Mechanistic Validation: Profile-Based Value Model (M3)
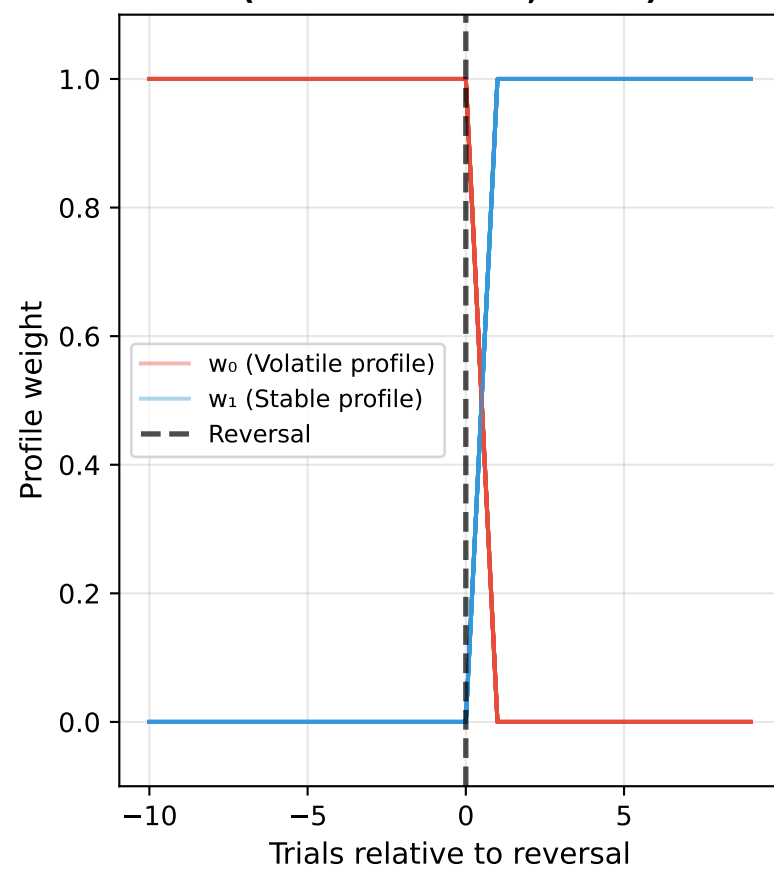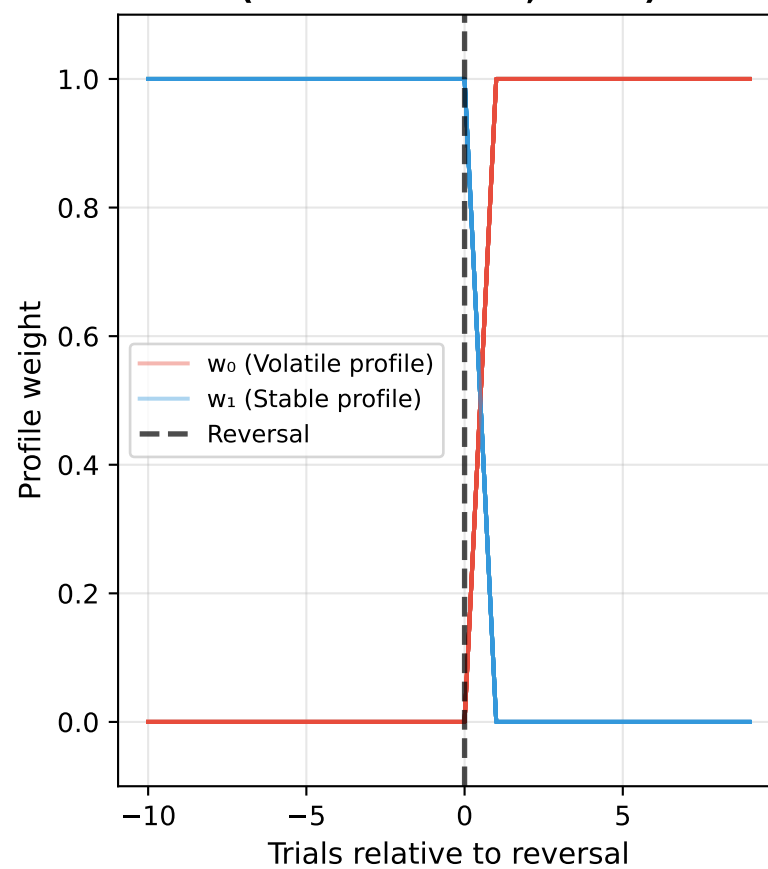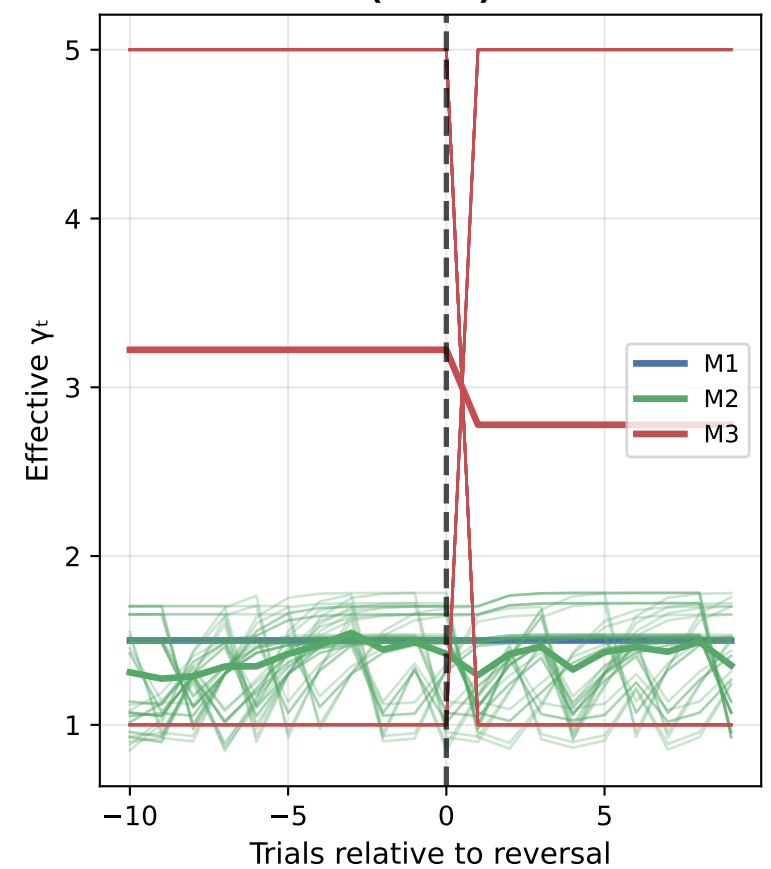

A. Profile Recruitment: Volatile → Stable (n=20 reversals, run 0)
- w_0 (Volatile profile)
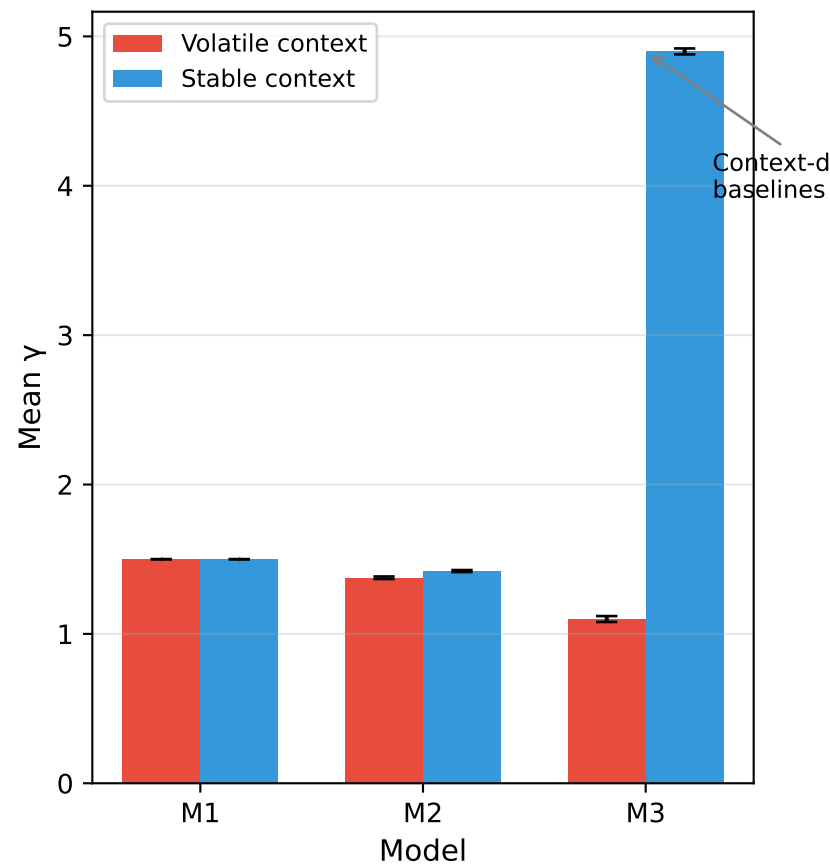- w_1 (Stable profile)
- Reversal

Profile weight vs Trials relative to reversal

B. Profile Recruitment: Stable → Volatile (n=25 reversals, run 0)
- w_0 (Volatile profile)
- w_1 (Stable profile)
- Reversal

Profile weight vs Trials relative to reversal

C. Precision Dynamics Around Reversals (run 0)
- M1
- M2
- M3

Effective γ_t vs Trials relative to reversal

D. Context-Conditional Precision (run 0)
- Volatile context
- Stable context

Mean γ vs Model (M1, M2, M3)

Context-dependent baselines

E. Context-Conditional Hint-Seeking (run 0)
- Volatile context
- Stable context

Hint-seeking rate vs Model (M1, M2, M3)

F. Profile Stability During Micro-Reversals (run 0)
- w_0 (Volatile profile)
- w_1 (Stable profile)

Profile weight vs Trial (within volatile context)

Profile tracks context, not all switches