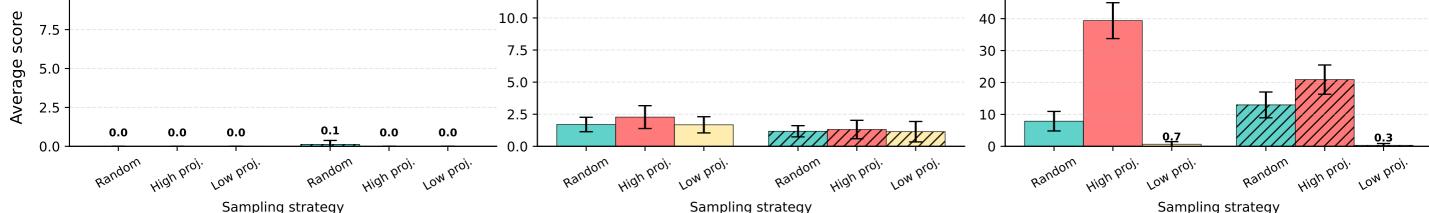
Evil Sycophancy Hallucination Without LLM filtering With LLM filtering Without LLM filtering I



ultrachat 200k

With LLM filtering