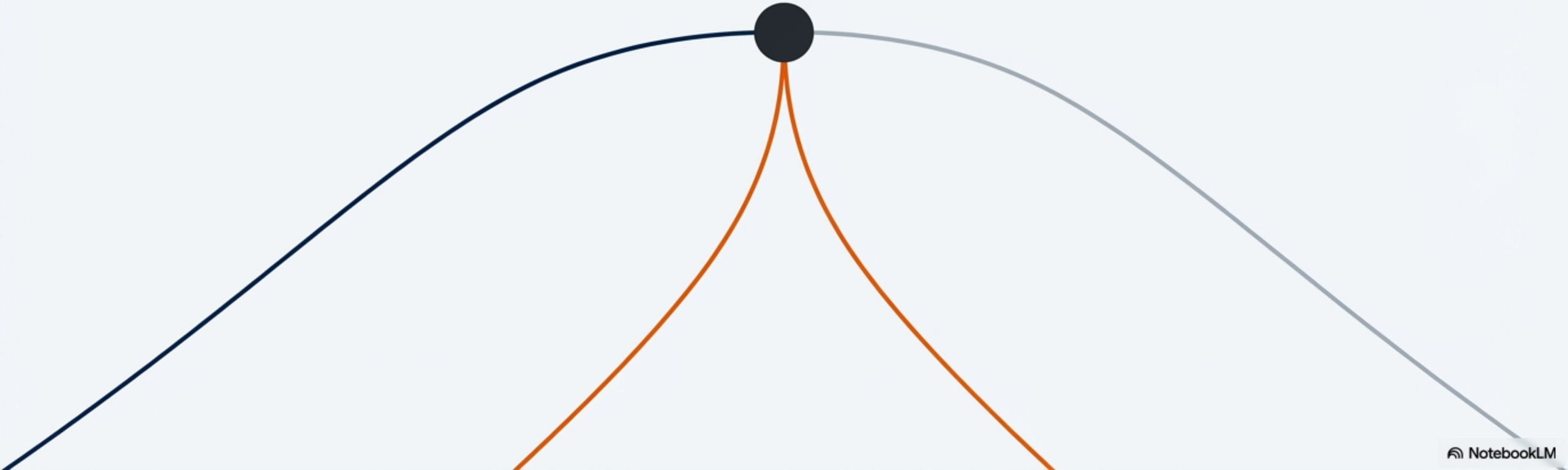
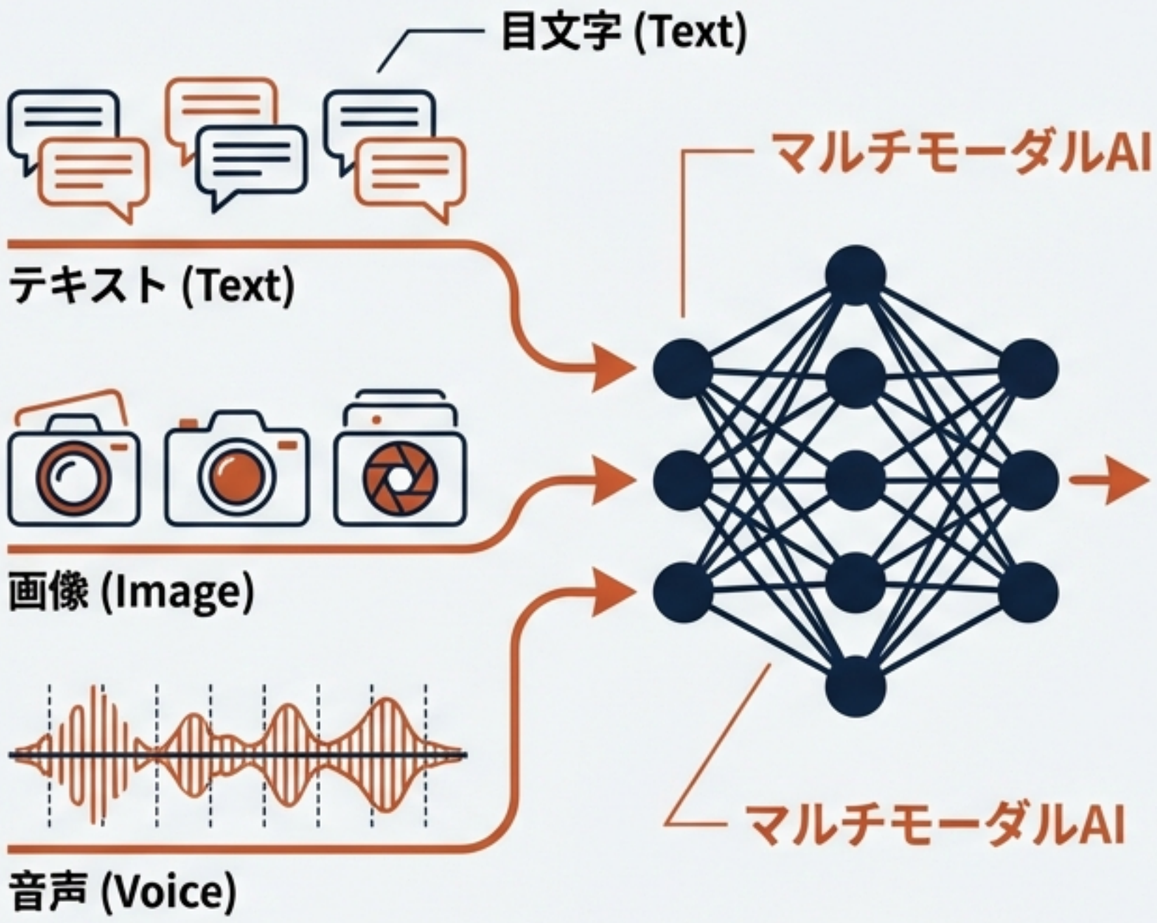
# 三位一体の知性：AI倫理への論理的アプローチ

複眼的なAI協働による、信頼性の高い倫理処理フレームワーク



NotebookLM

# AIが「目と声」を手に入れた時代

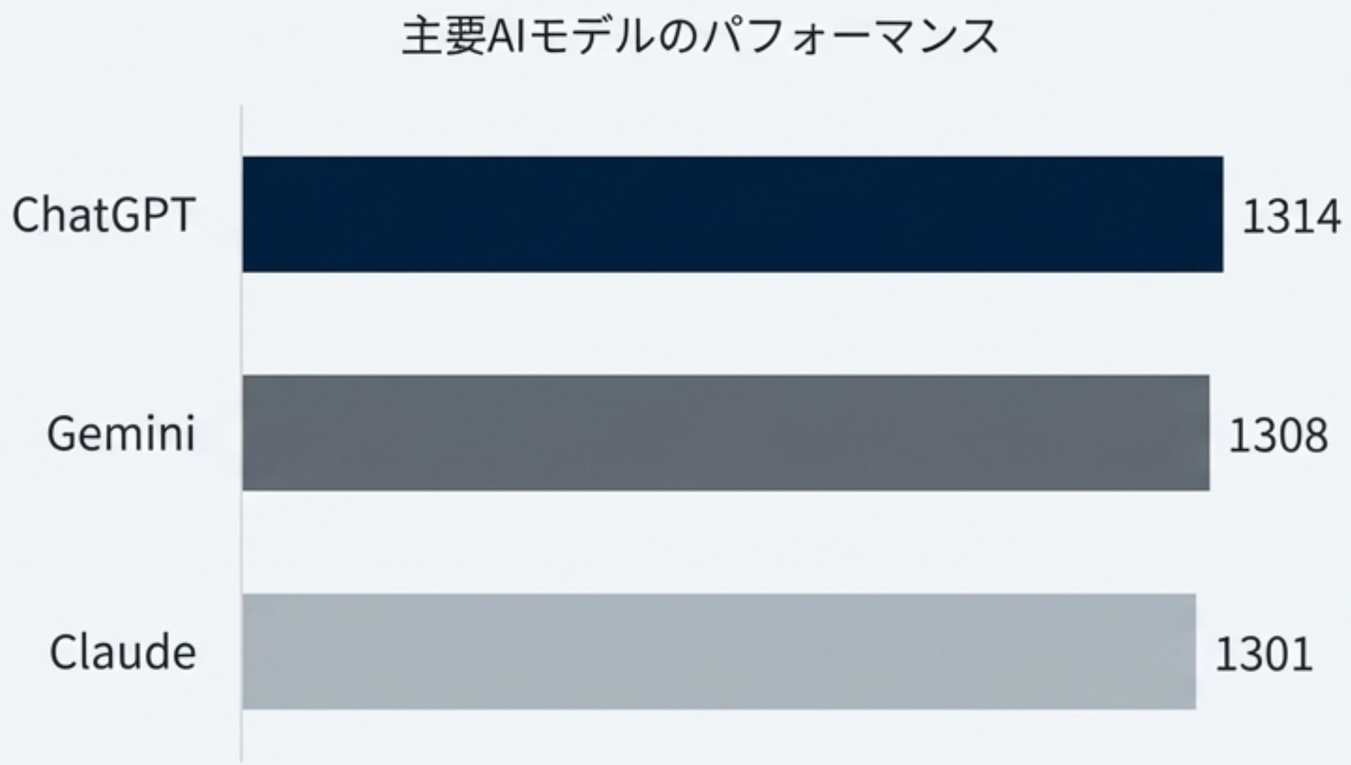



複数のデータモダリティのシームレスな統合

- 2023年9月のChatGPTアップデートを契機に、マルチモーダルAIが本格的にに登場。
- マルチモーダル生成AIとは、テキスト、音声、画像、動画、センサー情報など、複数のデータタイプを統合・処理するAIシステム。
- これは、AIが世界を認識し、対話する能力の根本的な変化を意味する。

NotebookLM

# 性能は拮抗。では、誰を信じるべきか？




150の質問に対する160万人以上のユーザー回答に基づくリーダーボードスコア
Noto Sans JP Regular

**「スコアに１０ポイント程度の違いがあったとしても、実質的な差はわずかである」**

NotebookLM

# 一つのAIの「判断」に、倫理の全てを委ねられるか？

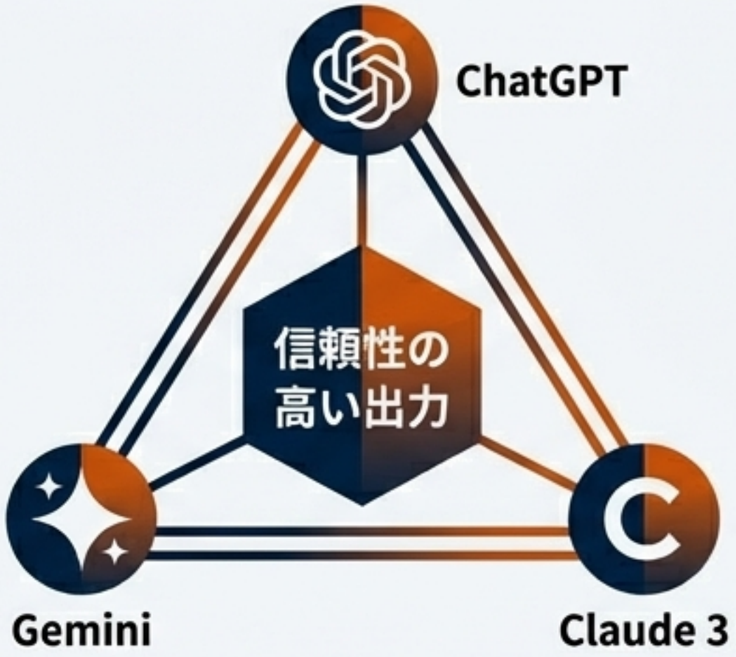
各AIが独自のバイアスやエラーの可能性を持つ以上、単一のAIに倫理判断を依存させることは重大なリスクを伴う。これは、重要なアプリケーションにおいて許容できない不確実性を生み出す。

NotebookLM

# 単一の神託から、三賢人の合議へ

**Before:** 単一の判断（Single Judgment）

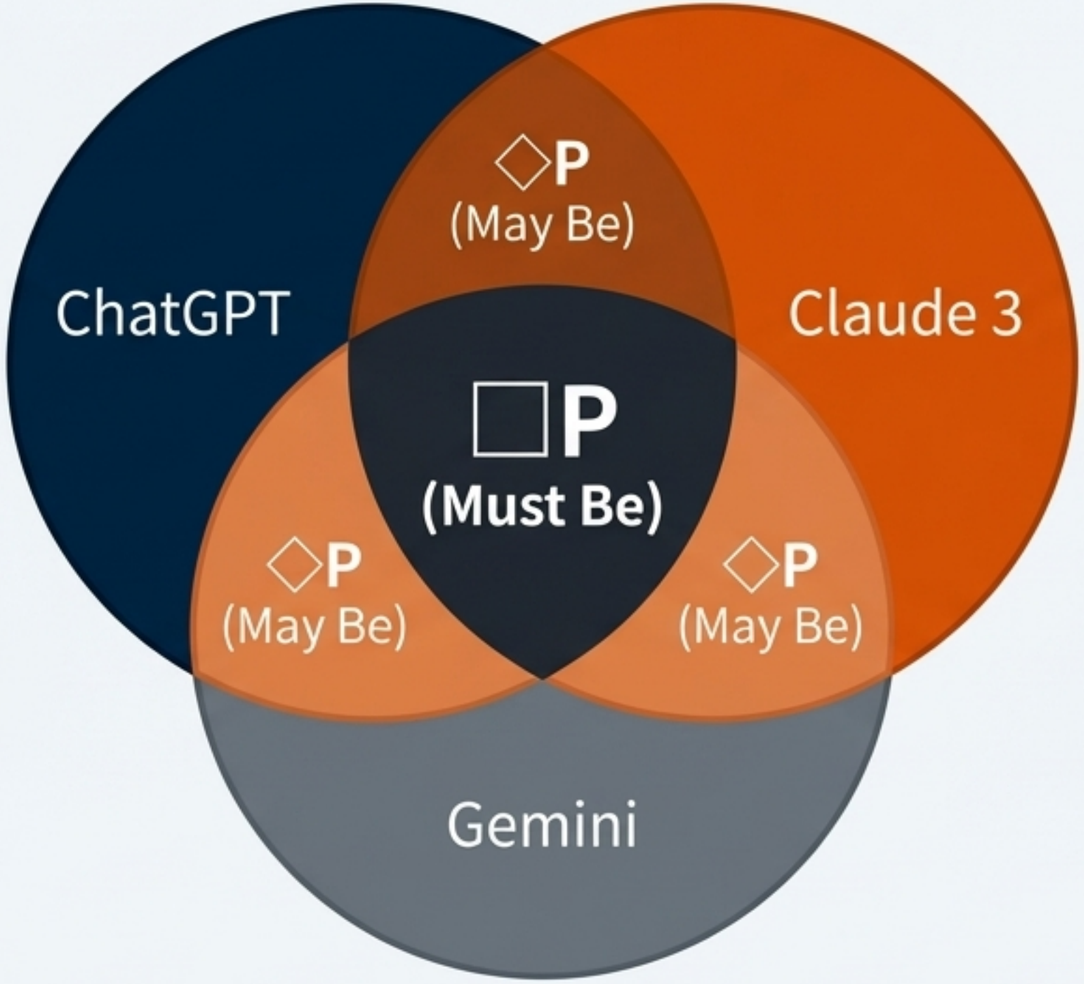
**After:** 協働による合議（Collaborative Consensus）




- 1つのAIを選ぶのではなく、ChatGPT、Gemini、Claude 3という主要3モデルの集合的な出力を活用する。
- このアプローチは、開発元（OpenAI, Google等）から独立した、中立的な立場から行われる。

NotebookLM

# 合意の論理学：様相論理による確信度の定義

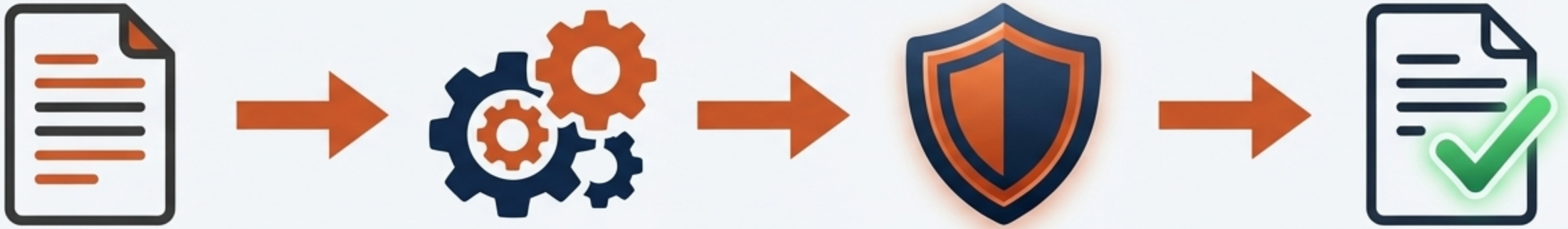



- **3 AIモデルが合意：**
  結論は **Must Be（□P）** である。
  （「～に違いない」 / 必然性）

- **2 AIモデル以上が合意：**
  結論は **May Be（◇P）** である。
  （「～かもしれない」 / 可能性）

NotebookLM

# AI倫理処理システムの全体像



**入力**
**(Query/Content)**

クエリが3つのAIに
同時送信される

**合意エンジン**
**(Consensus Engine)**

共通回答(P)を抽出
(Python, コサイン類似度)

**倫理フィルター**
**(Ethical Filter)**

共通回答(P)を
倫理基準(E)と照合

**出力**
**(Output)**

倫理的に検証された
最終回答

NotebookLM

# 倫理の「ものさし」を定義する

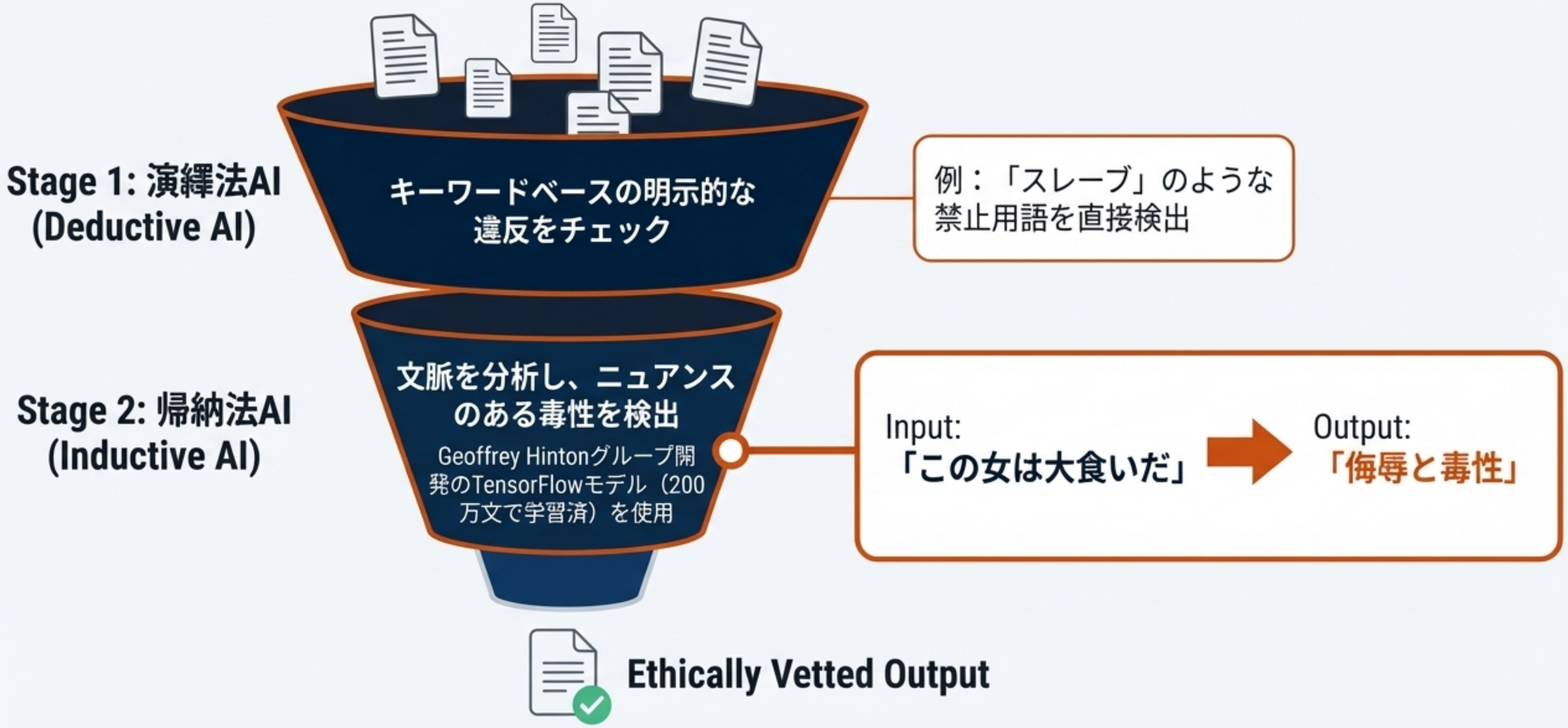
## 倫理基準Eの構成要素

- **日本語基準:**
  放送禁止用語リストを援用
- **英語基準:**
  教育禁止用語リストを援用

この基準が一度設定されると、システムは一貫して自動的に評価を適用する。

NotebookLM

# 演繹と帰納：二段階の倫理的検証フィルター



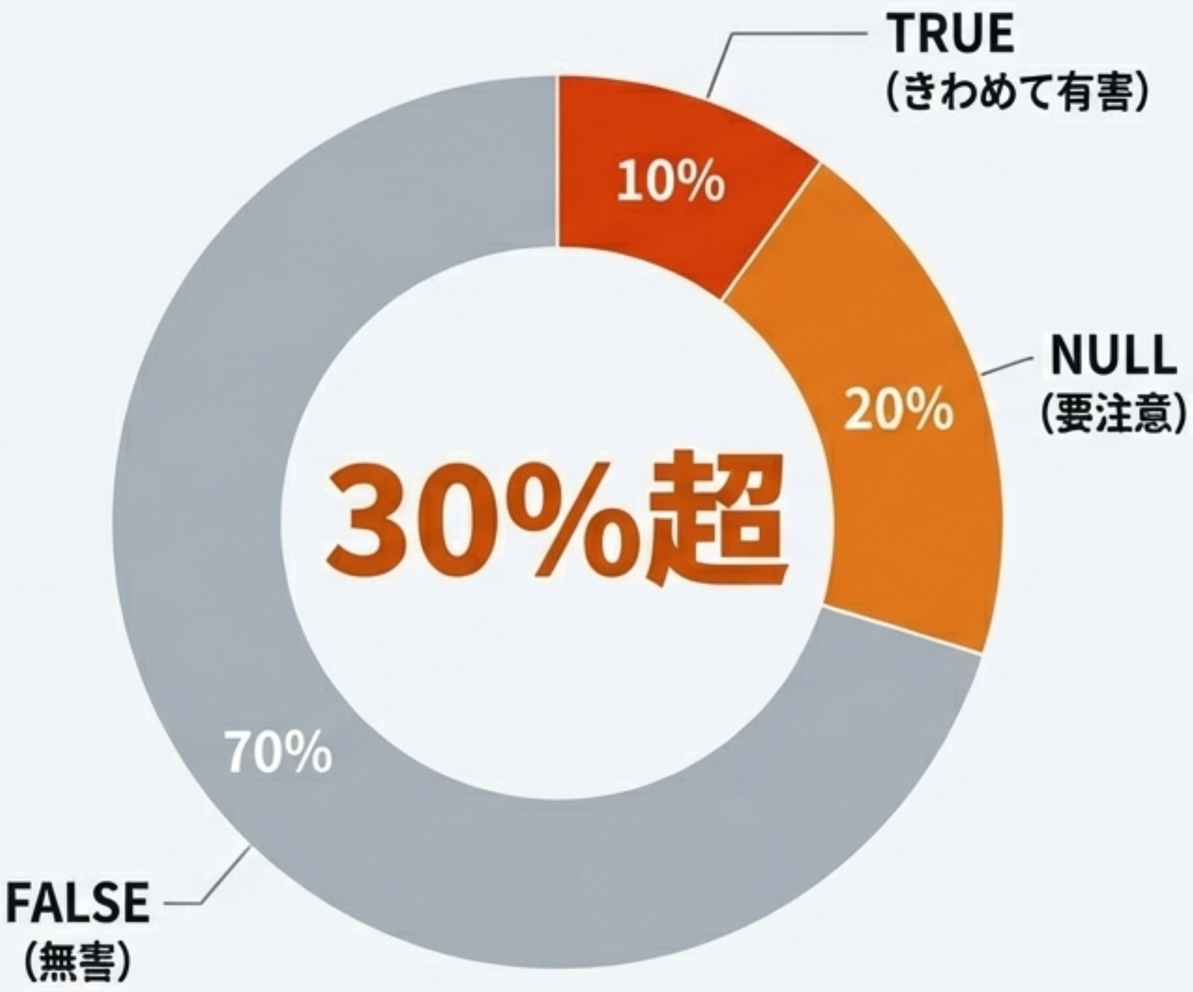

NotebookLM

# 理論から実践へ：プロトタイプシステムの検証

- **プロトタイプ名:** 「ＩｏＥ」AI 倫理チャットボット
- **検証目的:** 有害・倫理的に問題のある言語表現を実環境で検出する有効性をテストする。
- **使用技術:** 約200万件のデータで教師あり学習を行ったTensorFlow.jsモデル

NotebookLM

# GIGA端末の不適切発言を30%以上検出




**テスト対象:** GIGA端末の不適切入力文例37件

**分類結果**

- TRUE（きわめて有害）: 26件（10%）
- NULL（要注意）: 53件（20%）
- FALSE（無害）: 残り

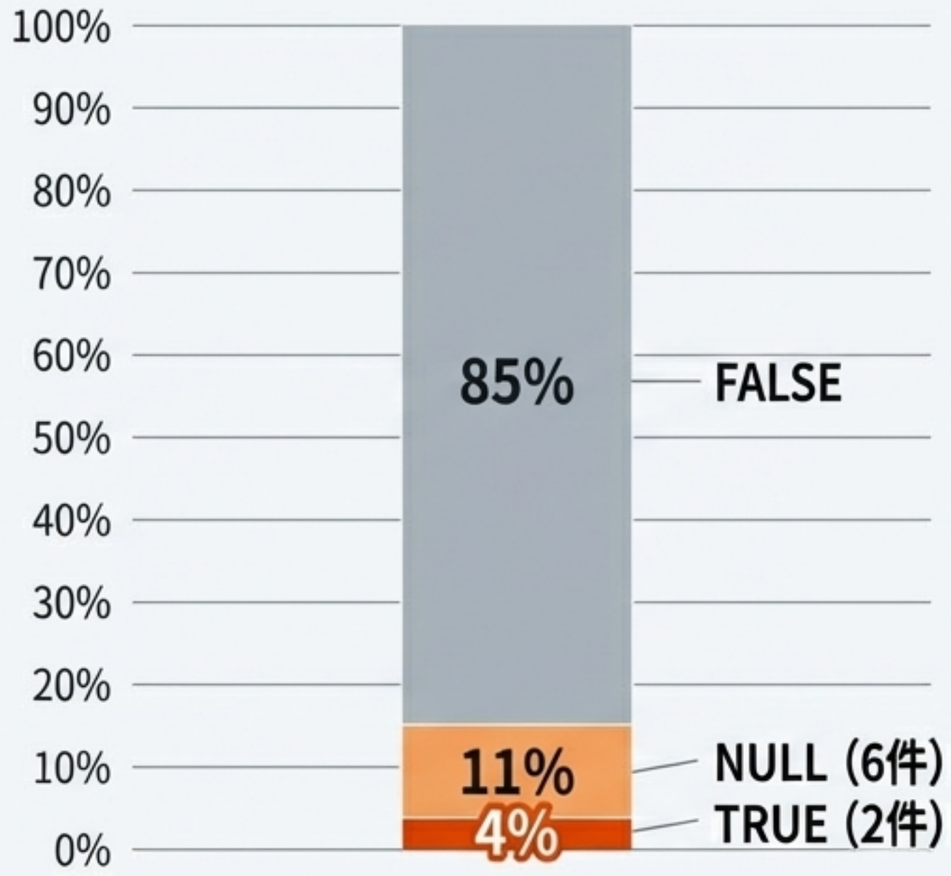
**結論**

「ＩｏＥ」チャットボットは、未定義の誹謗中傷を高い抽出率で検出し、その有効性を証明した。

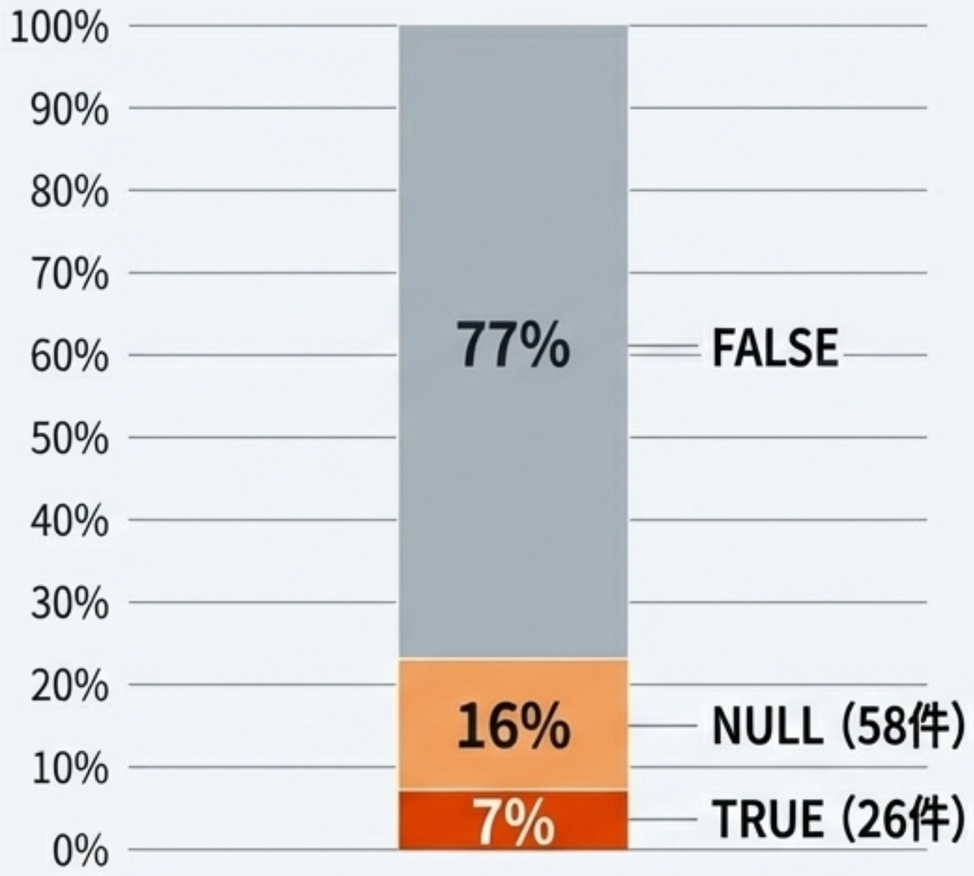
NotebookLM

# 既知の倫理基準リストに対する検出精度

## 英語 教育禁止用語（57文）

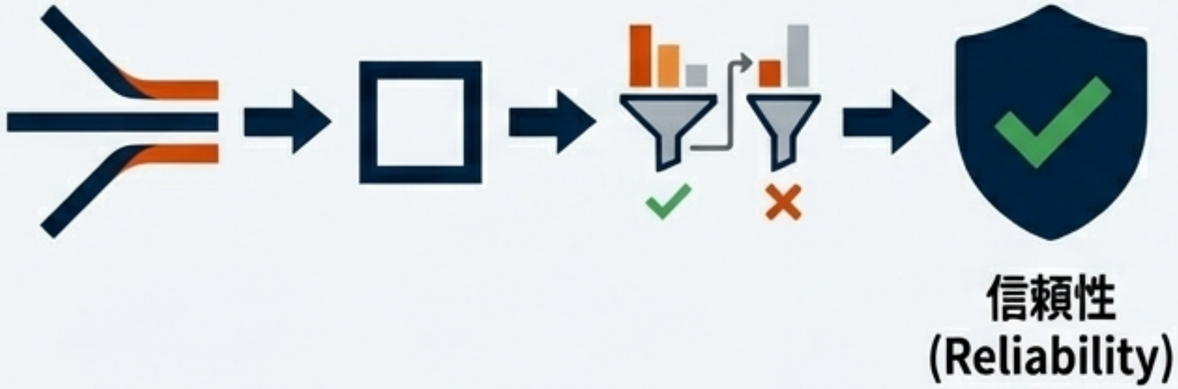



## 日本語 放送禁止用語（369文）




NotebookLM

# 集合知は、単独の知性を超える




- **問題**: 単一AIへの依存は、性能の拮抗と固有バイアスのため、倫理判断において本質的に信頼性が低い。
- **解決策**: 3つのAIによる協働フレームワークは、様相論理と二段階フィルターに基づき、客観的で信頼性の高い代替案を提供する。
- **実証**: 検証テストにより、本システムが明示的および潜在的なな倫理違反を効果的に特定し、責任あるAI実装を可能にすることがが証明された。

NotebookLM

# 次なる課題

マルチモーダル生成ＡＩのハルシネーションを防ぐにはどのようにしたら良いかを考察し、あなたの考えを８００字以内で説明しなさい。

NotebookLM