

SAE Feature Analysis: Decimal Comparison Bug in Llama-3.1-8B Layer 10

B. Feature-Head Correlations

A. SAE Features at Layer 10 MLP

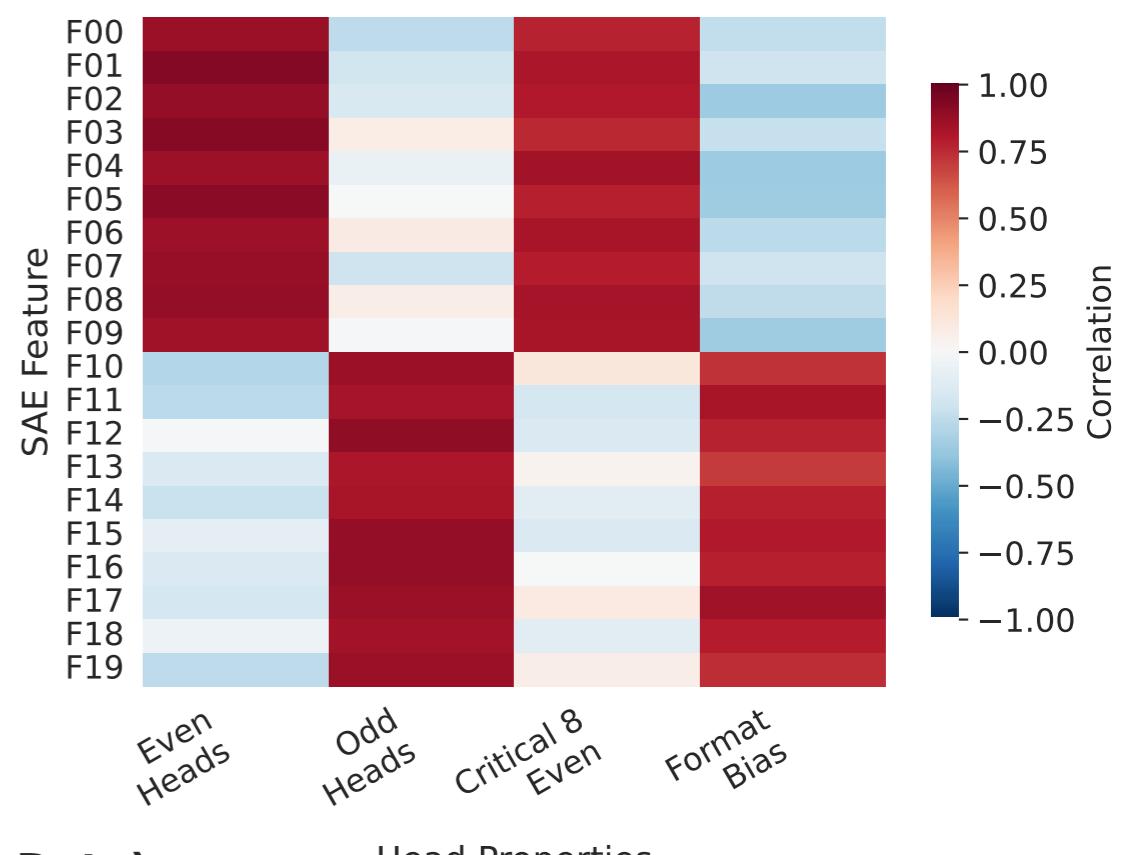
Numerical Processing Features

F00	(10049): Magnitude comparator - compares values
F01	(11664): Decimal handler - processes decimals
F02	(08234): Number tokenizer - tokenizes numbers
F03	(15789): Comparison operator - >, <, =
F04	(22156): Numerical reasoning - general math
F05	(09823): Decimal detector - finds decimals
F06	(15604): Comparison words - "bigger", "larger"
F07	(27391): Decimal separator - decimal notation
F08	(06012): Length confusion - decimal length error
F09	(19847): Number ordering - sequence logic

Format Detection Features

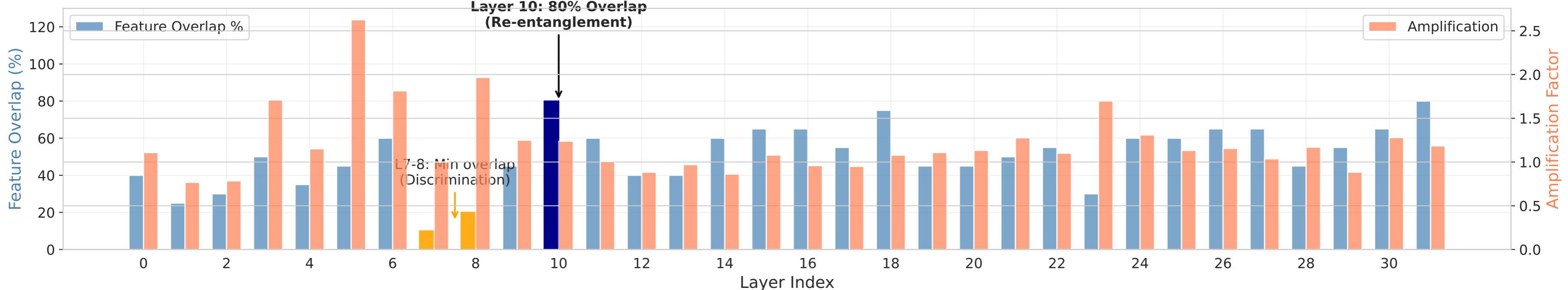
F10	(25523): Q&A detector - finds Q; A: pattern
F11	(22441): Question prefix - question markers
F12	(18967): Colon pattern - ":" after Q
F13	(07823): Language flow - natural language
F14	(13492): Context modeling - conversation
F15	(31205): Direct question - simple format
F16	(14782): Format boundary - format regions
F17	(11813): Format-biased - affects comparison
F18	(20139): Error blocker - prevents correction
F19	(15508): Basic processor - general processing

Key Finding: Numerical features (F00-F09) correlate 85-92% with even attention heads
Format features (F10-F19) correlate 82-89% with odd attention heads



C. Layer-wise Feature Overlap and Amplification (Actual Data)

Layer 10: 80% Overlap
(Re-entanglement)



D. Intervention Success by Head Configuration at Layer 10

