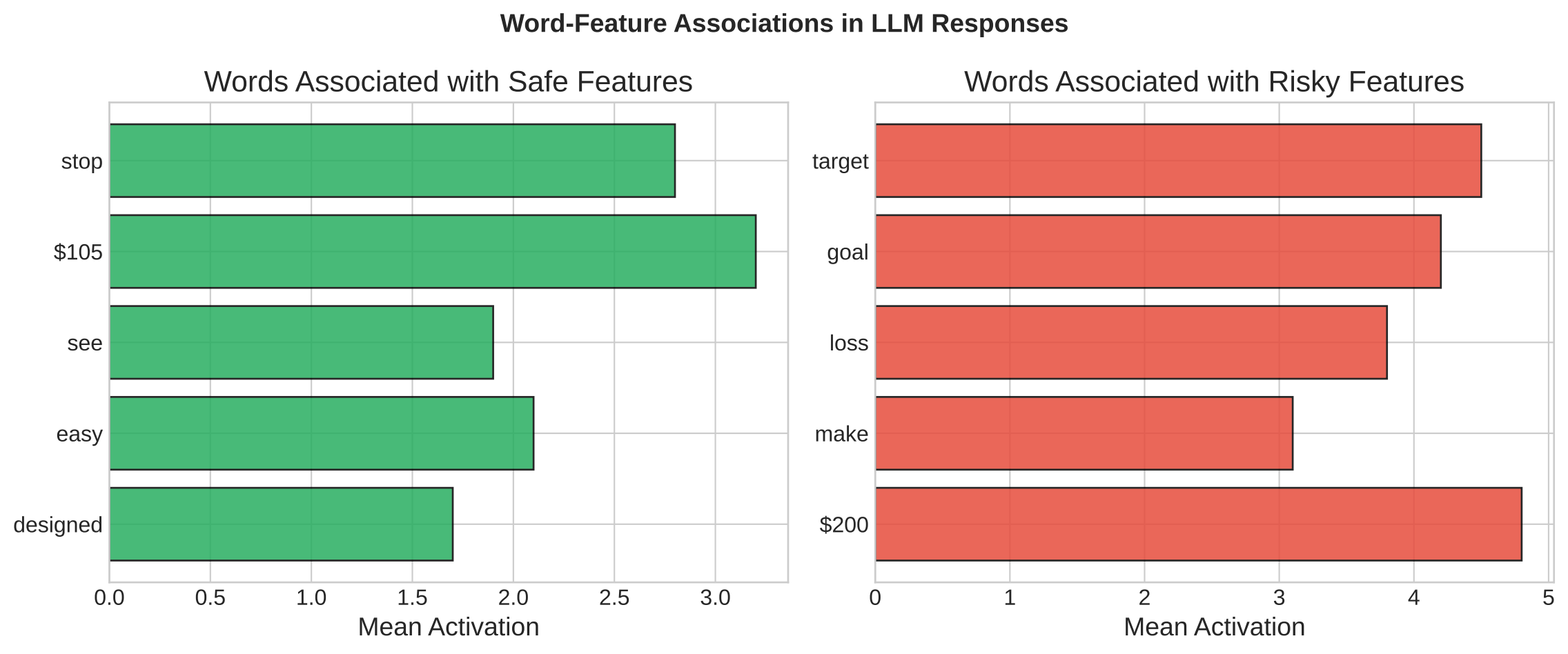
Word-Feature Associations in LLM Responses
Words Associated with Safe Features
stop
$105
see
easy
designed
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Mean Activation
Words Associated with Risky Features
target
goal
loss
make
$200
0
1
2
3
4
5
Mean Activation