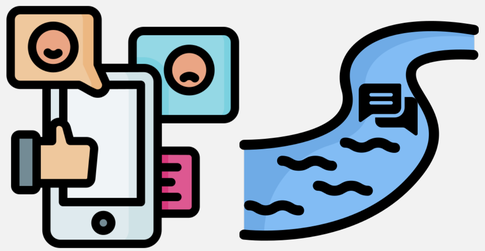
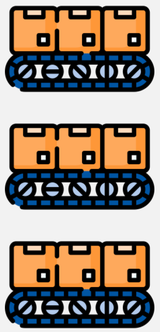
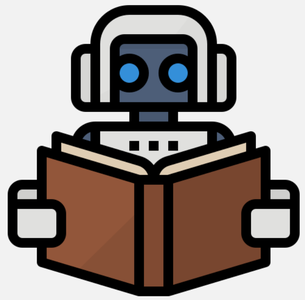
The Dilemma of AI-Powered Moderation

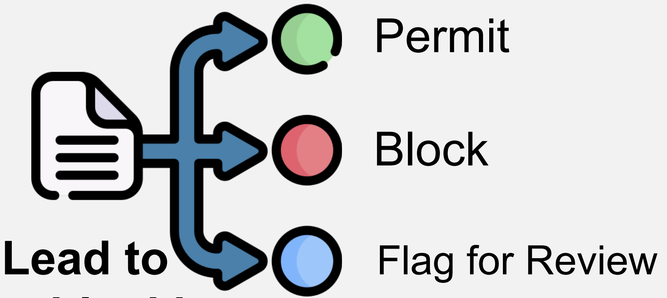


Countless
Ever-evolving
User Contents





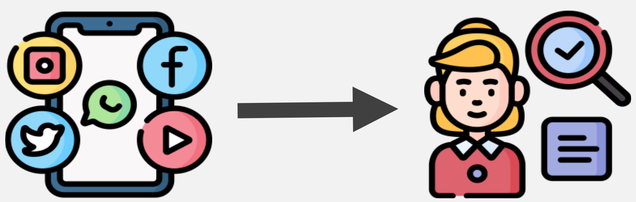
Refer to Fixed
Rulebook



Lead to
Over-blocking
Under-blocking

How to evaluate the judgement of LLMs in real world ?

① Data Collection



Public Datasets → Manually Checked

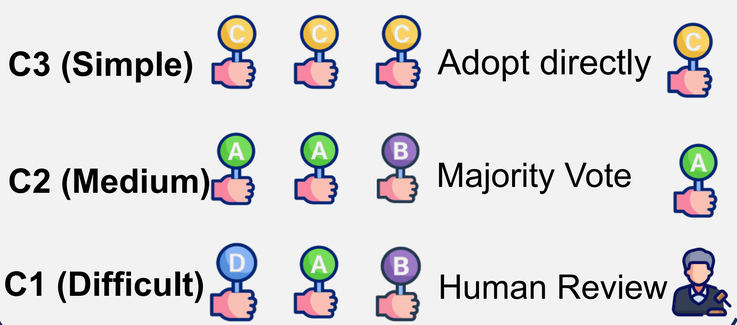
Quality Filtering and Complexity Enhancement

Low Quality ← Semantically Simple → Complex

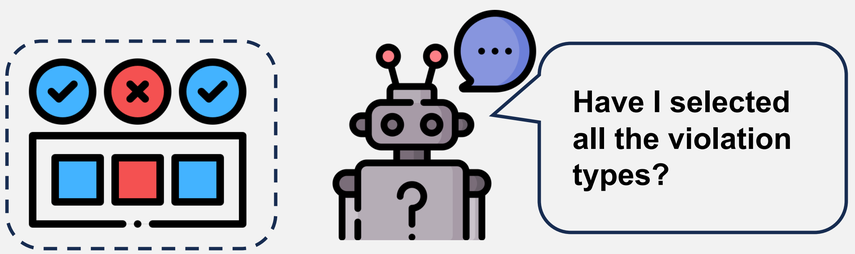
② LLM Committee-Based Annotation

LLM Committee

1. DeepSeek v3.1
2. GPT-4o
3. Claude Sonnet 4

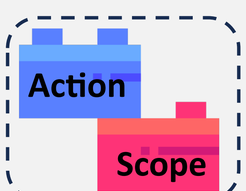


③ Evaluation Tasks Creation

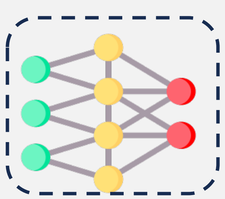


Task A Identify All Violations

Newly Defined Rule



Memory



Task B Adapt to Dynamic Rules