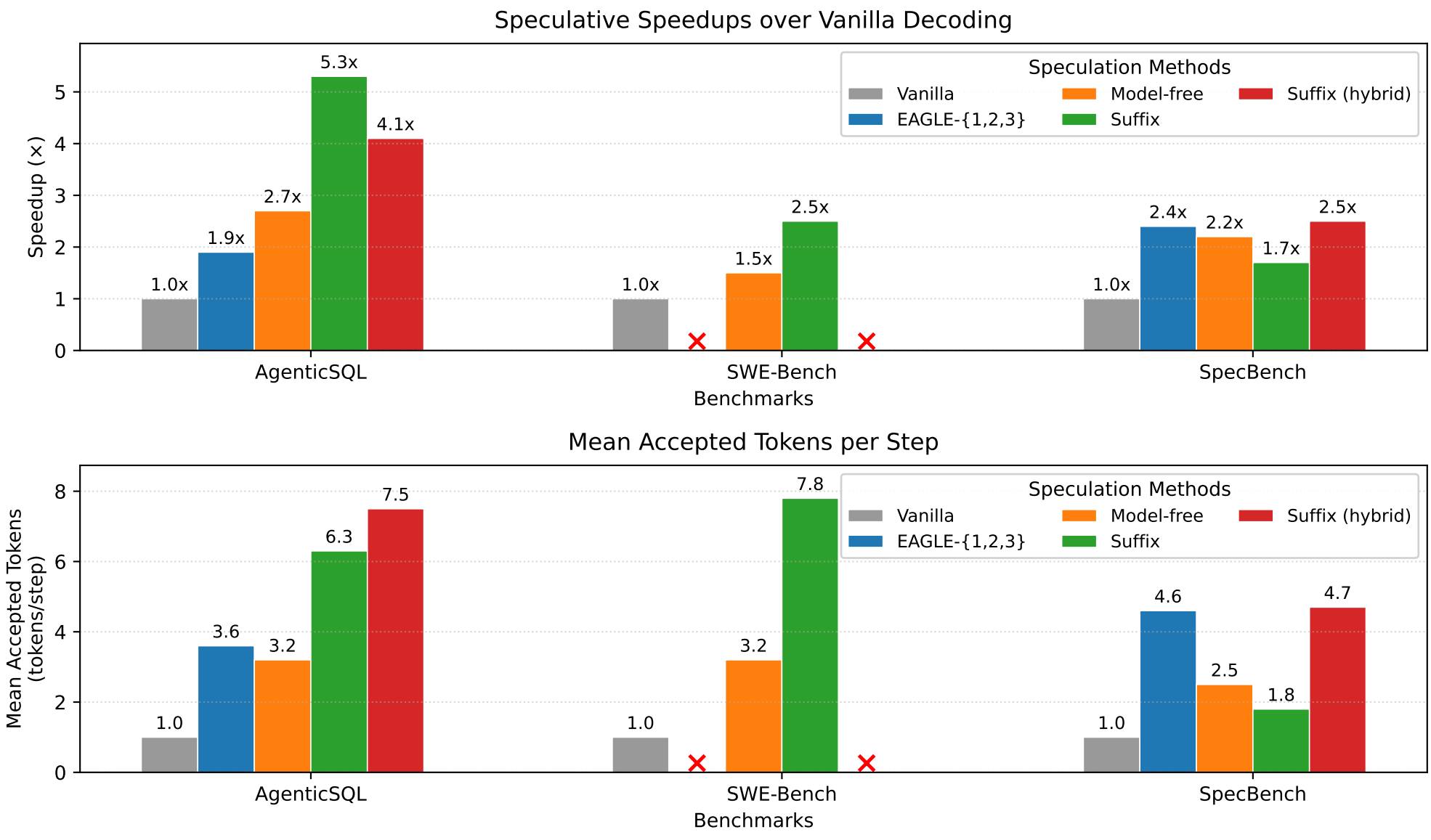
Speculative Speedups over Vanilla Decoding
Speedup (×)
Speculation Methods
Vanilla
EAGLE-{1,2,3}
Model-free
Suffix
Suffix (hybrid)
1.0x
1.9x
2.7x
5.3x
4.1x
1.0x
1.5x
2.5x
1.0x
2.4x
2.2x
1.7x
2.5x
AgenticSQL
SWE-Bench
SpecBench
Benchmarks
Mean Accepted Tokens per Step
Mean Accepted Tokens (tokens/step)
1.0
3.6
3.2
6.3
7.5
1.0
3.2
7.8
1.0
4.6
2.5
1.8
4.7