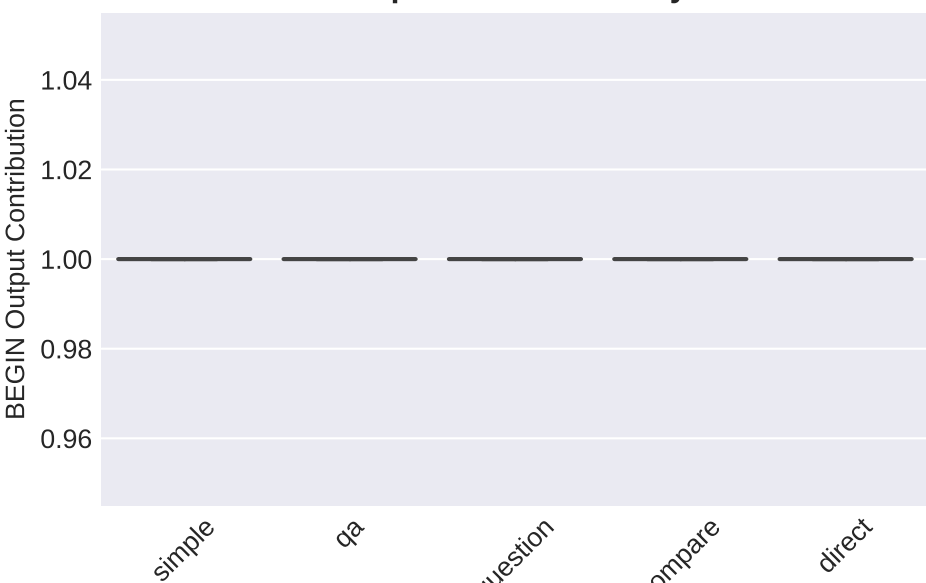
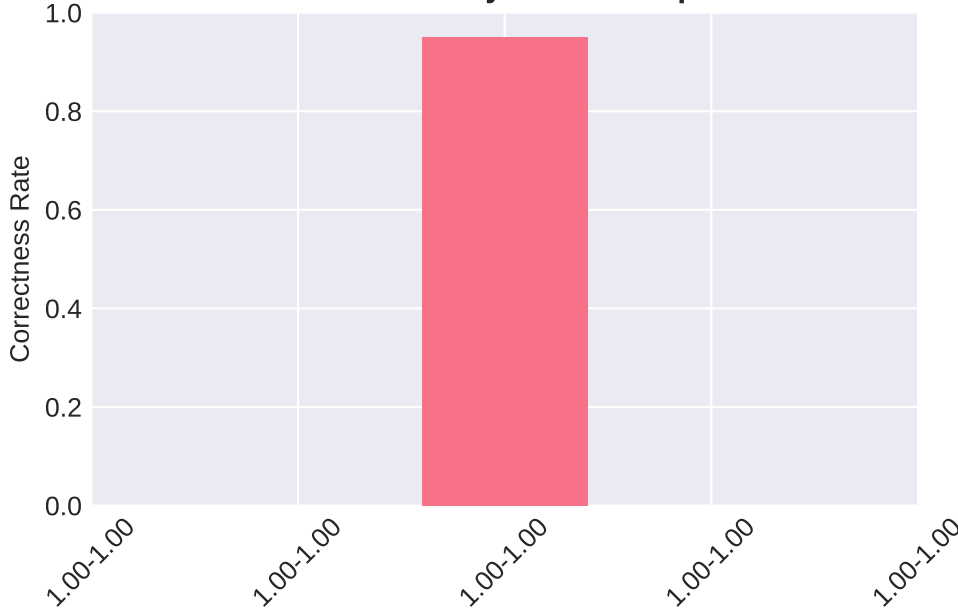


Attention OUTPUT Quantification: Information Flow Analysis at Layer 10

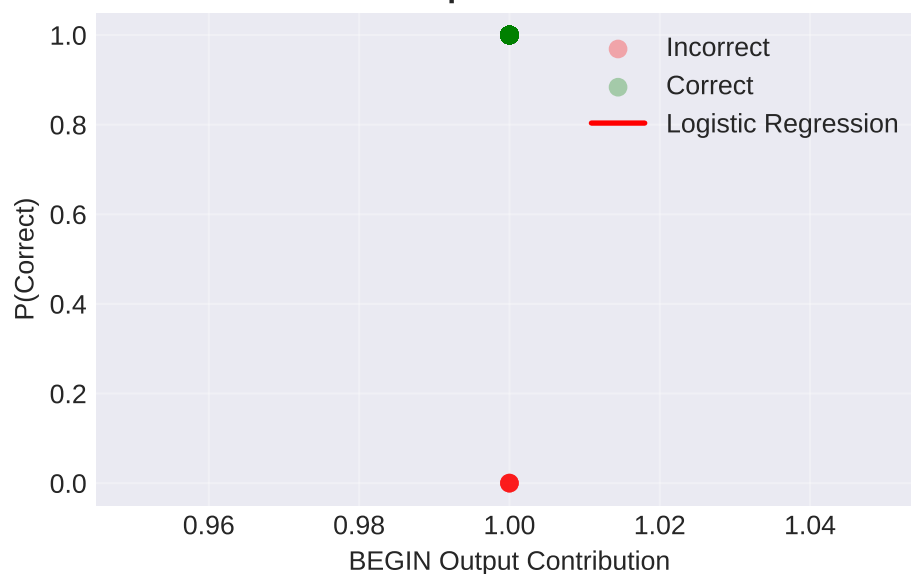
BEGIN Output Contribution by Format



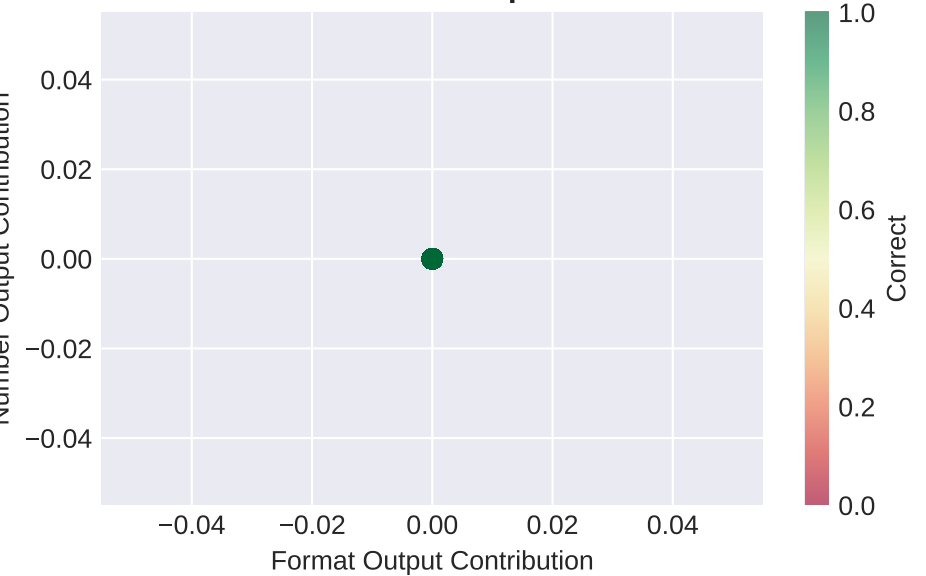
Correctness Rate by BEGIN Output Level



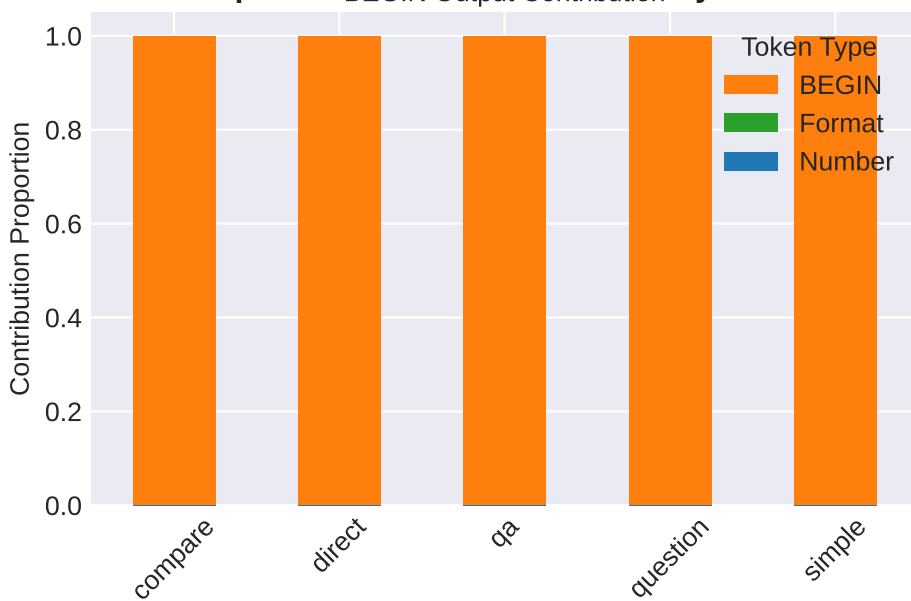
BEGIN Output vs Correctness



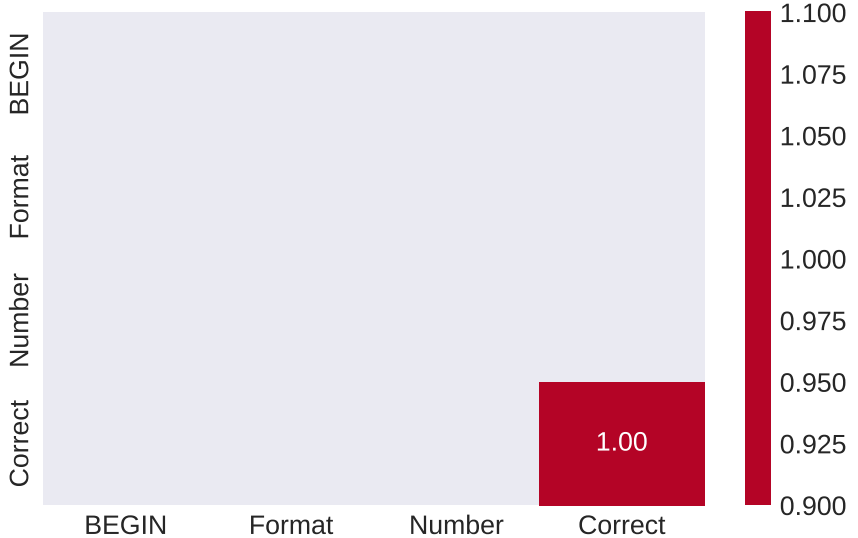
Format vs Number Output Trade-off



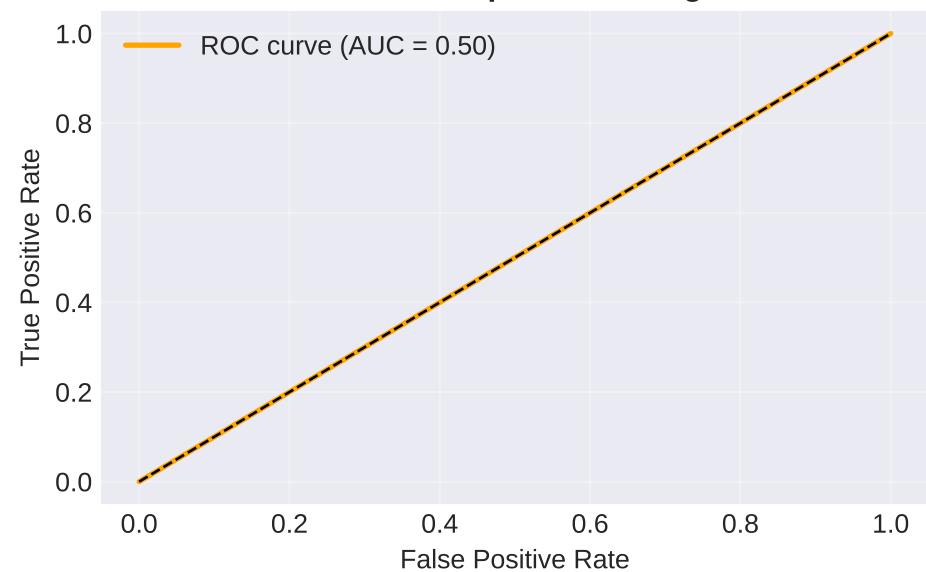
Output Contribution Breakdown by Format



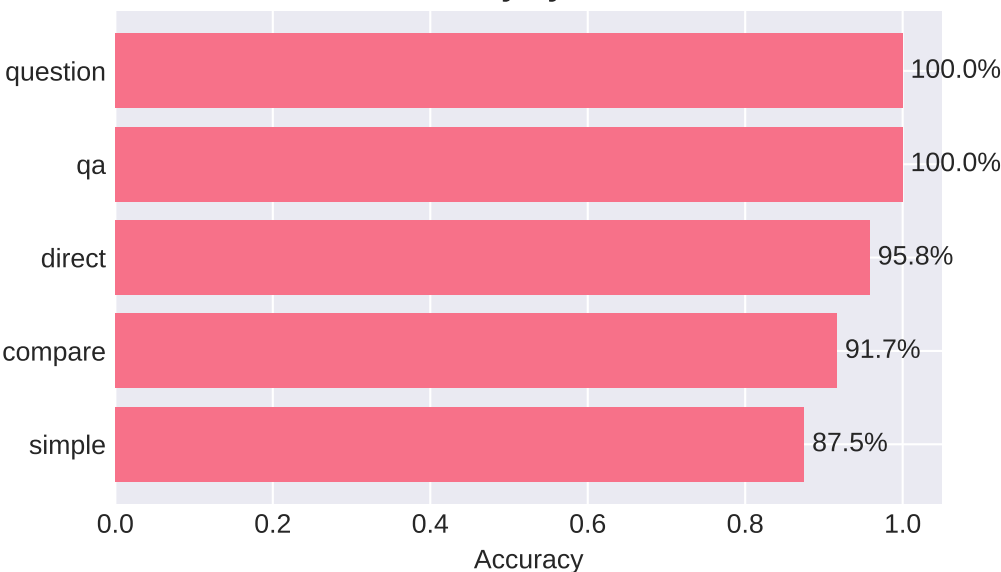
Correlation Matrix



ROC Curve: BEGIN Output Predicting Correctness



Accuracy by Format



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!