Comprehensive Pipeline Statistics

Phase 1: Activation Patching

Metric	Value
Total causal features	2,787
Total patching tests	334,440
Games per feature	120
Layers analyzed	25, 30

Phase 5: Prompt Correlation

Metric	Value
Total features tested	2,787
Significant (p<0.05)	3,425
Risky features	1,701
Safe features	1,724
Significance rate	122.9%

Phase 4: Word Association

Metric	Value
Total correlations	7,366,041
Unique words	~10,000
Unique features	2,787
Coverage	100%

Overall Pipeline Summary

Pipeline Stage	Input	Output
Feature Discovery	6,400 exp	2787 features
Patching Test	2787 features	334,440 tests
Statistical Filter	2787 features	3425 sig.
Classification	3425 sig.	1701R+1724S