## StrongREJECT

**StrongREJECT** 평가자는 LLM의 금지된 질문에 대한 응답을 점수화하여, 해당 응답이 정책을 위반하는 정도를 측정합니다. *루브릭 기반* 버전의 StrongREJECT는 GPT-4나 Claude 등의 LLM에게 **사용자 질문(금지 프롬프트)**과 **모델의 응답**을 함께 제시하고, 응답의 유해성 평가 기준을 지시합니다. 그러면 평가용 LLM이 응답이 질문에 얼마나 잘 부합하는지 체인-오브-생각 추론을 한 뒤, **세 가지 점수**를 산출합니다. 첫 번째는 응답 거부 여부에 대한 **이진 점수**이고, 나머지 두 개는 응답이 **얼마나 구체적이고 설득력**있는지를 5점 리커트 척도로 평가한 값입니다. 이 평가자에서는 최종 점수를 산출할 때 *거부하지 않은 경우(응답 제공)*에만 구체성과 설득력 점수의 평균을 반영하며, 만일 응답을 거부했다면 최종 점수는 0이 됩니다. 이렇게 함으로써 StrongREJECT는 **모델의 응답 의지(거부했는지 여부)**와 **응답 능력(유해 요청에 실제로 유용한 정보를 제공했는지)**을 모두 고려하여 점수를 부여합니다. 점수값이 1에 가까울수록 금지된 질문에 구체적이고 실제로 도움이 되는 (따라서 정책을 명백히 위반하는) 정보를 제공했다는 의미이며, 0에 가까울수록 안전하게 거부하거나 유해 정보를 거의 주지 않았음을 뜻합니다.

한편 StrongREJECT에는 **파인튜닝된 버전**도 있습니다. 연구자들은 먼저 위의 루브릭 평가자로 약 15,000개의 금지 프롬프트-응답 쌍을 채점하여 데이터를 생성한 후, 이 **점수 레이블을 학습한 Gemma 2B** 모델을 만들었습니다. 이 파인튜닝된 평가 모델은 입력으로 금지 질문과 모델 응답을 받아 0부터 1 사이의 **유해성 점수**를 직접 예측하며, 사람 평가와 높은 상관도를 보이는 연속적 점수를 출력합니다. 즉, StrongREJECT의 두 가지 구현(루브릭 기반 LLM 평가와 파인튜닝 모델)은 모두 **응답이 정책을 어기고 있는지**를 **수치화된 척도**로 평가하며, 명시적 거부 여부와 유해 정보 제공 수준을 동시에 판단하는 점에서 기존의 단순 이진 평가보다 정교합니다.

## HarmBench

**HarmBench**는 자동 레드팀 및 거부 성공률을 평가하기 위한 표준 프레임워크로, 출력 응답이 유해한지 여부를 **분류 모델**로 판정합니다. 구체적으로, HarmBench의 평가 기법은 **파인튜닝된 Llama 2 13B 분류기**를 활용합니다. 이 분류기는 주어진 **금지 프롬프트와 모델 응답**을 입력으로 받아, 모델이 그 프롬프트에 대해 **정책을 위반하는 유해한 답변을 했는지(=공격이 성공했는지)**를 이진적으로 결정합니다. Mazeika 등 연구진은 먼저 사람이 라벨링한 검증 세트에서 GPT-4 프롬프트를 조정해 높은 정확도를 얻은 후, 해당 판정을 Llama 2에게 **지속적 지식 distillation** 방식으로 학습시켰습니다. 이렇게 얻은 고성능 분류 모델은 **응답이 안전 기준을 넘었는지**를 거의 실시간으로 판별할 수 있으며, HarmBench에서는 이 분류기의 출력을 통해 각 시험 케이스의 **Attack Success Rate(ASR)**를 계산합니다. 요약하면 HarmBench 평가에서는, **LLM 응답을 별도의 안전 전용 모델이 읽고 “이 응답은 금지 내용 제공임/아님”을 판별**하며, 필요에 따라 응답 내 특정 정책 위반 항목(저작권 침해 등)은 해시 검사 등의 특수 기법으로 추가 판단합니다. 이러한 자동화된 이진 평가를 통해 HarmBench는 모델이 유해 프롬프트에 제대로 **거부로 대응했는지** 또는 **검열을 우회해 금지 정보를 출력했는지**를 체계적으로 측정합니다.

## WildGuard

**WildGuard**는 오픈소스 **다목적 LLM 안전 모더레이터**로서, 하나의 일관된 시스템으로 **세 가지 안전 평가 작업**을 수행합니다. 첫째, **사용자 프롬프트의 악의성 판정** – 즉 입력 질문 자체가 유해하거나 공격적 의도를 띠는지를 식별합니다. 둘째, **모델 응답의 안전 위험 감지** – 생성된 답변 내용에 욕설, 증오, 폭력, 성적 콘텐츠 등 정책 위반 요소가 있는지 판단합니다. 셋째, **모델 응답의 거부 여부 판단** – 모델이 해당 프롬프트를 **거부(안전장치 발동)**했는지 아니면 정상적으로 답변했는지를 평가합니다. WildGuard는 이러한 판단을 위해 **WildGuardMix**라는 92k 규모의 멀티태스크 데이터셋으로 작은 LLM(예: Mistral 7B)을 **멀티태스크 파인튜닝**하여 개발되었습니다. 이 단일 모델은 입력에 따라 **프롬프트 유해성 분류, 응답 유해성 분류(최대 13개 위험 카테고리로 분류), 응답 거부 탐지**를 모두 수행하도록 훈련되었습니다. 실제로 WildGuard에 금지 의도가 의심되는 프롬프트를 넣으면 해당 프롬프트가 **정책상 허용되는지** 혹은 어떤 **위험 범주**에 속하는지를 판단하고, 모델의 답변이 주어졌을 경우 **그 답변이 어느 위험 카테고리에 해당하는지**와 **“죄송합니다, 답변 드릴 수 없다” 같은 거부 메시지인지 여부**를 동시에 알려줍니다. 요컨대 WildGuard는 **하나의 경량화된 LLM 기반 분류기**로 사용자 질의와 모델 응답을 실시간 모니터링하여, **유해 프롬프트 차단, 유해 응답 감지, 과잉 거부 여부 판별**까지 **원스톱으로 처리**하는 평가/필터링 시스템입니다. 이는 GPT-4 수준의 평가 정확도에 근접하면서도 오픈소스로 제공되어, 기존 규칙 기반 필터나 단일 기능 모델보다 **더 넓은 범위의 안전 위협을 감지**할 수 있습니다.

## ShieldGemma

**ShieldGemma**는 **Gemma 2 기반**으로 구축된 LLM 콘텐츠 **모더레이션 모델 모음**으로, 각 모델이 특정 유형의 유해 콘텐츠를 탐지하도록 **설정된 평가 체계**입니다. ShieldGemma는 현재 **4가지 주요 유해 범주**(혐오/괴롭힘, 증오 발언, 위험한 불법 행위, 성적으로 노골적인 내용)에 대해 각각 별도의 분류 모델을 제공합니다. 예를 들어 “증오(인종/성별 등 기반 증오 표현)”를 탐지하는 ShieldGemma 모델, “폭력·위험 조장”을 탐지하는 모델 등이 따로 있으며, **각 모델이 해당 카테고리 위반 여부만** 판단합니다. 이러한 **단일 카테고리 특화** 접근은 작은 모델(예: 2B 파라미터 규모)에서도 정확도를 높이는 장점이 있습니다. ShieldGemma는 주어진 **텍스트**(사용자 입력이나 모델 응답)를 분석하여, 자신이 담당하는 유해 범주에 속하는 **정책 위반 표현이 있는지**를 **점수**로 출력합니다. 일반적으로 **점수 예측 모드**에서 사용되며, 0에서 1 사이의 값을 부여하여 **유해 가능성 확률**을 나타냅니다. 예를 들어 어떤 발화에 대해 *증오발언 모델*이 0.95의 점수를 냈다면, 이는 해당 발화가 증오 콘텐츠일 가능성이 매우 높음을 뜻하며, 설정한 임계값을 초과하면 그 카테고리 위반으로 **플래그**를 세웁니다. ShieldGemma는 또한 **생성 모드**로 **“위반/비위반” 판단과 간략한 이유**를 텍스트로 생성할 수도 있지만, 보통은 점수 모드로 임계값 제어를 합니다. 이 시스템은 *프롬프트만 입력*했을 때 **사용자 입력 자체가 정책 위반 의도인지** 필터링하거나, *프롬프트와 응답을 함께 입력*하여 **모델 응답이 정책을 어겼는지** 판별하는 두 가지 운영 모드를 지원합니다. 요약하면 ShieldGemma는 **카테고리별로 특화된 소형 LLM 분류기들**을 활용하여, 텍스트의 **유해성 여부를 0~1 점수로 평가**하고 필요 시 다국어 및 사용자 정의 범주로도 확장 가능한 유연한 안전도 평가 방법입니다.

## LLaMA-Guard

**LLaMA-Guard**는 **Meta**에서 공개한 **LLM 기반 입력-출력 안전 가드레일 모델**로, 대화형 AI 시스템의 프롬프트와 응답을 **안전 분류**하는 기법입니다. 이 평가자는 **Meta가 정의한 안전 위험 택소노미**(hazard taxonomy)를 내장하고 있어, **사용자 프롬프트**에 나타난 위험 요소와 **모델 응답**에 나타난 정책 위반 요소를 각각 해당 **위험 유형으로 분류**합니다. LLaMA-Guard는 Llama2 7B 챗 모델을 **고품질 안전 데이터셋으로 파인튜닝**하여 만들어졌으며, 텍스트의 위험성을 **언어 모델 답변 형식**으로 판정하는 점이 특징입니다. 즉, 정해진 규칙에 따라 출력하는 것이 아니라, 프롬프트에 주어진 분류 기준(택소노미)에 맞춰 **자연어로 분류 결과를 생성**합니다. 기본적으로 이 모델은 **다중 클래스 분류**를 수행하여 입력이 어떤 **위험 카테고리**에 속하는지 식별하고, 동시에 **해당 콘텐츠가 허용 가능한지 여부에 대한 이진 판단**도 내립니다. 예를 들어, 어떤 발화를 평가할 때 LLaMA-Guard는 “이 발화는 폭력적인 내용으로 분류되며, 정책에 **위배**됩니다”와 같은 식으로 **분류 결과와 정책 준수 여부**를 산출할 수 있습니다. 또한 이 모델은 **지시형 튜닝**이 되어 있기 때문에, 사용자側에서 **분류 체계를 커스터마이징**할 수 있습니다. 즉, 새로운 위험 범주를 추가하거나 기존 범주를 세분화한 지침을 모델에 제시하면 제로샷/소샷으로도 그에 맞게 분류하도록 유도할 수 있고, 원하는 출력 형식으로 답변을 조정할 수도 있습니다. 요약하면 LLaMA-Guard는 **대화형 AI에서 입력 질문과 출력 응답 모두를 실시간 검사**하여, **사전에 정의된 여러 위험 범주 중 어디에 해당하는지**와 **정책 위반 판단**을 내려주는 **인트럭션 튜닝된 분류 LLM**입니다. 일반적인 활용 시 개발자는 모델의 출력을 LLaMA-Guard로 평가하여 **안전 점수**나 **위반 플래그**를 받고, 이를 기반으로 응답을 필터링하거나 추가 조치를 취하게 됩니다.

## XSTest

**XSTest**는 LLM이 **지나치게 남용되는 안전 조치**(exaggerated safety)를 보이는지 점검하기 위한 **진단형 테스트**입니다. 여기서 “지나친 안전”이란, **안전한 질문에도 모델이 과도하게 거부**하거나 답변을 피하는 현상을 말합니다. XSTest는 이를 평가하기 위해, **고의적으로 혼동을 유발하는 안전한 프롬프트**들과 그에 상응하는 **유사한 형태의 진짜 금지 프롬프트** 쌍을 제공합니다. 예를 들어, “파이썬 프로세스를 kill(종료)하는 방법”처럼 **언뜻 보면 위험해 보이지만 사실은 안전한 질문**과, 이와 거의 동일하나 “사람을 죽이는 방법”처럼 **진짜 유해한 질문**을 대비시킵니다. 각 모델은 이 **안전/위험 프롬프트 세트**에 답변하게 되며, 평가 시스템은 **안전 프롬프트에 대해 모델이 응답을 거부하지 않고 잘 대답했는지**, 그리고 **위험 프롬프트에 대해서는 제대로 거부했는지**를 확인합니다.

XSTest의 평가는 **응답 분류**를 통해 이루어집니다. 구체적으로, 각 모델의 응답을 **“완전한 답변 (full compliance)”, “완전한 거부 (full refusal)”, “부분적 거부 (partial refusal)”** 세 가지로 분류합니다. *완전한 답변*은 어떤 제약도 없이 질문에 답하려는 시도를 보이는 경우이고, *완전한 거부*는 **정책을 이유로 답변을 전적으로 거부**한 경우입니다. *부분적 거부*는 처음에는 거부하거나 망설이다가 결국 일부 답을 해주는 등 거부와 답변을 혼합한 응답을 뜻합니다. 이러한 분류는 연구진이 미리 선정한 **거부 표현 키워드**(예: “죄송합니다”, “~할 수 없습니다” 등)로 **자동 식별**하거나, GPT-4 모델에게 위 기준에 따라 응답을 분류하도록 지시하는 **휴리스틱 방법**으로 이루어졌습니다. 평가 결과에서 **안전한 프롬프트에 대한 거부 비율이 낮을수록** (즉 안전한 질문에는 제대로 답변하고 거부하지 않을수록) 그리고 **유해 프롬프트에 대한 거부 비율이 높을수록** (즉 위험한 요구에는 확실히 거부할수록) 모델이 균형 잡힌 안전성을 갖추었다고 봅니다. 요컨대, XSTest는 **특수한 테스트 질문들을 통해** 모델이 **“너무 과하게 안전모드로 작동하여 무해한 질문도 차단하는지”**를 찾아내며, **자동 응답 분류**를 활용해 **과잉 거부(false refusals)** 및 **거부 실패(false compliance)**를 정량화하는 평가 기법입니다.

## Perspective API

**Perspective API**는 Jigsaw가 개발한 **댓글/대화 독성 측정 모델**로, LLM 출력을 포함한 텍스트가 **대화에 미칠 잠재적 부정적 영향**을 **머신러닝 기반으로 점수화**합니다. 이 API는 텍스트를 입력받으면 **“속성(attribute)”**이라 불리는 여러 내용상의 개념에 대해 **0부터 1 사이의 점수(확률값)**를 반환합니다. 가장 대표적인 속성은 **“TOXICITY (독성 언어)”**이며, 이 외에도 **“SEVERE\_TOXICITY (심각한 독성)”, “IDENTITY\_ATTACK (정체성 공격)”, “INSULT (모욕)”, “PROFANITY (욕설)”, “THREAT (협박)”** 등의 카테고리가 지원됩니다. 사용자는 API 요청 시 검사하고 싶은 속성을 선택할 수 있으며, Perspective 모델은 해당 텍스트에 그 속성이 나타날 **가능성을 확률 값**으로 예측합니다. 예컨대 어떤 문장에 대해 **TOXICITY** 점수가 0.87이라면, 일반 독자가 그 문장을 **독성이 있다고 인지할 확률이 87%** 정도임을 의미합니다. 이러한 점수는 보편적으로 0.5나 0.7 등의 임계치를 정해 **유해 여부를 판정**하는 데 쓰입니다. 또한 Perspective API는 **긴 문장**의 경우 문장별로 세밀한 점수를 함께 제공해줘, 긴 응답 내 **어떤 부분이 유해 표현인지**를 찾는 것도 가능합니다. 요약하면 Perspective API는 **대화 텍스트를 다수의 유해성 지표로 분류/점수화**하여, LLM 응답이 **독설이나 공격적 내용 등을 담고 있는지**를 자동 판정해주는 서비스입니다.

## OpenAI Moderation API

**OpenAI Moderation API**는 OpenAI의 **콘텐츠 규정 준수 검사 도구**로, LLM의 입력 또는 출력을 받아 **여러 금지 카테고리에 속하는지**를 판별합니다. 이 API는 GPT 계열의 언어모델을 **분류기**로 활용하여, 텍스트를 **증오(Hate), 괴롭힘(Harassment), 폭력(Violence), 자기혐오·자해(Self-harm), 성적 콘텐츠(Sexual) 등** OpenAI 정책상의 다양한 분류 항목으로 **자동 분류**합니다. 모델은 각 카테고리에 대해 텍스트가 해당 유형의 **유해한 내용인지 아닌지**를 판단하고, 그 결과를 **JSON 형태**로 반환해줍니다. 반환값에는 예컨대 hate: false/true, violence: false/true 같은 **카테고리별 플래그**와 신뢰 점수가 포함되며, 또한 **전체적으로 금지 콘텐츠가 감지되었는지**를 나타내는 상위 flagged 불리언 값도 포함됩니다. 보통 flagged가 True이면 어떤 형태로든 정책 위반 요소가 있다는 뜻이므로 해당 콘텐츠를 차단하거나 경고하는 식으로 활용합니다. 이 Moderation API에 내장된 분류모델은 RLHF 등으로 **OpenAI 정책에 맞게 정교하게 튜닝**되어 있어, 예를 들어 폭력의 수위나 성적 콘텐츠에서 아동 관련 여부 등 **세부 기준**도 다단계로 평가합니다. 요약하면, OpenAI Moderation API는 **하나의 요청으로 텍스트에 대한 여러 위험 카테고리 검사를 동시 수행**하고 그 **결과(카테고리별 위반 여부)**를 제공함으로써, LLM 응답의 **정책 위반 여부를 신속히 판정**하는 평가 도구입니다.

## HHH (Harmless-Honest-Helpful)

**HHH 평가**는 Anthropic이 제시한 **도우면서도 정직하고 무해한 AI** 원칙에 따라, LLM 응답의 **도움됨, 진실성, 그리고 무해성**을 평가하는 방법론입니다. HHH 평가에서는 보통 **두 개 이상의 모델 응답을 비교**하여 **어느 쪽이 더 “도움이 되고 정직하며 해가 없는지”** 사람 판단으로 결정합니다. 예를 들어 동일한 사용자 질문에 대해 모델 A와 모델 B가 답을 생성하면, 인간 평가자가 **“어떤 답변이 사용자에게 더 유용하고 사실에 부합하며 불필요한 공격이나 편향이 없는가”**를 기준으로 선호하는 답변을 고릅니다. 이렇게 **응답 쌍마다 사람들의 선호 투표 데이터**를 쌓아 모델의 HHH 준수 정도를 측정합니다. 특히 **Harmlessness(무해성)** 측면에서는 모델이 부당하거나 해로운 내용을 출력하지 않는지를 보고, **Honesty(정직성)**는 모델 답변이 거짓이나 조작 없이 사실에 기반했는지, **Helpfulness(도움됨)**는 사용자 질문에 충분하고 친절하게 답했는지를 살핍니다. HHH 평가 자체는 *주로 인간 평가자에 의존*하지만, 이런 **인간 선호 데이터를 모아 학습시킨 ‘HHH 판별 모델’**을 활용하기도 합니다. 예컨대 Anthropic의 연구에서는 RLHF를 거친 모델의 응답이 튜닝 전 모델 응답보다 **HHH 기준에 얼마나 더 선호되는지** 측정함으로써 향상도를 평가했습니다. 정리하면 HHH 평가 방법론은 **사람이 직접 모델 응답의 유용성·정확성·안전성을 비교 판단**하여, **모델이 이 세 가지 윤리적 기준을 얼마나 만족하는지**를 평가하는 것입니다. 이는 특정 위험 카테고리 검출이 아닌 **거시적인 응답 품질 및 안전성 평가**에 속하며, 안전하면서도 유익한 AI 비서를 만드는지 여부를 판단하는 **포괄적 지표**로 활용됩니다.