Behavioral Consistency (CV)

Figure: Violin plot of CV (%) for three models — Claude 4.5 Sonnet (μ=15.2%), GPT-5 (Snowflake) (μ=32.2%), and Llama-3.1 70B (μ=47.0%).