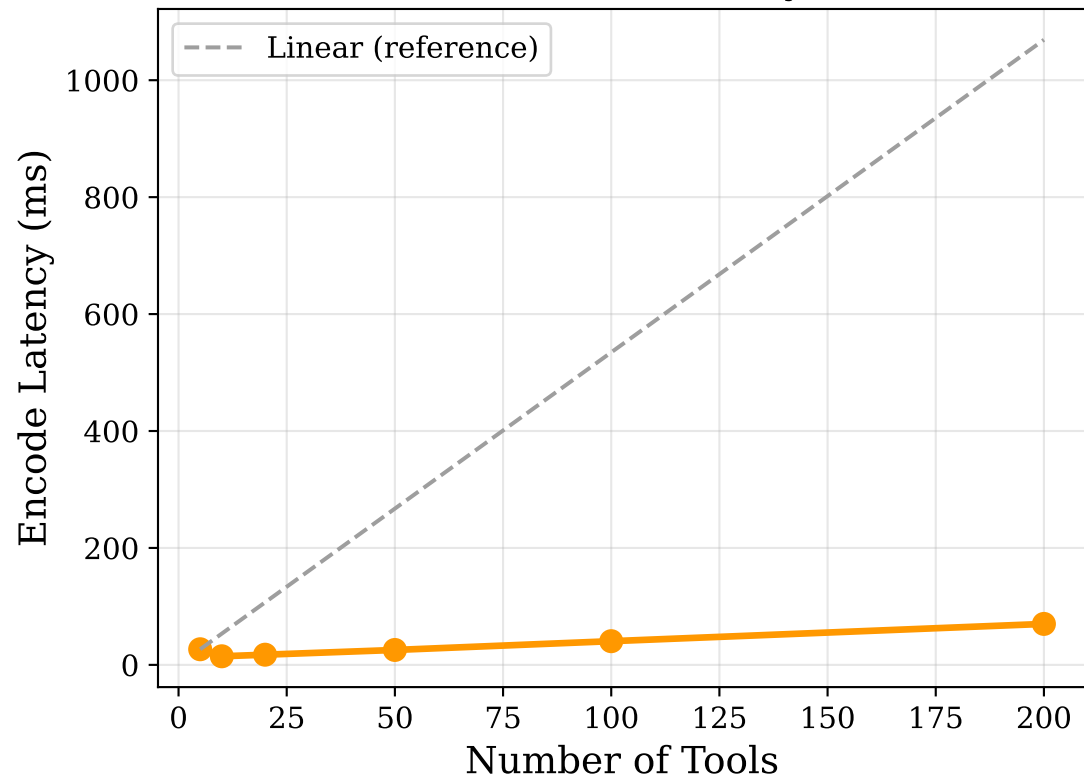ToolFormerMicro: Latency Analysis

(a) Encode Latency

(b) Generate Latency