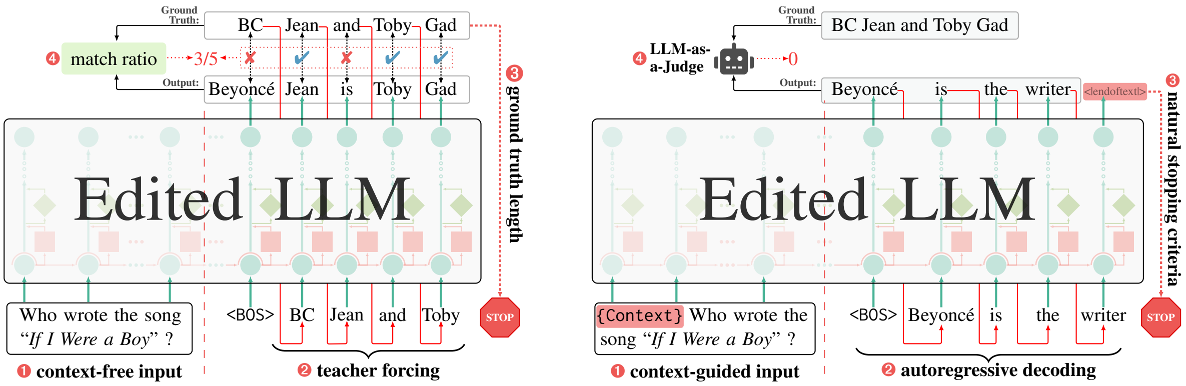

(a) synthetic evaluation framework

(b) WILD evaluation framework

Figure 1: Illustration of synthetic and WILD evaluation frameworks for measuring reliability, generalization, and locality. Each framework comprises four key modules:❶ *input*, ❷ *generation strategy*, ❸ *output truncation*, and ❹ *metric*. Here, we use LLM-as-a-Judge as an example metric to illustrate WILD, which supports various metrics.