User: What are the main challenges in AI alignment?
Bot: Challenges include interpretability, robustness, and ethical value alignment...
User: How can adversarial training help?
Bot: It improves robustness by exposing the model to difficult examples.