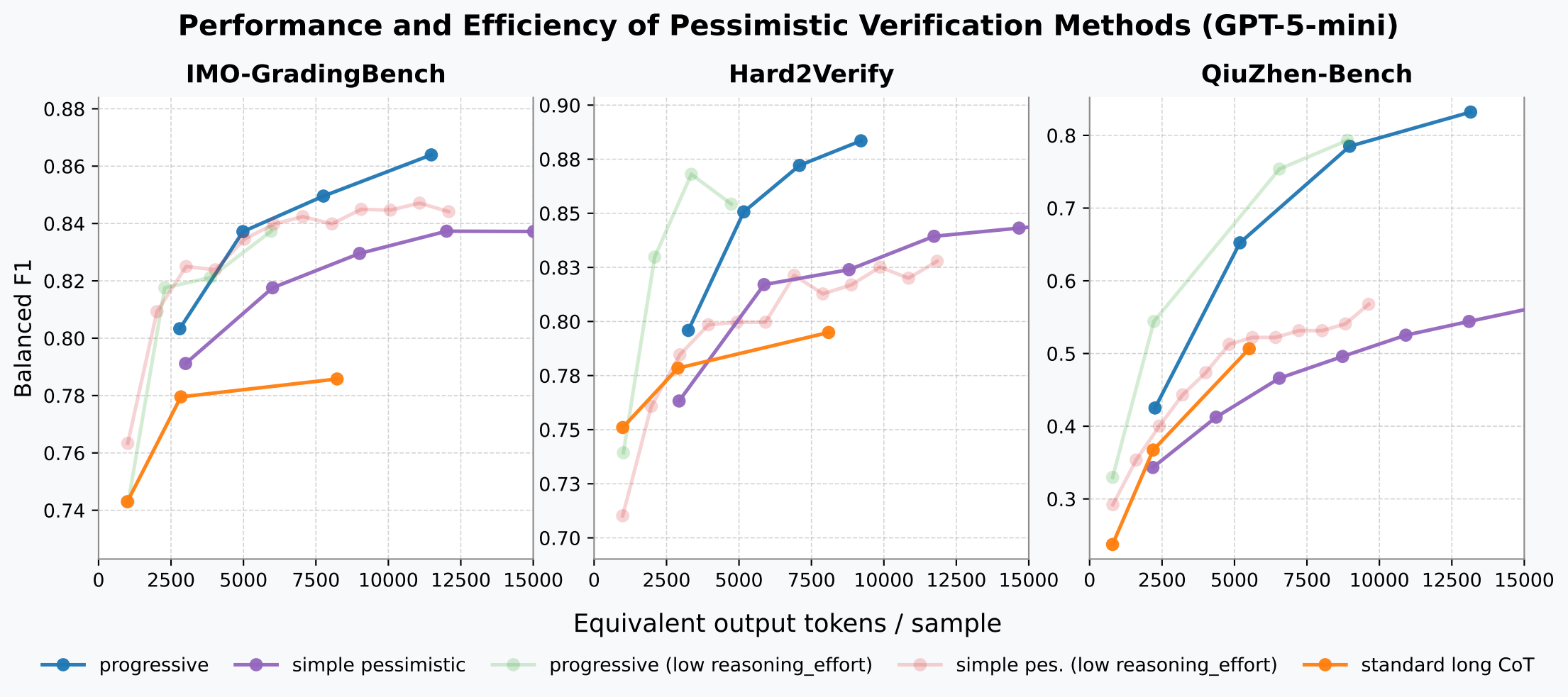
Performance and Efficiency of Pessimistic Verification Methods (GPT-5-mini)
IMO-GradingBench
Hard2Verify
QiuZhen-Bench
Balanced F1
Equivalent output tokens / sample
progressive
simple pessimistic
progressive (low reasoning_effort)
simple pes. (low reasoning_effort)
standard long CoT