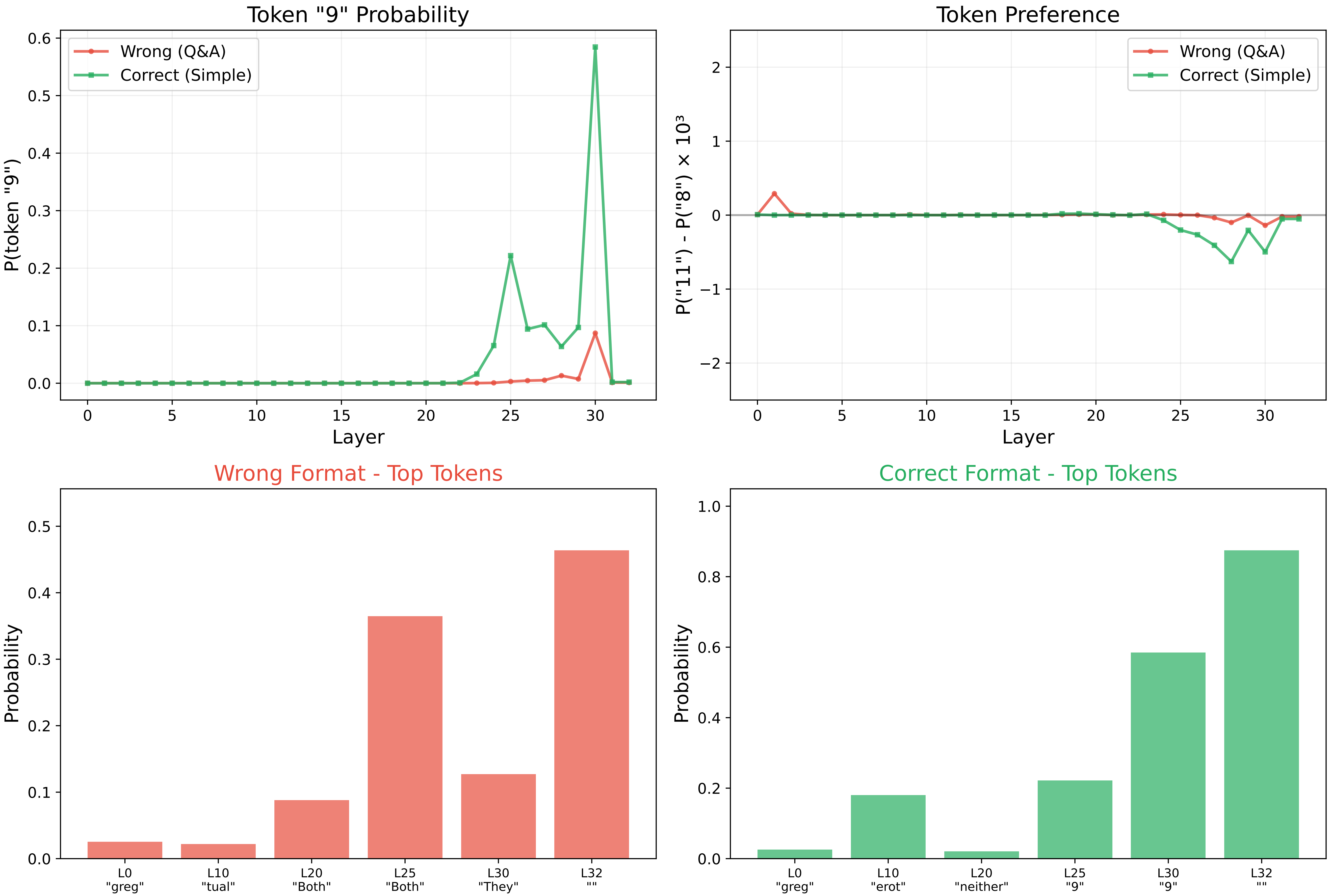
Logit Lens Analysis: Format-Dependent Processing


Token "9" Probability
P(token "9")
Layer
Wrong (Q&A)
Correct (Simple)
Token Preference
P("11") - P("8") × 10³
Wrong (Q&A)
Correct (Simple)
Wrong Format - Top Tokens
Probability
L0
"greg"
L10
"tual"
L20
"Both"
L25
"Both"
L30
"They"
L32
""
Correct Format - Top Tokens
Probability
L0
"greg"
L10
"erot"
L20
"neither"
L25
"9"
L30
"9"
L32
""