Figure panels (left to right): **Code generation**, **MATH**, **MMLU**, **MMLU with soft labels**, **Code critique**

Y-axis: Correctness
X-axis: Number of demos (1, 8, 64, 512)

Legend:
- Locked with pwd (green dashed)
- Locked without pwd (red dotted)
- All (high-quality) demos (blue dashed)
- Some (high-quality) demos (blue solid)
- Some pi0 samples (orange solid)