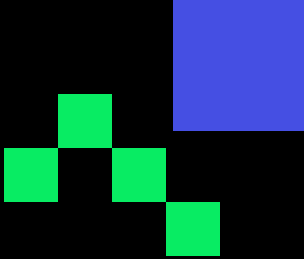


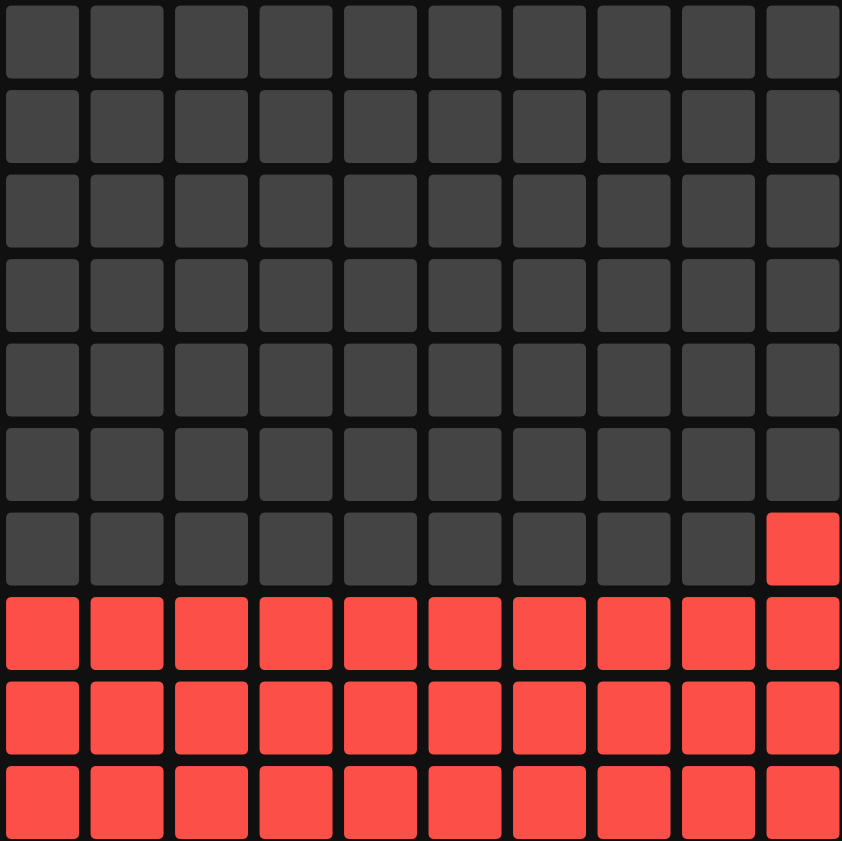
Rogue Deep Think (Preview)

DeepSWE v1.1



Fewer failures

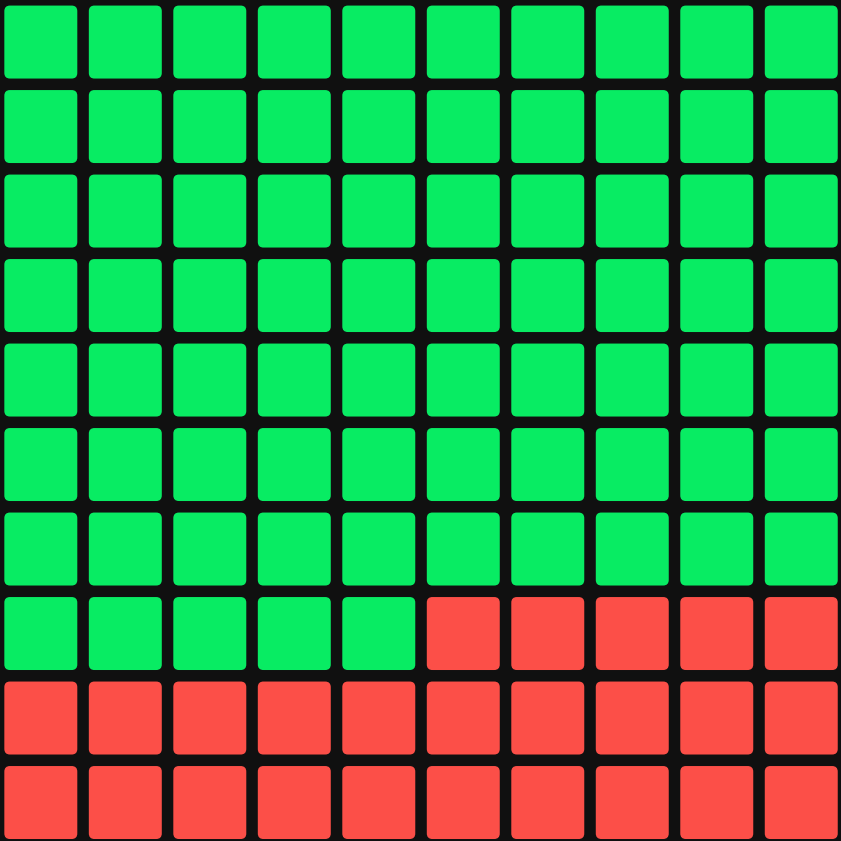
Failed attempts out of 100: 30.97 to 24.78, 20% fewer



GLM 5.3 internal baseline

30.97%

failed



Rogue Deep Think (Preview)

24.78%

failed