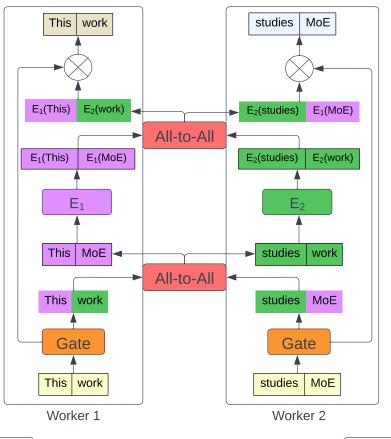
## Distributed MoE Token Flow, where |W| = EP = 2



Gate Scale 2<sup>nd</sup> All-to-All 1st All-to-All Expert E<sub>i</sub> is routes applies reconstitutes disseminates an FFN probability tokens to processed routed block weights experts tokens tokens