Compiled success rate by parameter count across LLMs (baseline filtered)