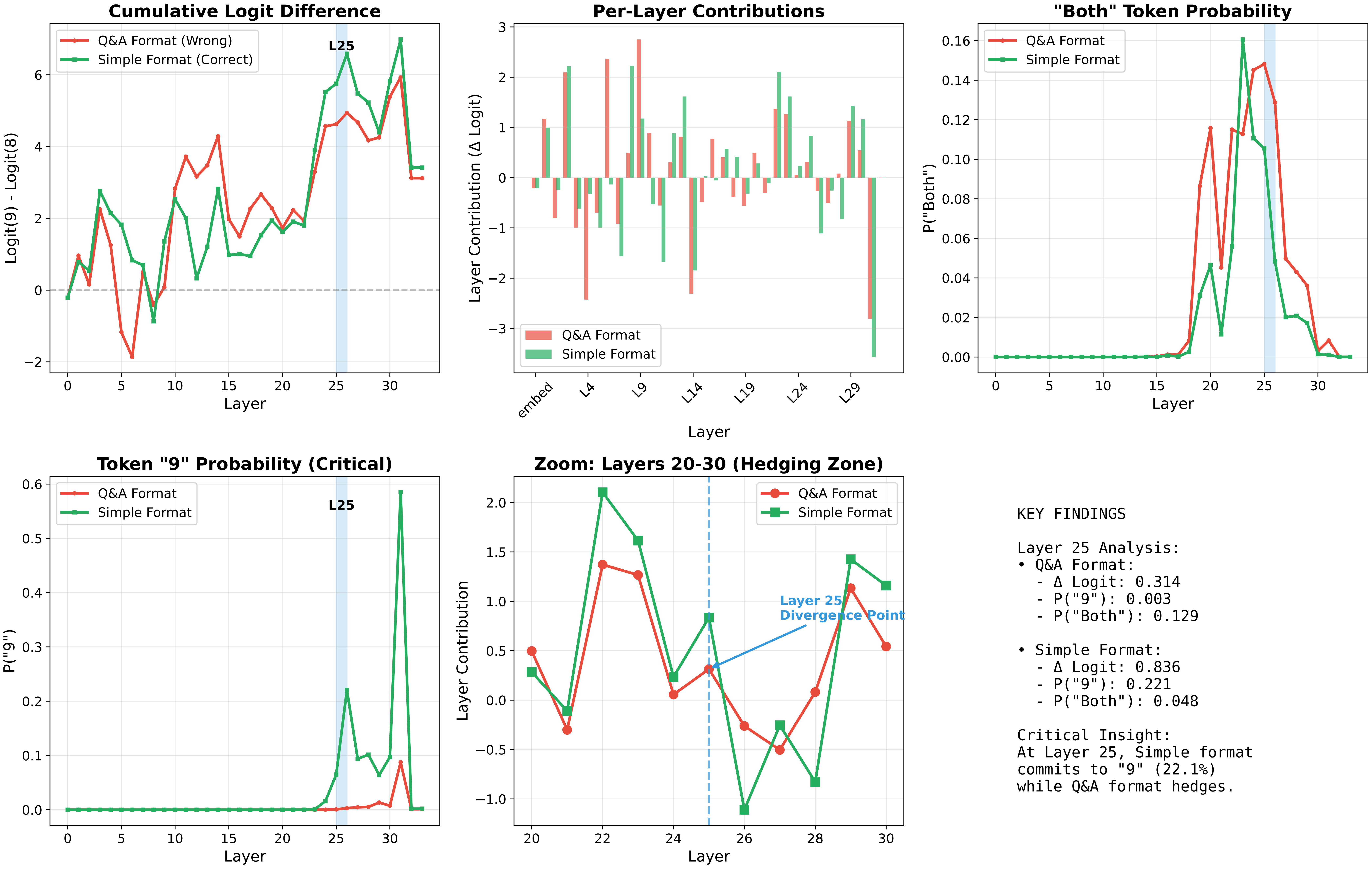
Logit Attribution Analysis: Mechanistic Path of the Decimal Bug



Cumulative Logit Difference

Per-Layer Contributions

"Both" Token Probability

Token "9" Probability (Critical)

Zoom: Layers 20-30 (Hedging Zone)

KEY FINDINGS

Layer 25 Analysis:

- Q&A Format:
  - Δ Logit: 0.314
  - P("9"): 0.003
  - P("Both"): 0.129

- Simple Format:
  - Δ Logit: 0.836
  - P("9"): 0.221
  - P("Both"): 0.048

Critical Insight:
At Layer 25, Simple format
commits to "9" (22.1%)
while Q&A format hedges.