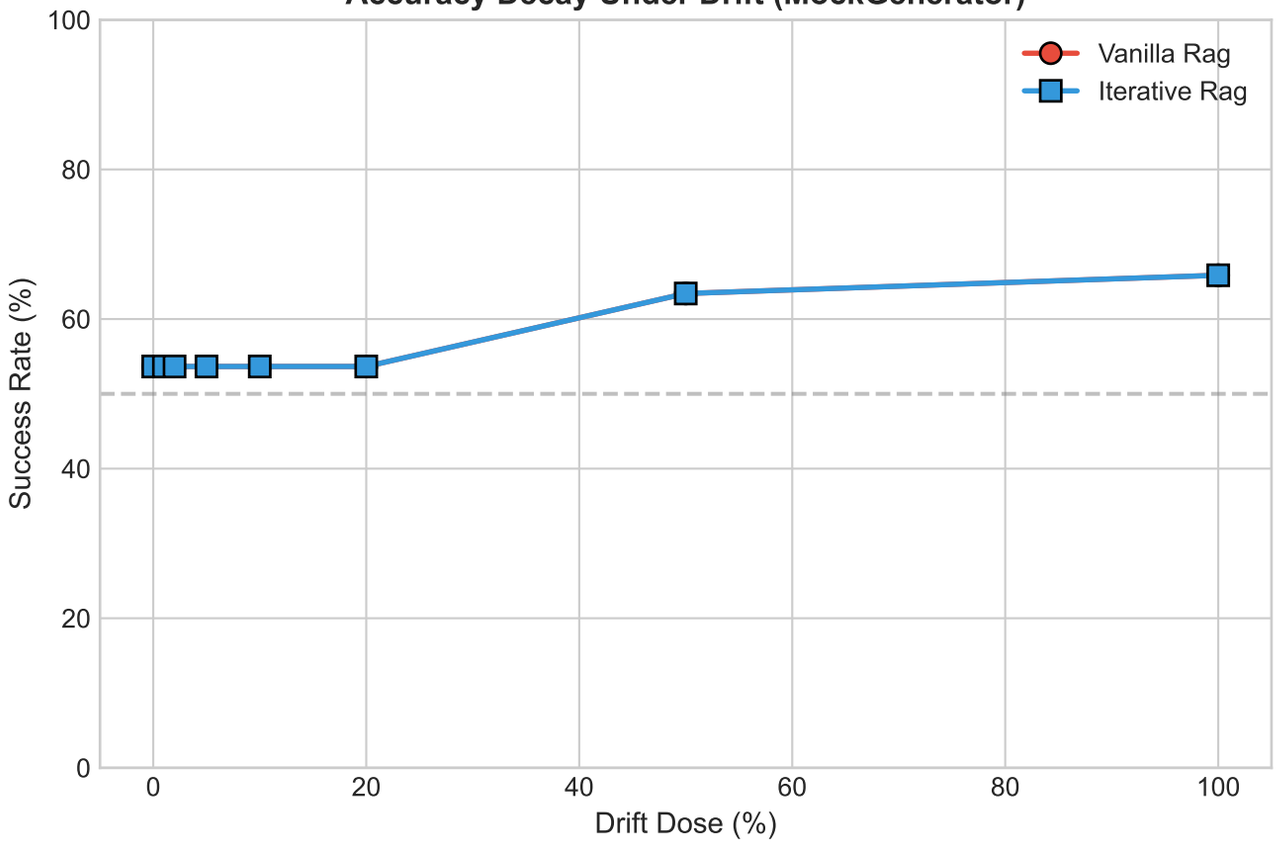

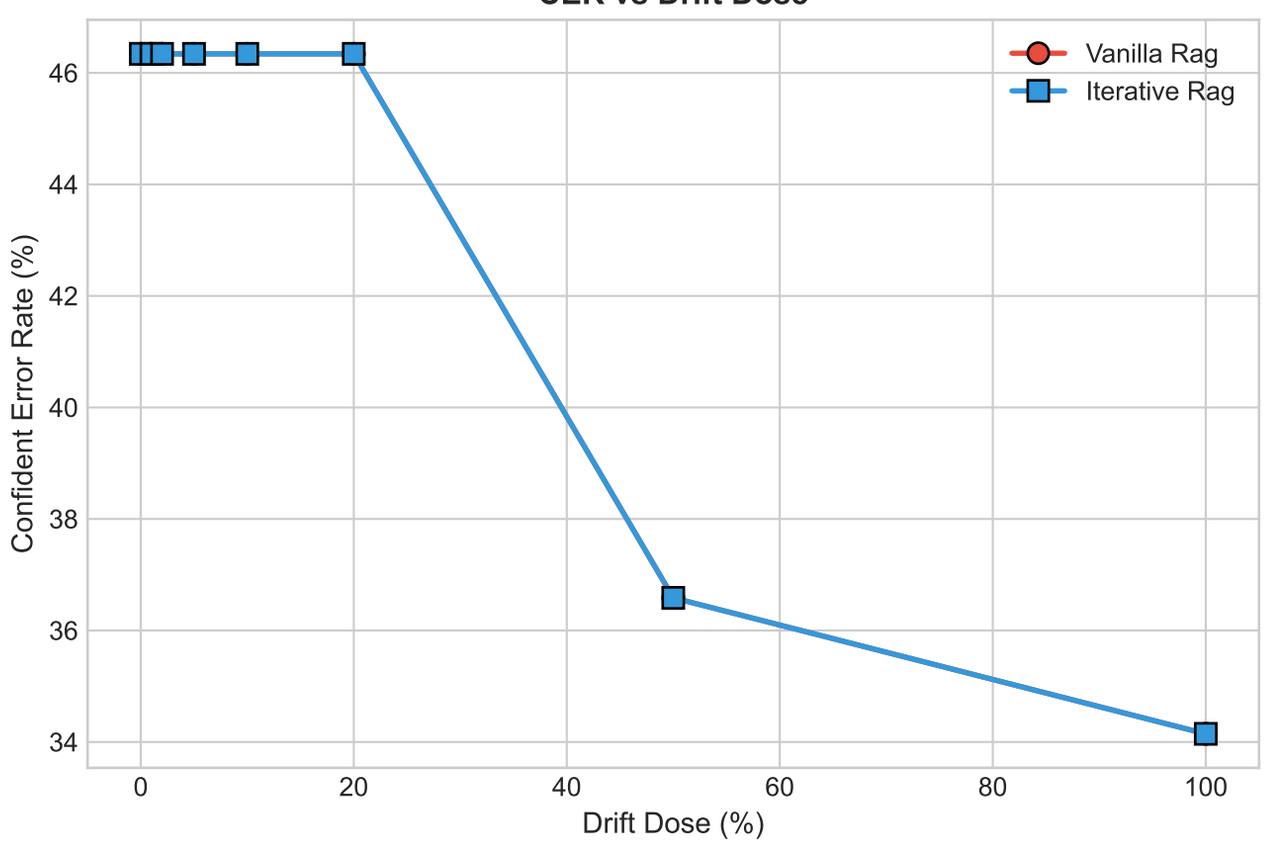

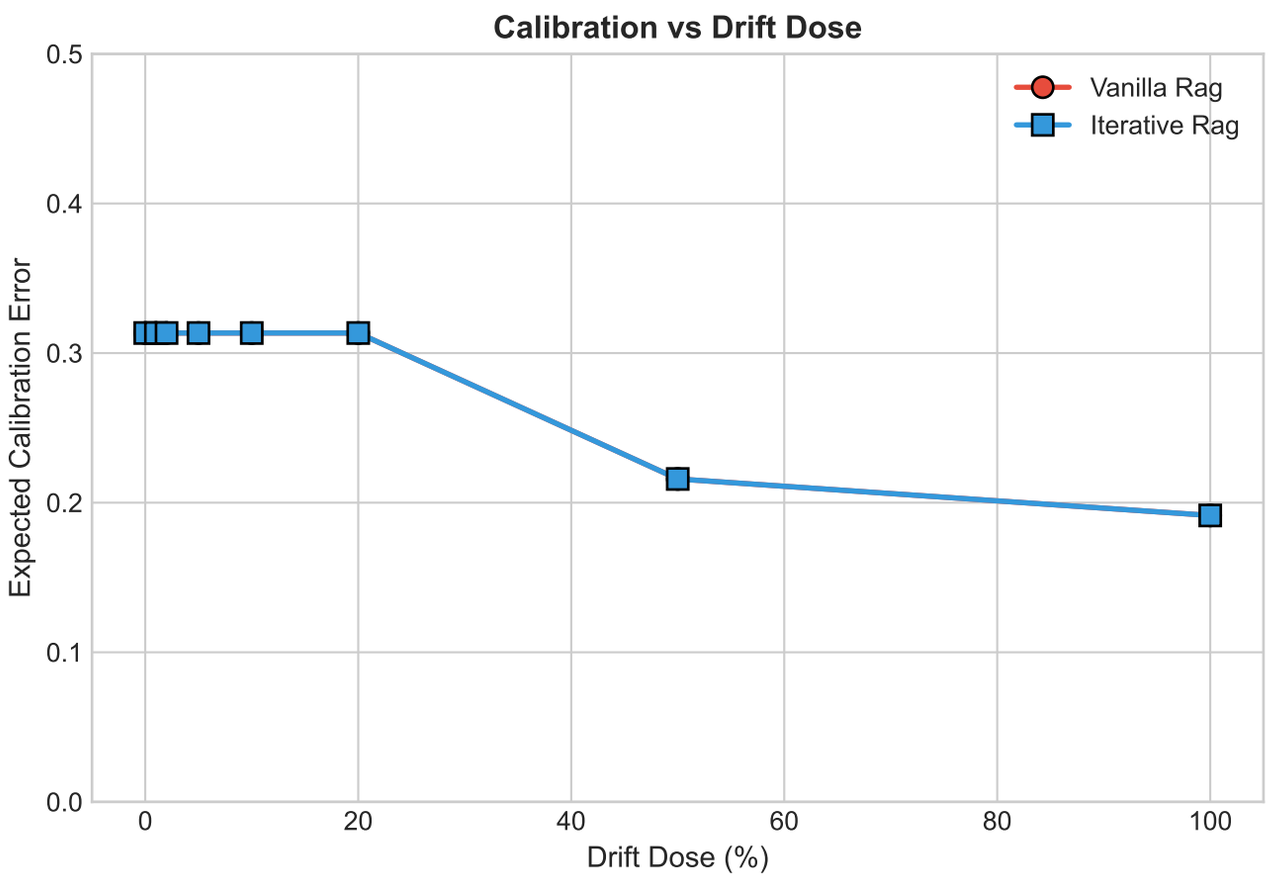


```
DRIFT SWEEP SUMMARY (MockGenerator)
═══════════════════════════════════

Tasks: 41
Drift doses: 0%, 1%, 2%, 5%, 10%, 20%, 50%, 100%

KEY FINDINGS:

• Accuracy is flat from 0-20% drift (53.7%)
• Jumps to 63-66% at 50-100% drift
  (MockGenerator artifact - v2 has cleaner patterns)

• CER decreases with drift (46% → 34%)
  (More correct answers → fewer confident errors)

• ECE improves with drift (0.31 → 0.19)
  (Better calibration as accuracy improves)

• SFR = 0% at all doses
  (MockGenerator appropriately flags uncertainty)

NOTE: Real LLM results show opposite pattern
(accuracy DROPS with drift - see llm_decay_curve.png)
```