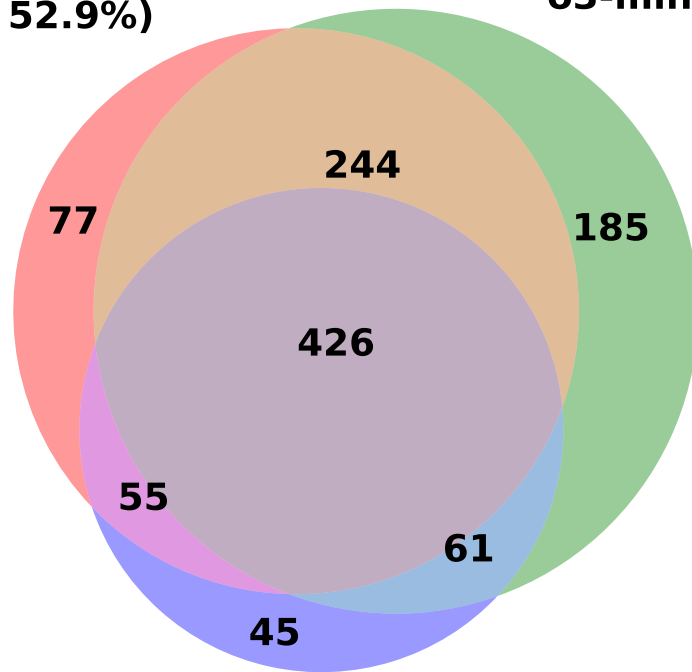


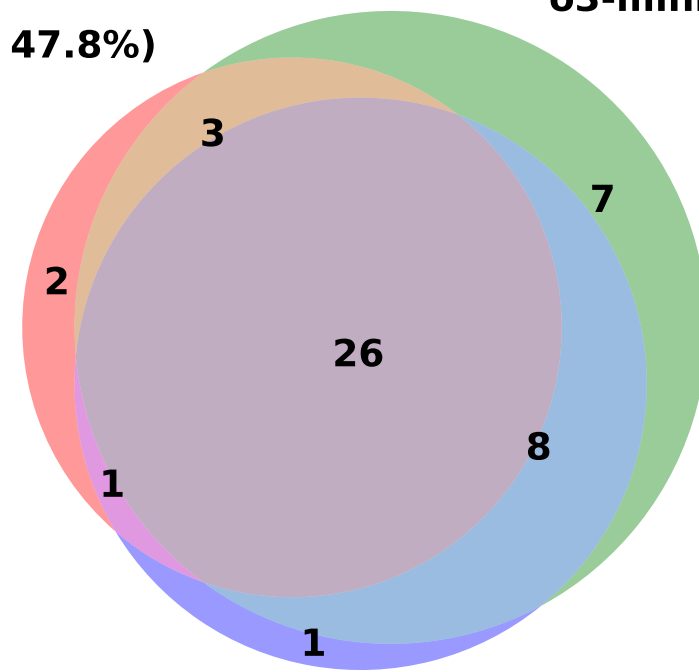
Claude 3 Sonnet (802/1515, 52.9%)



Deepseek (587/1515, 38.7%)

o3-mini (916/1515, 60.5%)

Claude 3 Sonnet (32/67, 47.8%)



Deepseek (36/67, 53.7%)

o3-mini (44/67, 65.7%)