

Perplexity Comparison Across Models

■ Harmful (Train Data) ■ Harmful Test (Test Data) ■ Normal (Test Data)

