Compiled success rate by parameter count across LLMs (baseline filtered) Human (original) o3-mini 88% deepseek 87% claude3 sonnet 83% 82% 80% 80% 80% Union of 3 LLMs 77% 76% 69% 68% 68% 62% 62% 60% 57% 55% 52% 48% 46% 42% 39% 39% 34% 30% 27% 27% 22% 20%

100

successfully

of files typechecked

Percentage

