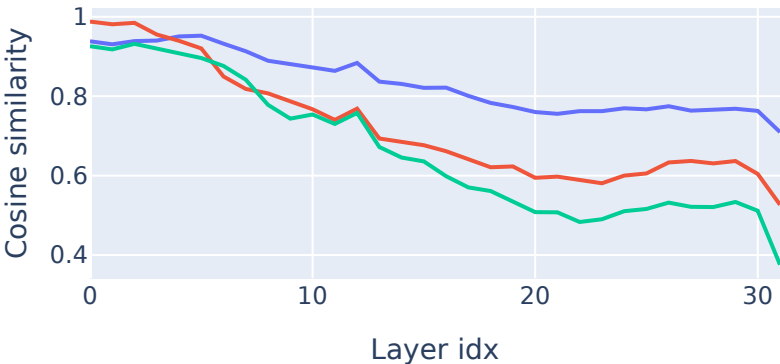Cosine Similarity Between Representations

Cosine similarity vs Layer idx

- Sim. between examples w/ instr
- Sim. between examples w/o instr
- Sim between same ex. w/ and w/o instr