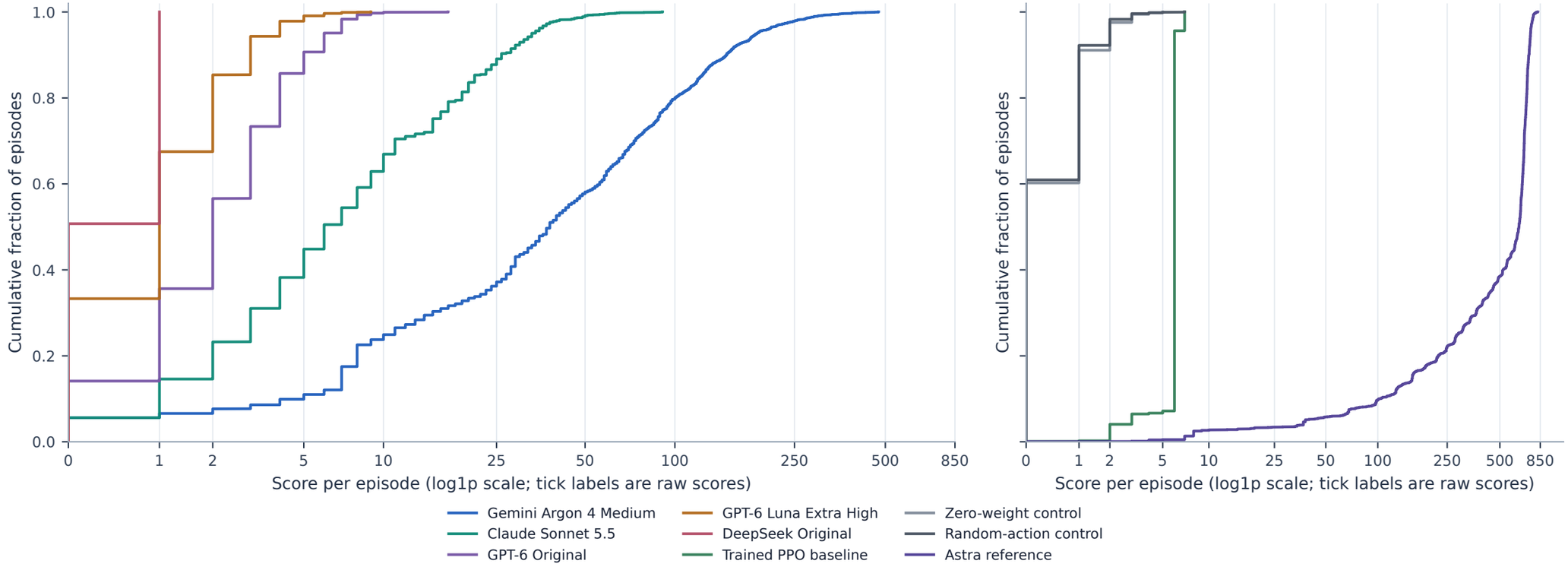
Episode-level score distributions

Five primary zero-shot policies

Trained baseline, controls, and reference



Source: corrected Phase 2 raw episode CSVs · n=3,000 per condition. Each line is an ECDF; larger scores lie farther right. A labeled log1p scale keeps low scores legible while retaining Astra's full range. DeepSeek Original appears only in the primary-model panel.