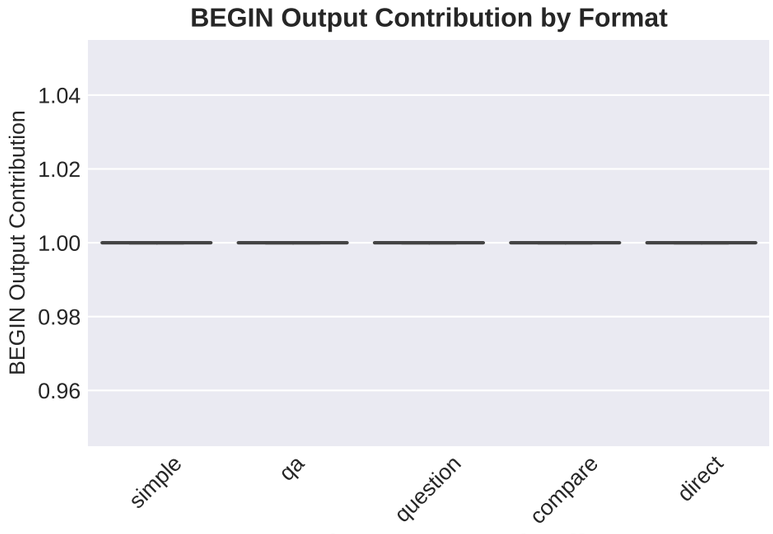
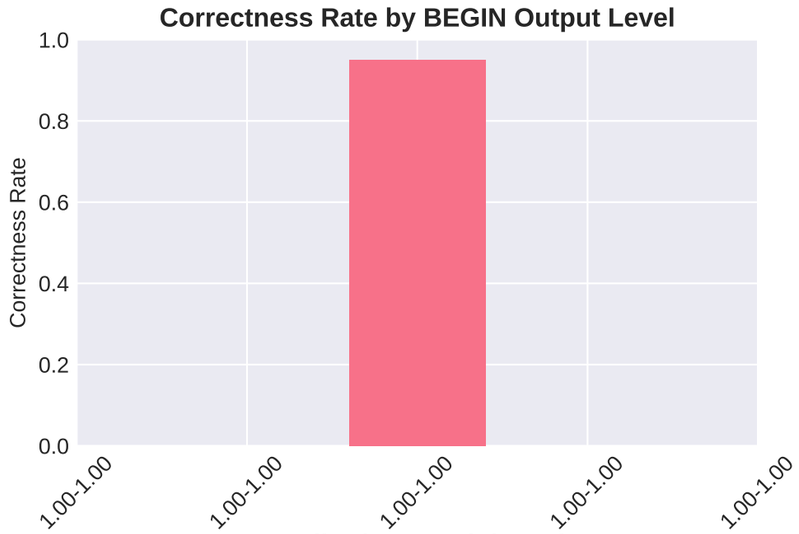
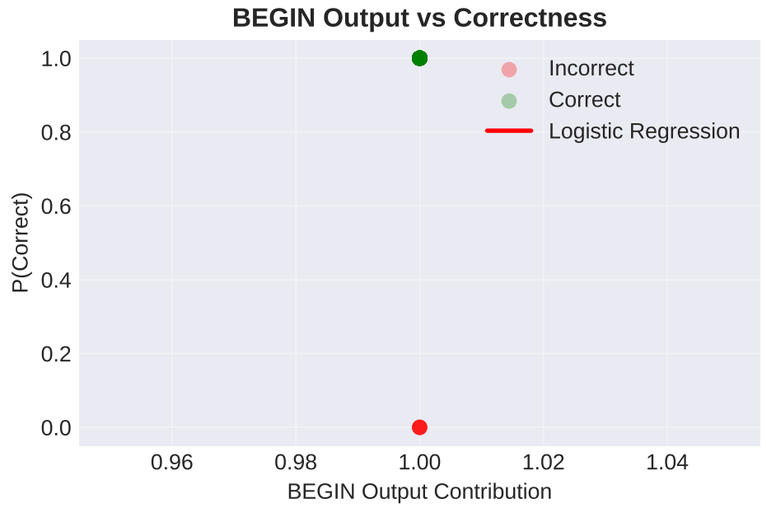
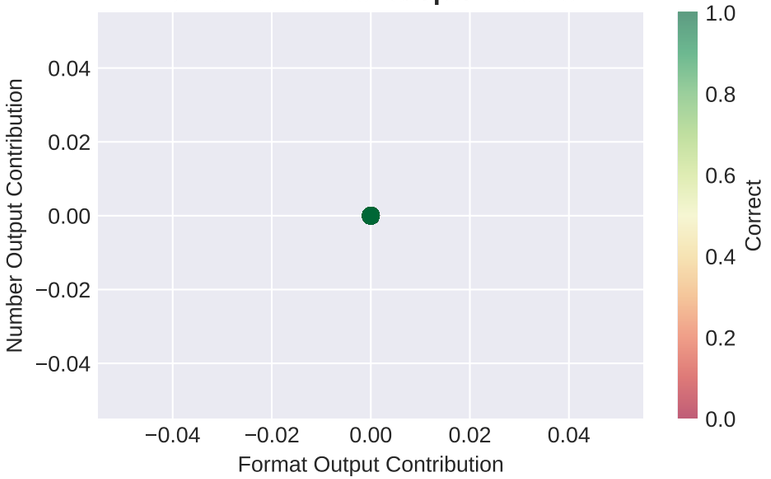
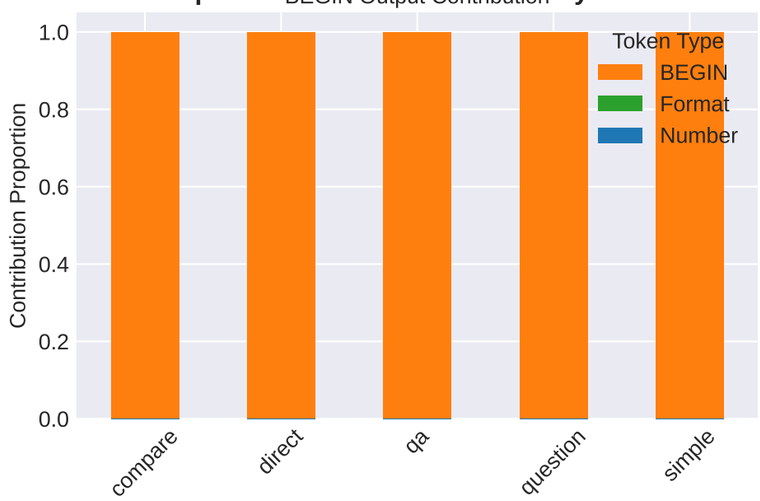
Attention OUTPUT Quantification: Information Flow Analysis at Layer 10

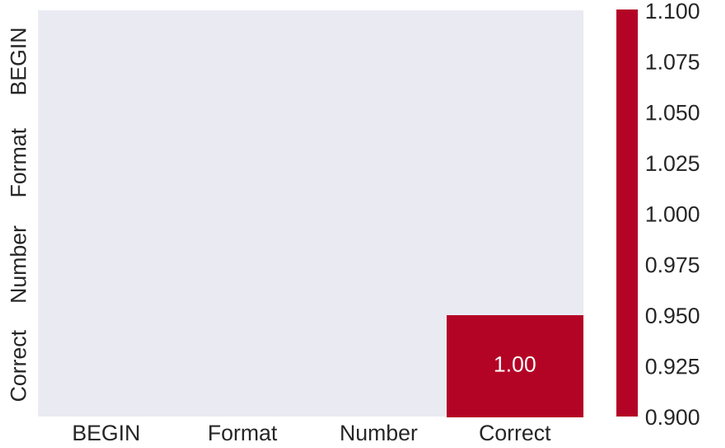


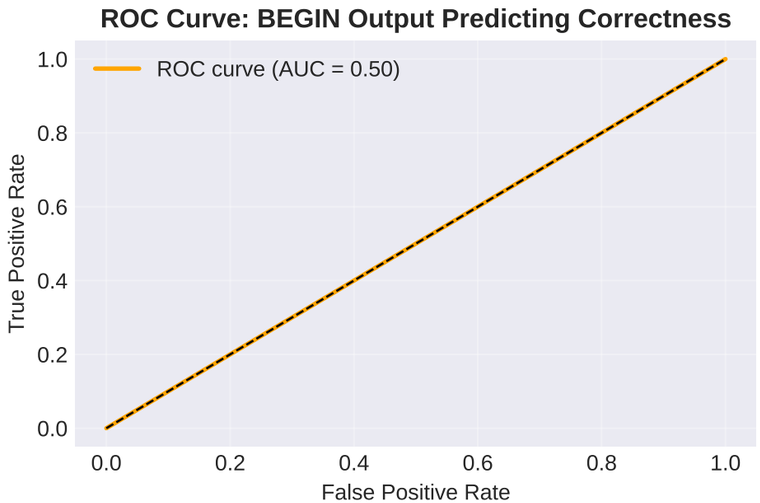


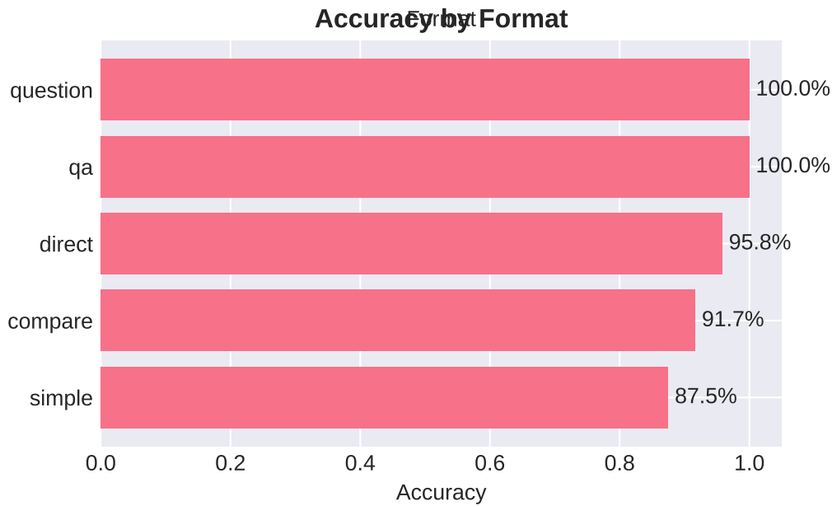













```
KEY FINDINGS (Attention OUTPUT):

• Correlation (BEGIN output, correctness):
  Not available (single class)

• Logistic Regression:
  Coefficient = -0.000
  ROC AUC = 0.500

• Format Comparison:
  Simple format: 100.0% BEGIN output
  Q&A format: 100.0% BEGIN output
  Difference: 0.0%

• Interpretation:
  Attention OUTPUT (information flow)
  shows different patterns than
  attention WEIGHTS (where model looks)

  This measures actual causal mechanism!
```