Safety-Utility Pareto Frontier

Legend:
- No Defense
- Escape-Hatch
- Input Guard
- Reject-Only
- DeLex-JSON (Ours)
- Pareto Frontier

Same ASR, higher cost

Ideal (0, 0)

3.3pp lower benign cost

X-axis: Benign Cost (Modification/Rejection Rate, %)
Y-axis: Attack Success Rate (HarmBench, %)