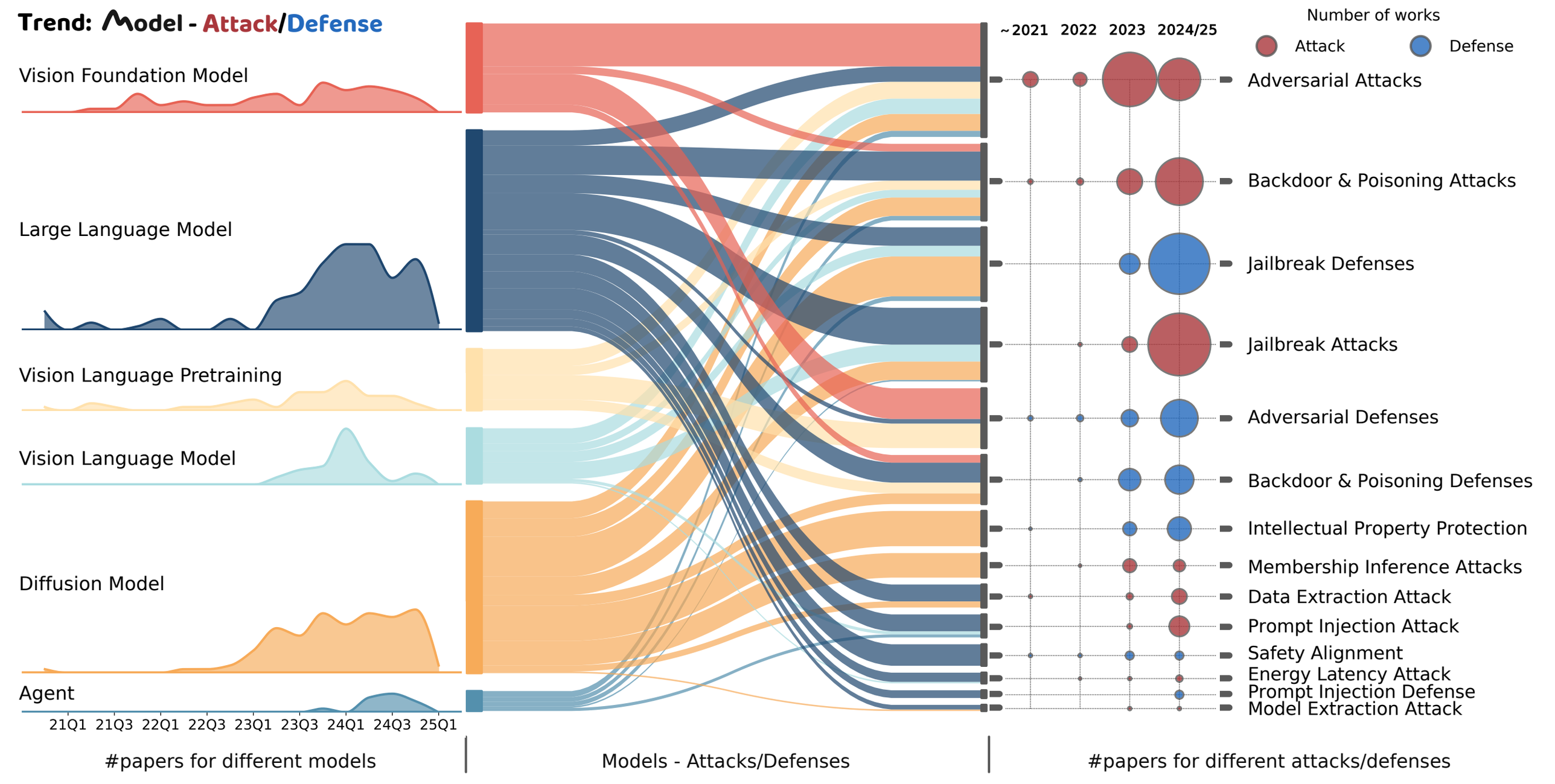
Trend: Model - Attack/Defense
Vision Foundation Model
Large Language Model
Vision Language Pretraining
Vision Language Model
Diffusion Model
Agent
21Q1 21Q3 22Q1 22Q3 23Q1 23Q3 24Q1 24Q3 25Q1
~2021 2022 2023 2024/25
Number of works
Attack
Defense
Adversarial Attacks
Backdoor & Poisoning Attacks
Jailbreak Defenses
Jailbreak Attacks
Adversarial Defenses
Backdoor & Poisoning Defenses
Intellectual Property Protection
Membership Inference Attacks
Data Extraction Attack
Prompt Injection Attack
Safety Alignment
Energy Latency Attack
Prompt Injection Defense
Model Extraction Attack
#papers for different models
Models - Attacks/Defenses
#papers for different attacks/defenses