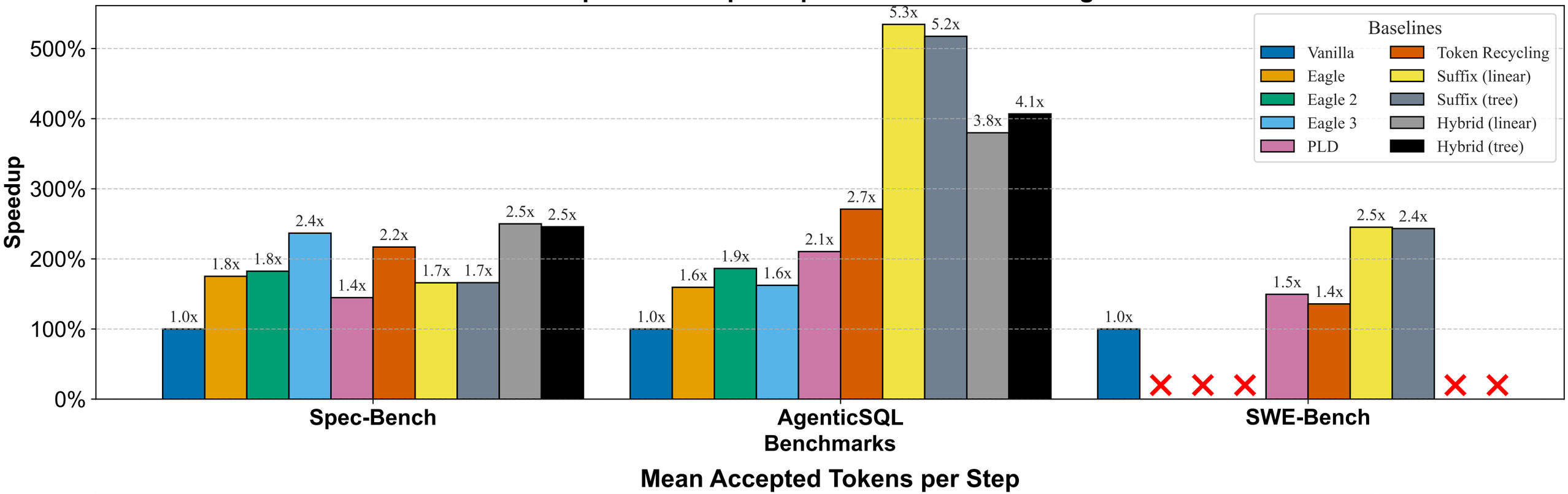
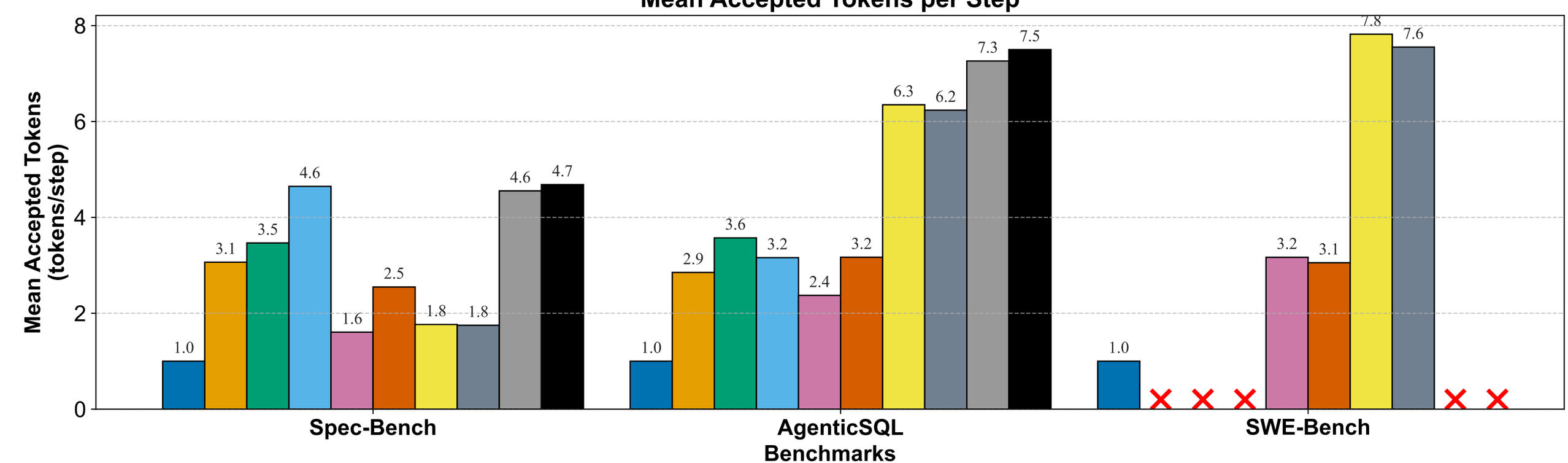
Speculative Speedups over Vanilla Decoding



Mean Accepted Tokens per Step