Language Learning and Understanding for AI Scientists and Linguists

Does AI See Like We Speak? The Egocentric–Allocentric Axis in Vector Spaces

Yasuhiro Kondo

Aoyama Gakuin University
yhkondo@cl.aoyama.ac.jp

Abstract

The distinction between egocentric and allocentric perspectives — whether a scene or event is construed from the speaker's own viewpoint or from an external, observer-independent frame of reference — is a fundamental axis in human cognition, and one that is grammatically encoded with particular richness in Japanese (for example, in directional verbs such as iku / kuru, in spontaneous forms such as mieru and kikoeru, and in benefactive and deictic systems more broadly). This talk asks whether the same axis is also latent in the internal representations of contemporary AI systems, and whether it takes a comparable form across modalities.

We first show that, in text-based large language models, a clear egocentric–allocentric dimension can be extracted from contextual embeddings of perspectival vocabulary. Using multilingual-e5 and the hidden states of decoder-only models such as Llama 3, principal component analysis on a curated set of directional and viewpoint-related Japanese expressions yields a leading component that cleanly separates egocentric from allocentric items. The resulting axis behaves as a genuine semantic dimension: it generalizes beyond the seed vocabulary, aligns with linguistic intuitions about spontaneous and deictic forms, and can be used as a steering direction in the model's representation space.

We then show that an analogous structure emerges in vision foundation models. Applying the same analytical procedure to image embeddings from CLIP and DINO/DINOv2 — using image sets contrasting first-person and third-person depictions of comparable scenes — we again recover a low-dimensional axis that separates egocentric from allocentric construals of visual content. The axis derived from language and the axis derived from vision are not identical, but they are structurally parallel and partially aligned, suggesting that the egocentric–allocentric contrast is not a peculiarity of linguistic encoding but a more general organizing principle that self-supervised models recover from their respective input distributions.

Taken together, these findings support a view in which a perspectival axis long studied in linguistics and cognitive science — and grammaticalized with unusual transparency in Japanese — surfaces as a shared geometric feature across text and vision models. We discuss implications for the comparison of human and machine language understanding, for cross-modal grounding, and for the use of grammatically rich languages as probes into the representational structure of foundation models.

Keywords: egocentric and allocentric perspective; large language models; vision foundation

models; Japanese linguistics; representation geometry