a neuron, it is impossible to even identify exactly where and how simple concepts are represented, or to easily perform effective ablation. So how can one investigate these questions? And further, how would one prove that one's claims are valid? I describe some explorations that students, colleagues, and I have been conducting over the past 2 years.

Biography

Dr. Hovy completed a Ph.D. in Computer Science (Artificial Intelligence) at Yale University and was awarded honorary doctorates from the National Distance Education University (UNED) in Madrid in 2013 and the University of Antwerp in 2015. He is one of the initial 17 Fellows of the Association for Computational Linguistics (ACL) and is also a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI). Dr. Hovy's research focuses on computational semantics of language and addresses various areas in Natural Language Processing, Machine Learning, and Data Analytics, including in-depth machine reading of text, information extraction, automated text summarization, question answering, the semi-automated construction of large lexicons and ontologies, and machine translation. In early 2026 his Google h-index was 113, with over 75,000 citations.

For more information:
http://www.iis.sinica.edu.tw/

2026 DISTINGUISHED LECTURE SERIES