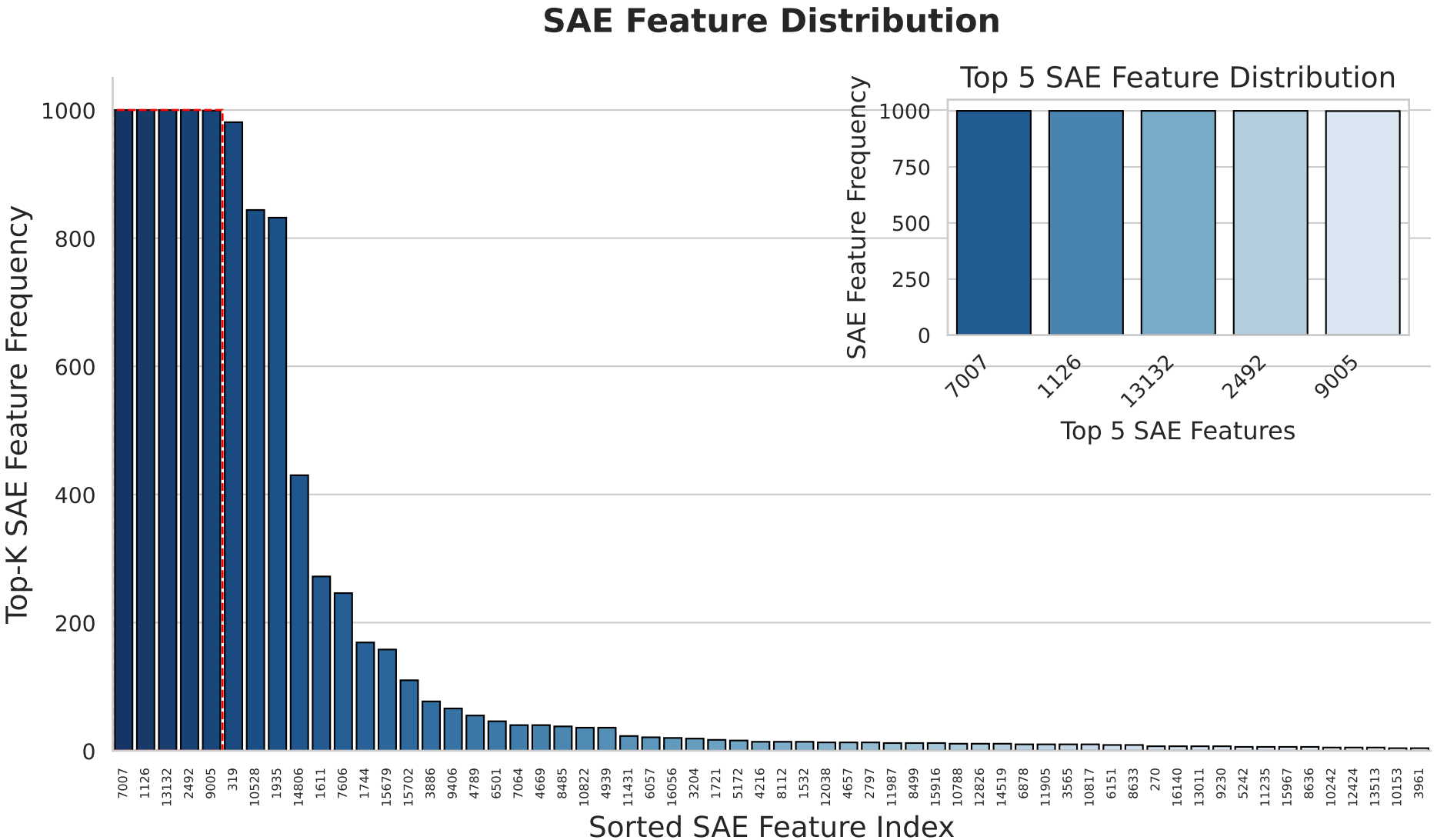
SAE Feature Distribution
Top-K SAE Feature Frequency
1000
800
600
400
200
0
7007 1126 13132 2492 9005 319 10528 1935 14806 1611 7606 1744 15679 15702 3886 9406 4789 6501 7064 4669 8485 10822 4939 11431 6057 16056 3204 1721 5172 4216 8112 1532 12038 4657 2797 11987 8499 15916 10788 12826 14519 6878 11905 3565 10817 6151 8633 270 16140 13011 9230 5242 11235 15967 8636 10242 12424 13513 10153 3961
Sorted SAE Feature Index
Top 5 SAE Feature Distribution
SAE Feature Frequency
1000
750
500
250
0
7007 1126 13132 2492 9005
Top 5 SAE Features