

**Model Efficiency: F1 Score vs Model Size**  
(Marker size  $\propto 1/\text{Latency}$ )

