

# 大規模言語モデルを用いた オープンソースライセンスの権利・義務の推定

肥後研究室

B4 長崎智人



# OSSの利用拡大とオープンソースライセンス

- ・コスト削減のため、既存のプログラムの再利用が効果的
- ・近年、**OSS（オープンソースソフトウェア）**の利用が拡大
- ・OSSの利用には**オープンソースライセンス**の遵守が必須

# ライセンス遵守における課題

- ・ **詳細かつ複雑**なライセンス
- ・ **解説などの二次情報が不十分な**ライセンス
- ・ 複数のソフトウェアを組み合わせることによる  
ライセンスの衝突の可能性（**互換性**の問題）

# ライセンス分析手法の必要性

- ・ ライセンスの遵守を徹底するには  
**多くの時間と労力が必要**
- ・ **開発初期段階でライセンス違反の可能性を  
検出する必要性<sup>[1]</sup>**



**ライセンス分析・遵守を支援する手段の導入**

[1] 東裕之輔. コンテナ開発におけるOSSの法的リスク特定自動化に関する研究, 博士論文, 和歌山大学, 2023.

# 定義：条項，権利条項，義務条項

条項：ライセンスが規定する行為

権利条項：

許可（“**Can**”）・禁止（“**Cannot**”）の対象となる，  
利用者の権利に関わる条項

例：商用利用（Commercial Use），配布（Distribute）など

義務条項：

義務付け（“**Must**”）の対象となる条項

例：著作権表示の記載（“Include Copyright”），  
変更箇所の明記（“State Changes”）など

# tl;drLegal<sup>[2]</sup>における権利・義務の判定

ライセンスに明記された**権利**（**Can, Cannot**）,  
**義務**（**Must**）のリストを公開

tl;drLegal による権利・義務のリスト： Apache License 2.0 <sup>[2]</sup>

| Can               |
|-------------------|
| Commercial Use    |
| Modify            |
| Distribute        |
| Sublicense        |
| Place Warranty    |
| Private Use       |
| Use Patent Claims |

| Cannot        |
|---------------|
| Hold Liable   |
| Use Trademark |

| Must              |
|-------------------|
| Include Copyright |
| Include License   |
| State Changes     |
| Include Notice    |

[2] FOSSA. tl;drlegal. <https://www.tldrlegal.com>. 2025年1月30日参照.

# ライセンス分析手法の先行研究 LiDetector<sup>[3]</sup>

- ・ 自然言語処理を利用した，ライセンスの非互換性を検出する手法
- ・ 対象のライセンスに対し**権利・義務を推定**
- ・ tldrLegalと同様のリストを生成
- ・ 従来の自然言語処理手法より高い精度を計測

(再現率 69.70%,  
適合率 85.88%,  
F1スコア 76.95%)

| Can               |
|-------------------|
| Commercial Use    |
| Modify            |
| Distribute        |
| Sublicense        |
| Place Warranty    |
| Private Use       |
| Use Patent Claims |

| Cannot        |
|---------------|
| Hold Liable   |
| Use Trademark |

| Must              |
|-------------------|
| Include Copyright |
| Include License   |
| State Changes     |
| Include Notice    |

[3] Sihan Xu, Ya Gao, Lingling Fan, Zheli Liu, Yang Liu, and Hua Ji. Lidetector: License incompatibility detection for open source software. ACM Trans. Softw. Eng. Methodol., Vol. 32, No. 1, feb 2023.

# 大規模言語モデル（LLM）

- ・ディープラーニング技術に基づく高度な自然言語処理モデル
- ・ **GPT, Llama**などのモデルが代表例
- ・ 様々な自然言語処理タスクを実行可能  
⇒ **権利・義務の推定を含む，ライセンス分析にもLLMを適用できる可能性**
- ・ LLMを用いたライセンス分析の研究は進んでいない



# 実験の概要

目的：

**LLMによるライセンスの権利・義務の推定の精度を調べる。**

方法：

LLMにライセンス文書と質問を与え、  
対象のライセンスの

**権利**（“Can”，“Cannot”）・**義務**（“Must”）を  
推定し，その精度を評価

# 推定方法：権利条項

## “Distribute”（配布）の例：

1. 「ライセンスは**配布を明示的に許可している**か？」をLLMに問う

⇒“Yes”なら“Distribute”は“**Can**”であると判定

2. 「ライセンスは**配布を明示的に禁止している**か？」をLLMに問う

⇒“Yes”なら“Distribute”は“**Cannot**”であると判定

# 推定方法：義務条項

“**Include Copyright**”（著作権表示の記載）の例：

「ライセンスは**著作権表示の記載を明示的に義務付けているか？**」をLLMに問う

⇒“Yes”なら“**Include Copyright**”は  
“**Must**”であると判定

# 対象の条項

tl;drLegalで定義されている23の条項

11の権利条項 (“Can”, “Cannot”を判定)

12の義務条項 (“Must”を判定)

⇒ライセンスごとに34個の判定項目

# 評価用データセット

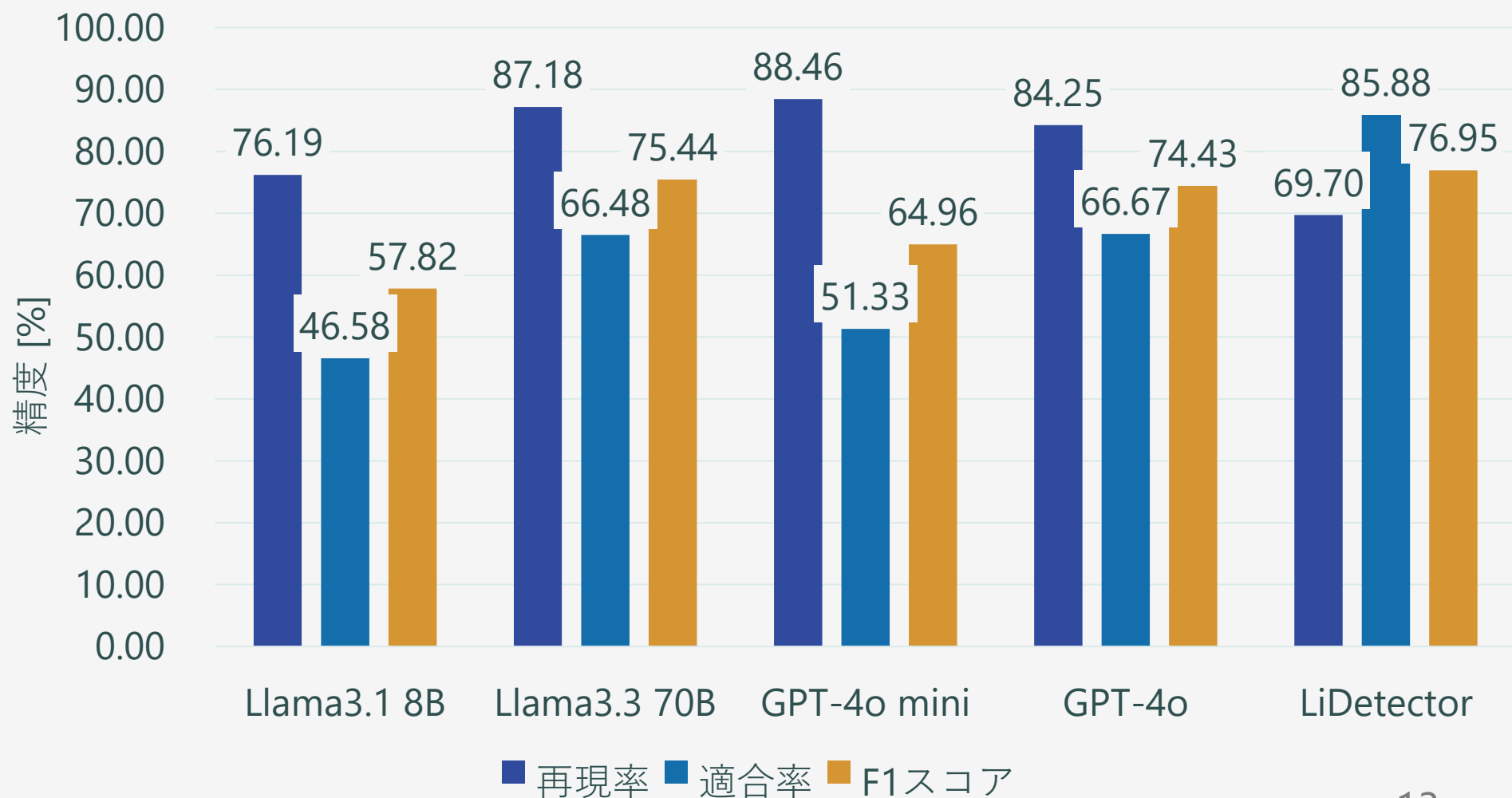
tl;drLegal で OSI-Approved とタグ付けされた  
58のライセンス

- ・ OSI-Approved : OSIにより認可されたOSS
- ・ すでに権利・義務の判定がされており、  
正解データとして扱える

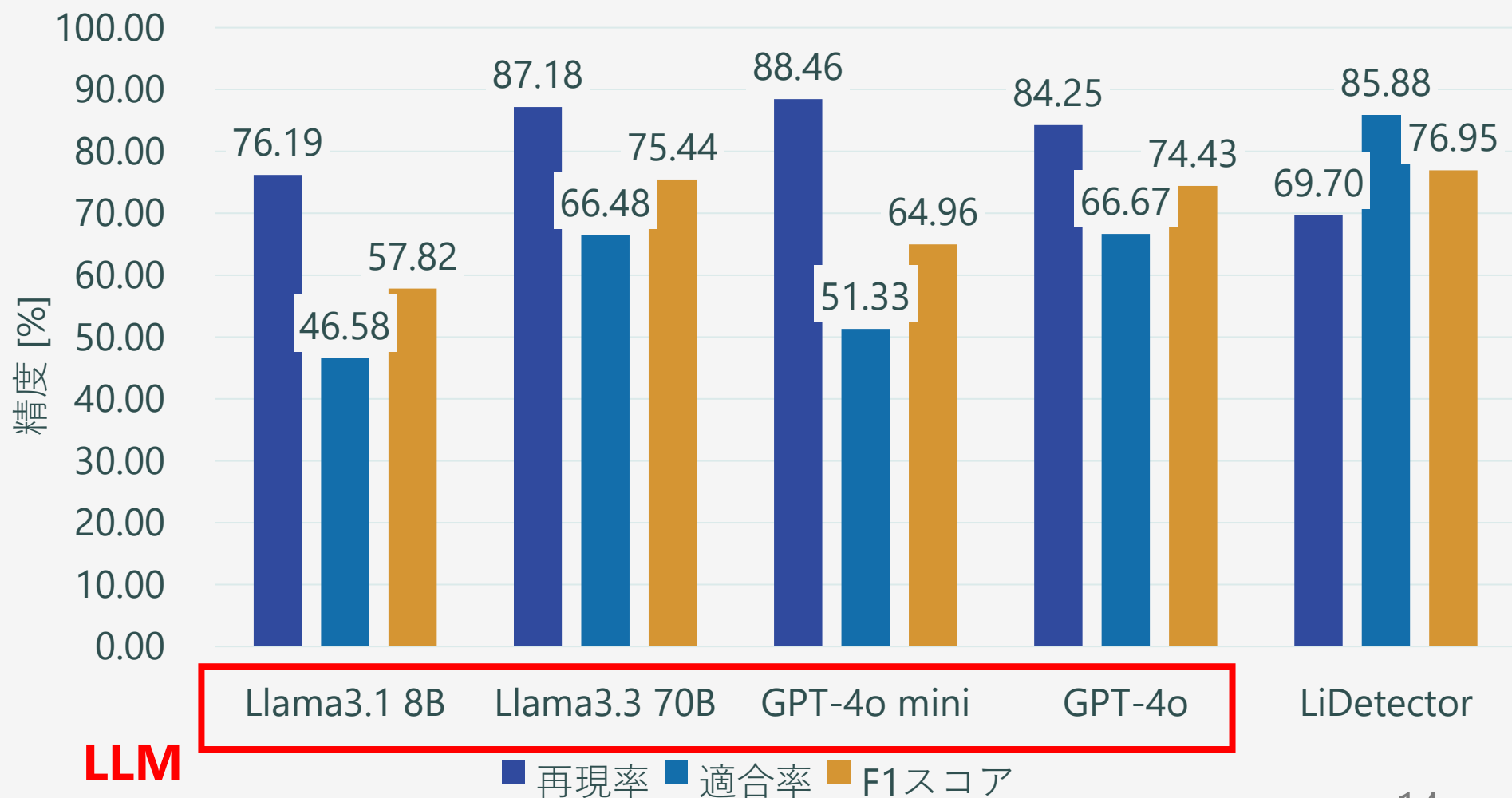
文書中の固有名詞を，意味を損なわないように  
別の文字列に置き換えて匿名化

- ・ LLMの学習データにライセンスの二次情報が  
含まれている可能性があるため

# 実験結果

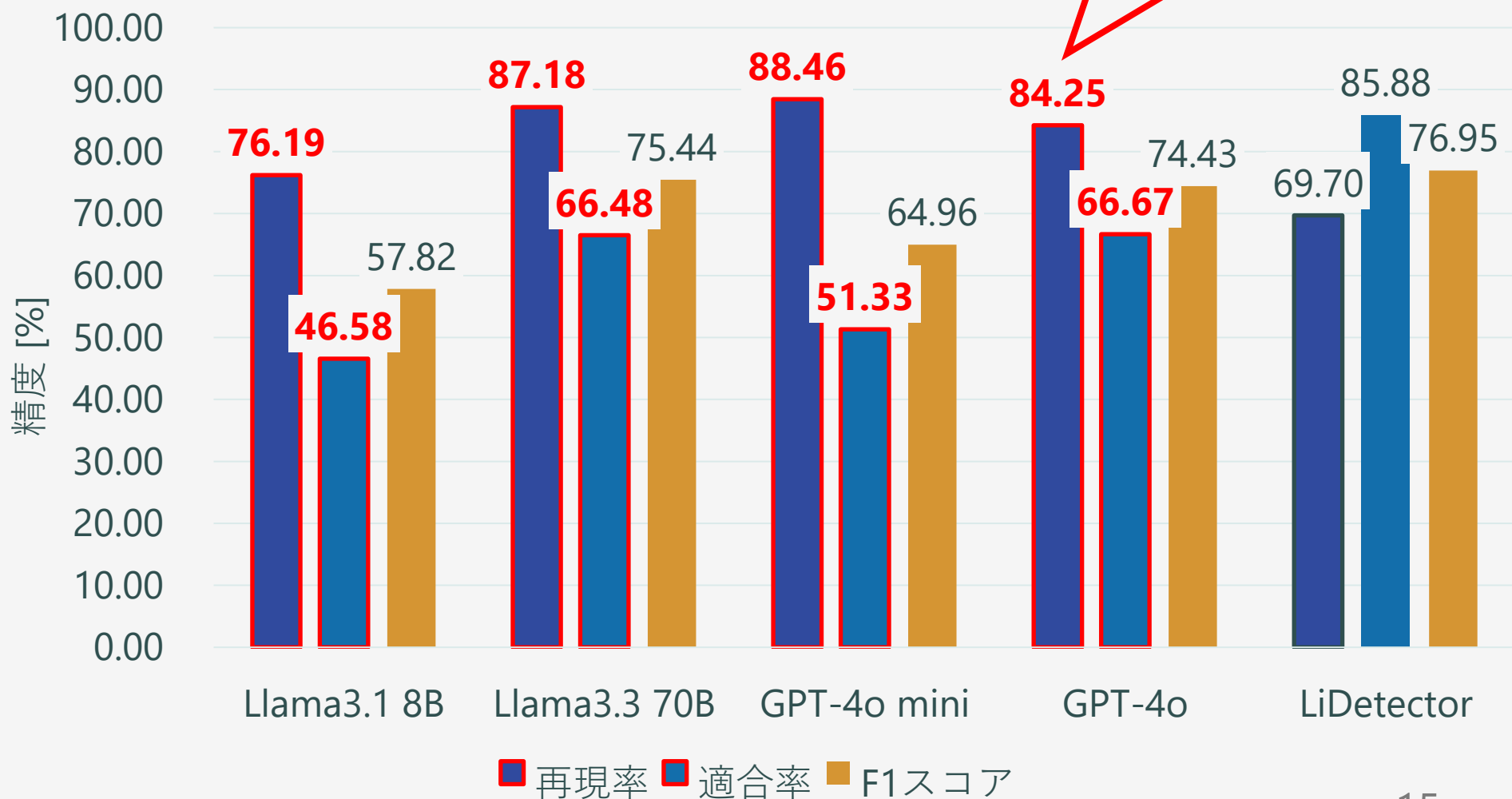


# 実験結果



# 実験結果

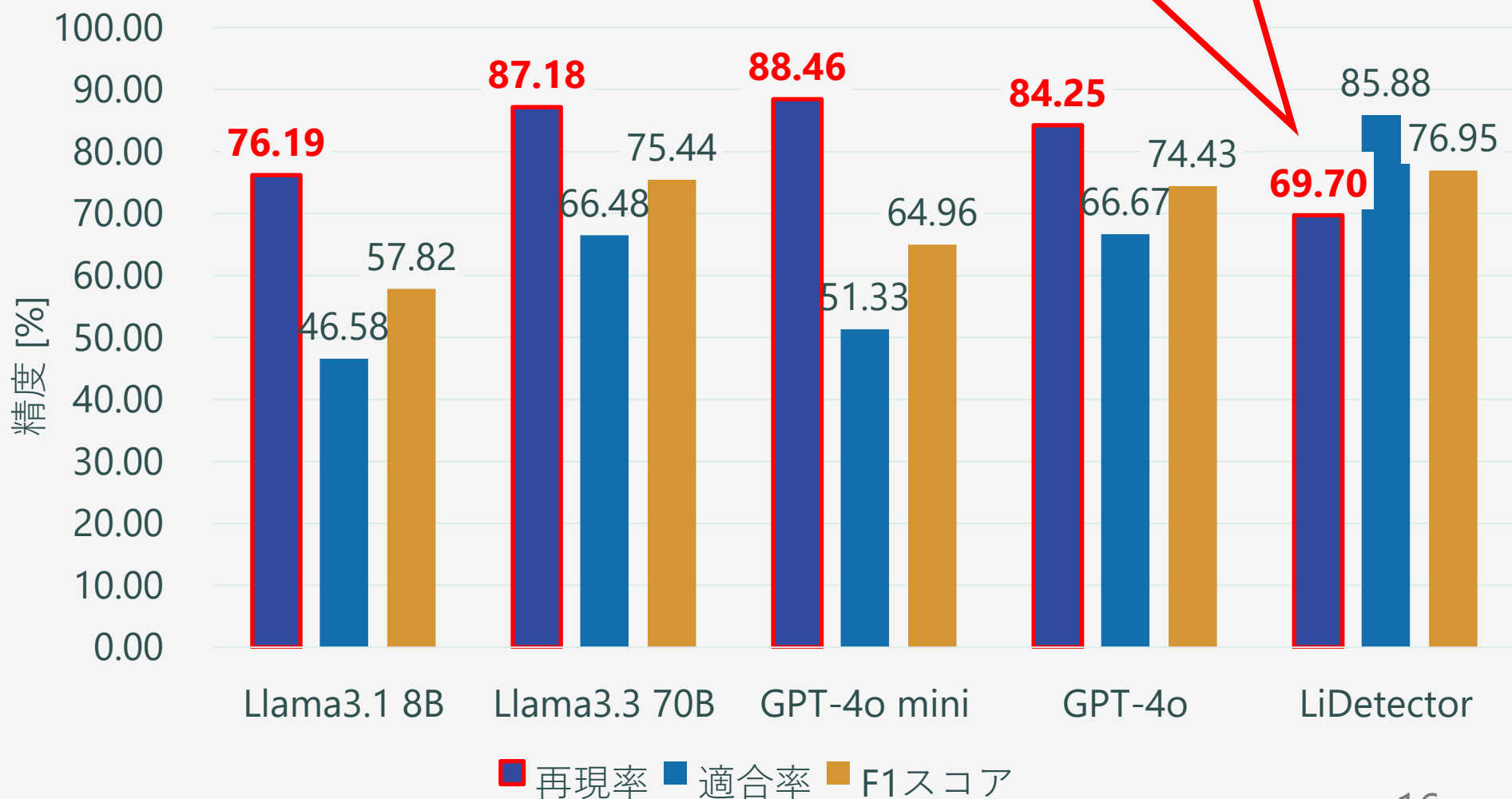
LLMはいずれも  
再現率 > 適合率





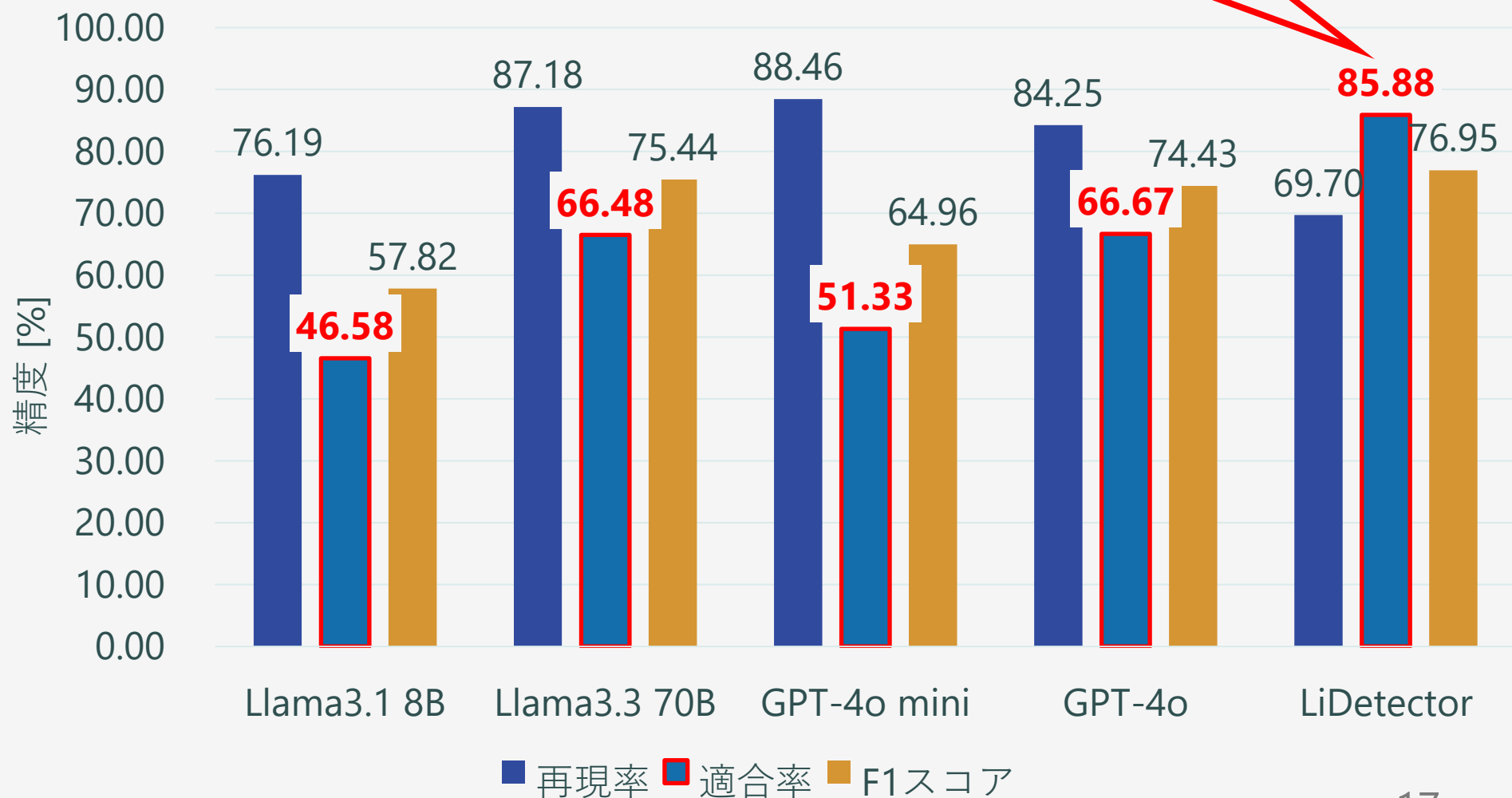
# 実験結果

再現率は  
**LLM > LiDetector**



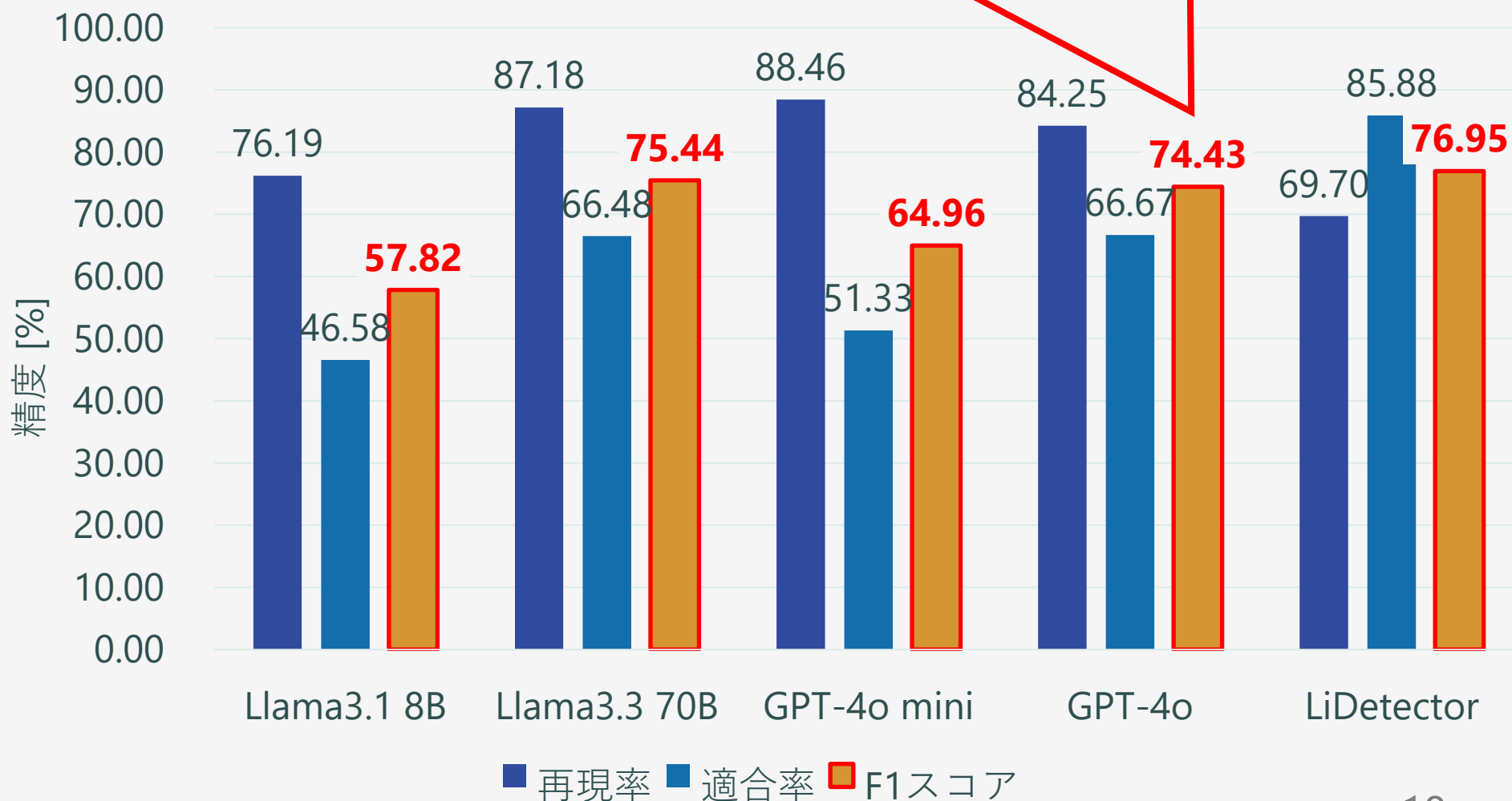
# 実験結果

適合率は  
**LLM < LiDetector**



# 実験結果

F1スコアは  
**LLM < LiDetector** だが  
**Llama3.3 70B, GPT-4o** が近い値に



# 分析 1：ライセンス文書の長さ と推定精度の関係

「ライセンス文書が長いほど推定精度が低い」  
と仮定し、トークン数と精度の相関係数を計算

・ 4つのモデル・3つの評価指標ごとに計算



負の相関は確認できなかった

・  $-0.3613$  から  $0.3489$  の範囲であり、負の傾向は無い

## 分析 2：判定項目ごとの推定精度の差

「判定項目ごとに正答率に差がある」  
と仮定し検定を実施



**有意な差があることを確認**

- ・ 得意な判定項目，不得意な判定項目が存在
- ・ 推定精度の低い判定項目は  
条項の定義の不十分さなどが考えられる

# 考察 1：ライセンスの権利・義務の推定 へのLLMの利用可能性

## ライセンスの権利・義務の判定にLLMの利用が 有用であるケースが存在

- ・ 再現率が求められる禁止・義務の推定
- ・ 精度の高い一部の判定項目

# 考察 2 : 推定精度の向上方法

## プロンプトの改善による精度向上の可能性

- ・ 各条項の厳密な定義を与える（専門的な知識が必要）
- ・ Chain-of-Thought<sup>[5]</sup>等の技術の活用

## ファインチューニングによる精度向上の可能性

- ・ ただし，学習データの収集が困難

[5] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Minh Pham, Gerson C. Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncarencu, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexan der Miserlis Hoyle, and Philip Resnik. The prompt report: A systematic survey of prompting techniques. ArXiv, Vol. abs/2407.21783, , 2024.

# まとめ

- ・ LLMを利用してライセンスの権利・義務を推定
- ・ LLMによる推定は，再現率で先行研究を上回ったが，適合率で下回った
- ・ LLMの利用が有用であるケースが存在する
- ・ いくつかの方法により，精度の向上が見込める