Compiled success rate by parameter count across LLMs (baseline filtered) 100 gpt-3.5 gpt-4o o1-mini o3-mini 80% deepseek 80 claude3 sonnet 76% Union(o3-mini, deepseek, claude) 68% 68% Intersection(o3-mini, deepseek, claude) 67% typechecked Union(o3-mini, o1-mini, gpt-4o) 62% 61%0% 60% 57% 55% 52% 52% 492/18% 49% 47% 46% 45% 42% 39% 39% 38% 34%5% **3**5% 34% 30% 28% 26% 25% Percentage 23% 22% 212/19 20% 18% 17% 16% 15% 14% 13% 12% 12% 5%

successfully

of files

Parameter count bins (from untyped baseline)