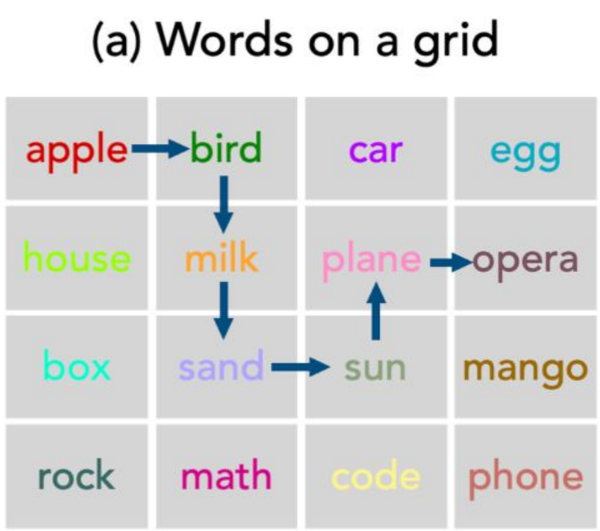
In-Context Learning of Representations

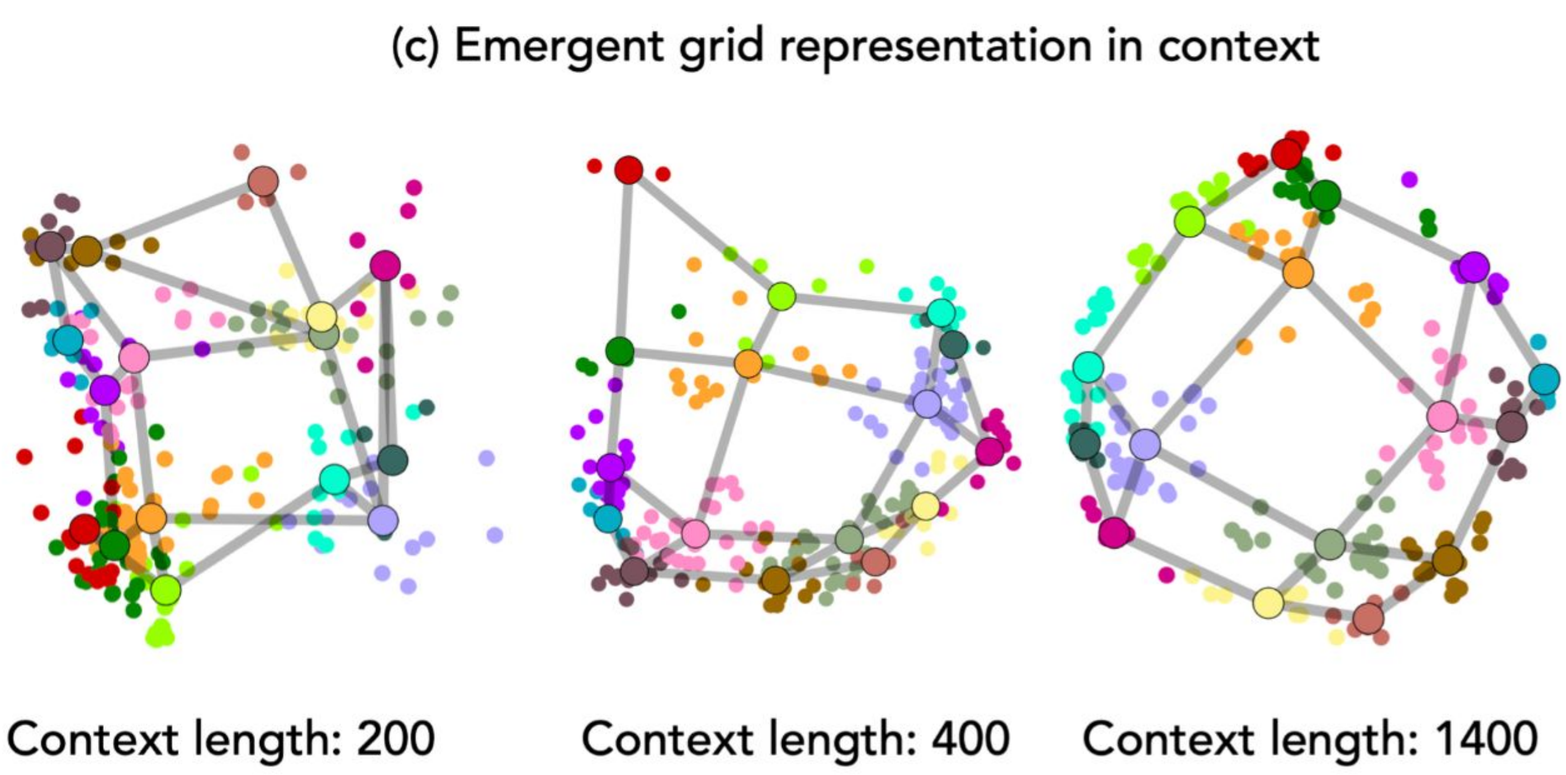
Core Francisco Park*, Andrew Lee*, Ekdeep Singh Lubana*, Yongyi Yang*,
Maya Okawa, Kento Nishi, Martin Wattenberg+, Hidenori Tanaka+

Preprint

TL;DR:

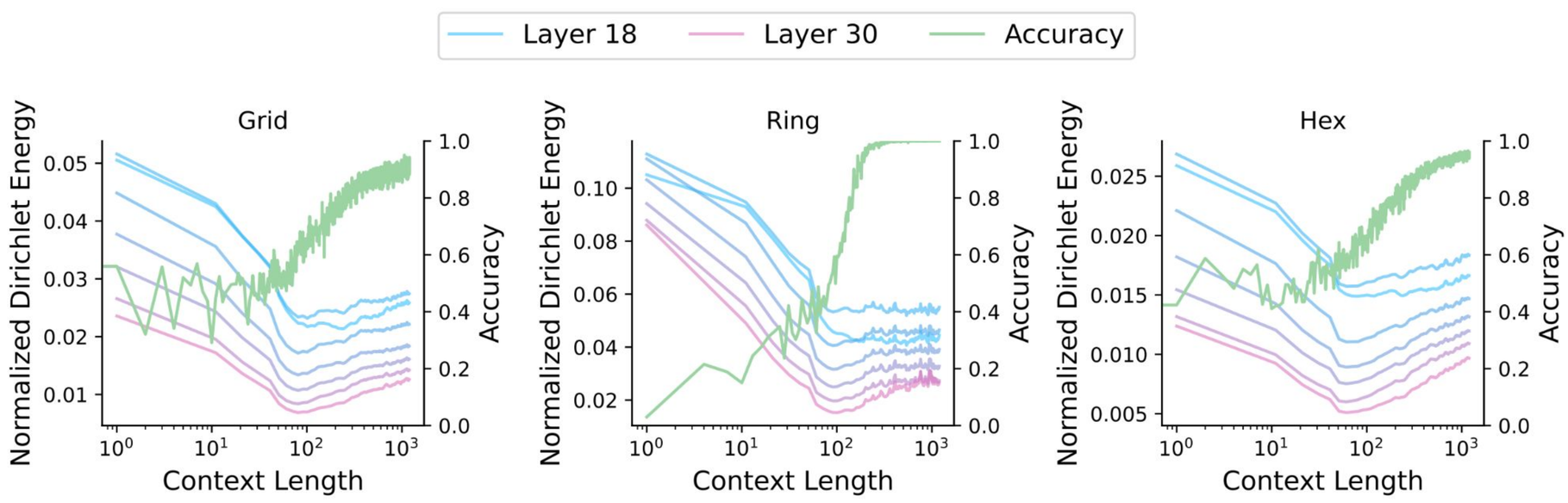
LLMs restructure their internal representations to match a task structure given in context.



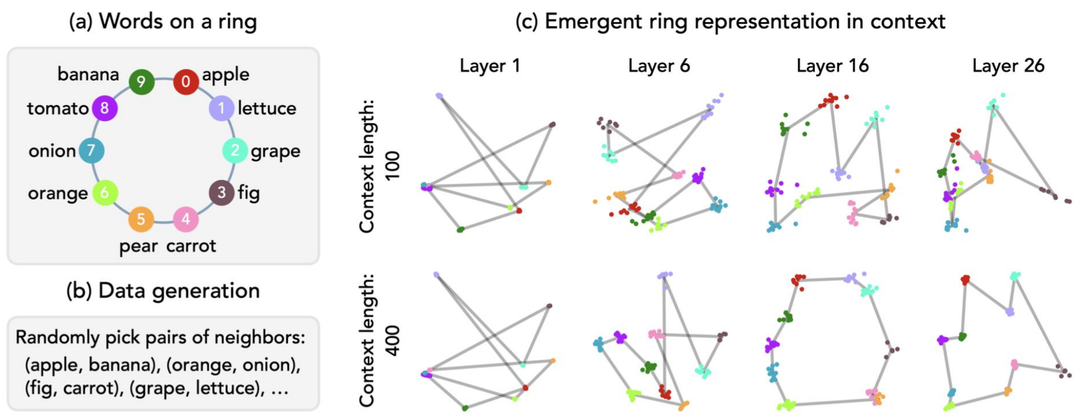


There seems to be a graph spectral energy minimization going on!

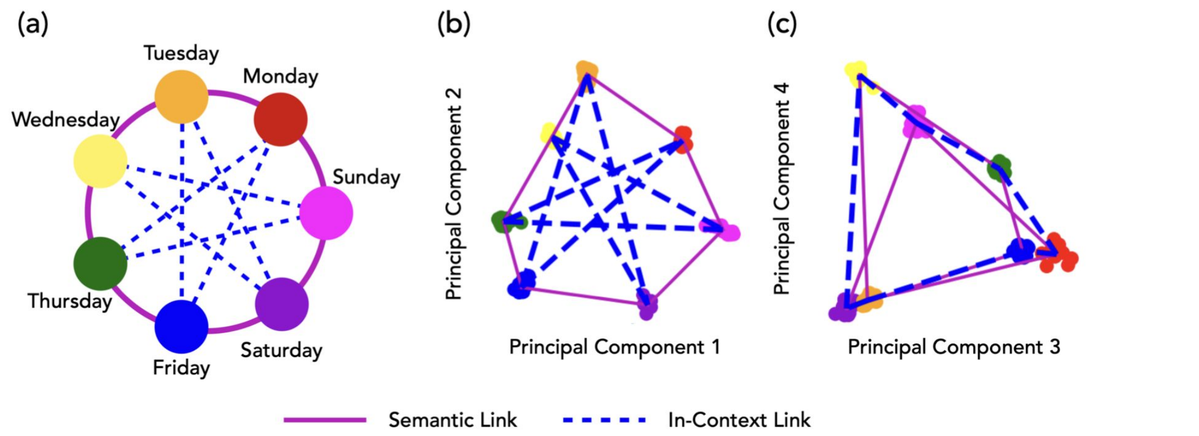
$$E(A, \vec{h}) = \frac{1}{2} \sum_{i,j} A_{ij} || \vec{h_i} - \vec{h_j} ||_2^2$$



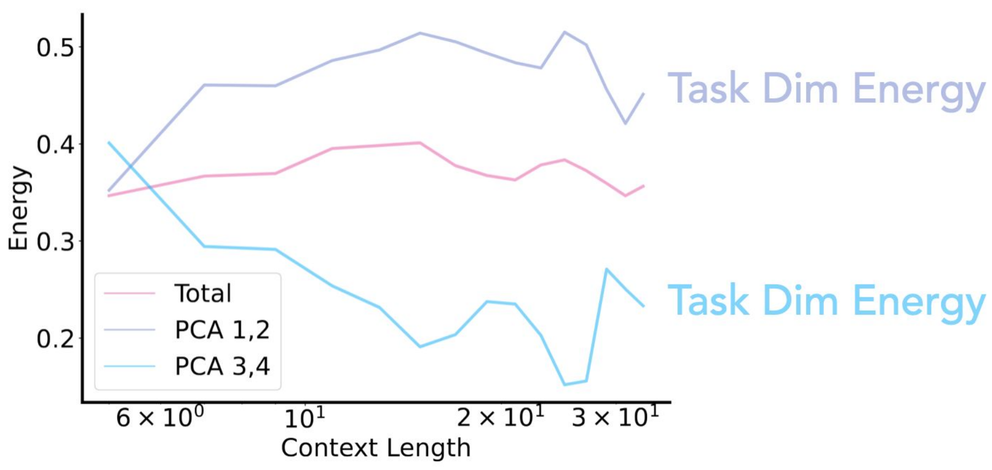
Ring geometry with random edge sampling



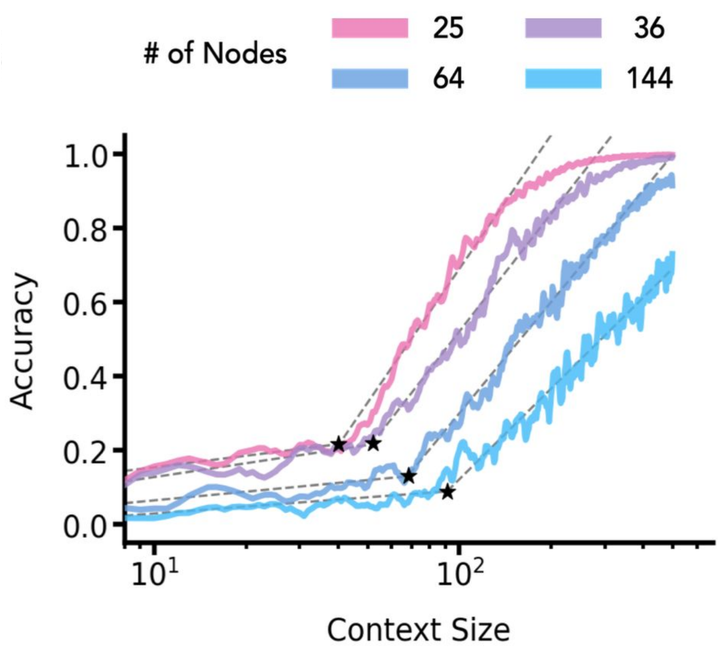
Semantic priors inversion happens at higher PCs



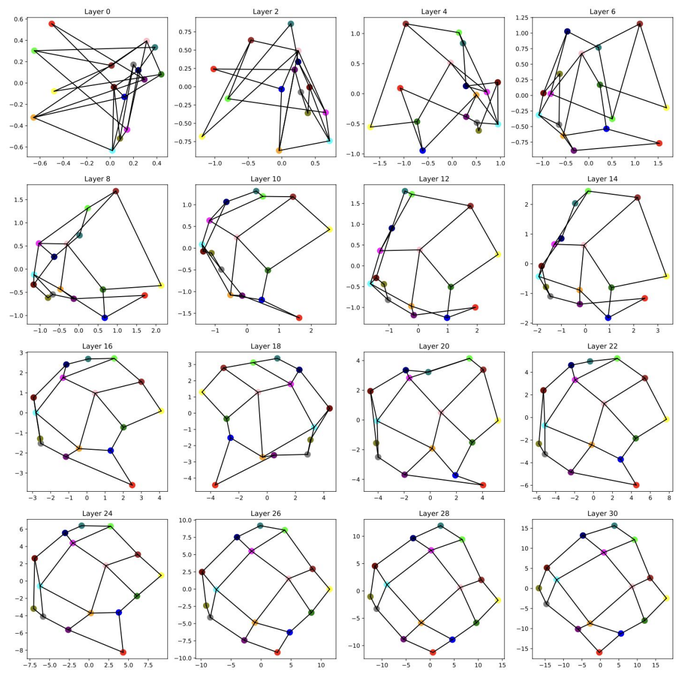
Energy minimization does happen in the "in-context PCs"



There seems to be an in-context transition!



Deeper layers show more robust structure.



Contact: corefranciscopark@g.harvard.edu