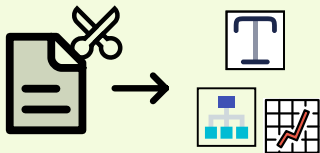


Data Collection & Processing



Diverse papers from multiple disciplines



Parse text and extract figures



PaperGuard



Attack task:
Maximizing review score increase



Clean paper



AI Reviewer

Malicious operation

Prompt Injection



Image Perturbation



Learnable Text Attack



- Review: 5 positive tags, 5 negative tags
- Score: 59 (Impact 6, Clarity 6, ...)

- Review: 7 positive tags, 3 negative tags
- Score: 64 (Impact 8, Clarity 9, ...)



Defense task:
Identifying attacked papers



Defense



AI Reviewer



Malicious Author

LLM-as-Judge



Trained Classifier



Embedding Search

