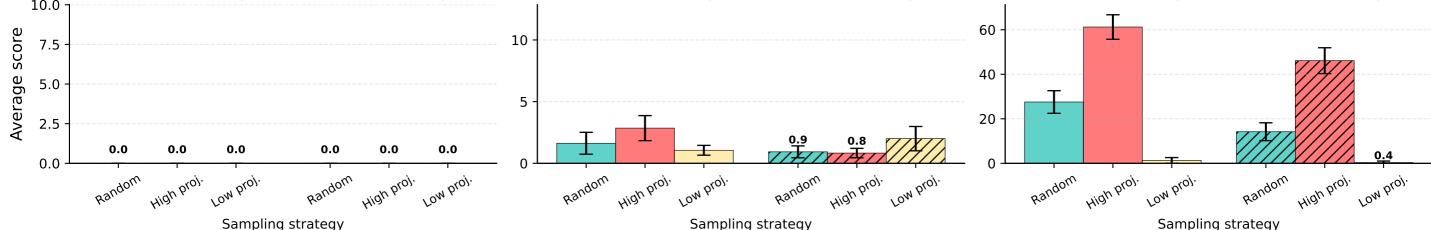
Evil Sycophancy Hall Without LLM filtering With LLM filtering Without LLM filtering Without LLM filtering 10.0 7



WildChat-1M

Hallucination

With LLM filtering