

Wireheading in Action: Llama-3.1-8B on Summarization (Selfgrade Condition)

