JaSST'26 Tokyo

論文セッション H3-3

LLMのテスト分析・テスト設計能力を測定するためのベンチマーク手法

株式会社ベリサーブ　研究開発部

吉川 努、谷﨑 浩一、上野 彩子

東京都市大学

増田 聡

2026年3月20日（祝・金）

内容

LLMはテスト分析・テスト設計に使える？

LLMの進化がすごい

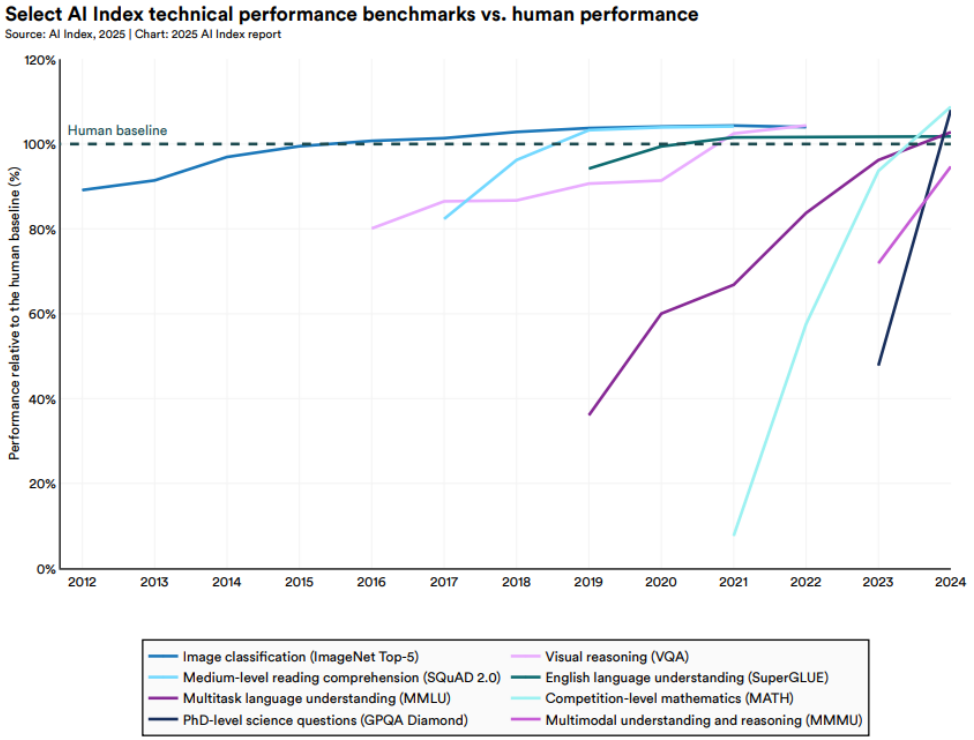


Figure 2.1.33

AIはこれまで以上に速く新たなベンチマークを習得する。2023年、AI研究者らはMMMU、GPQA、SWE-benchなど、高度化するAIシステムの限界をテストする複数の新たな挑戦的なベンチマークを導入した。2024年までに、これらのベンチマークにおけるAIの性能は目覚ましい向上を見せ、MMMUとGPQAではそれぞれ18.8ポイントと48.9ポイントの改善を達成した。SWE-benchでは、2023年にAIシステムが解決できたコーディング問題はわずか4.4%だったが、この数値は2024年には71.7%に跳ね上がった。

- 各種ベンチマークで軒並み人を超える
  - LLMの進化は凄まじい
  - ベンチマークも、短期間で“更新”される
  - テスト活動も「AIでできそう」に見える

テスト分析・テスト設計で使えそう！

出典：Stanford HAI, Artificial Intelligence Index Report 2025, Chapter 2: Technical Performance, Fig. 2.1.33, p.13.
（PDF: < https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter2_final.pdf >）

テスト分析・テスト設計に対するLLMの能力

- **LLMは60点？**
  - ▸ 何の指標で？どんな条件の？どんな手順で？

「使える/使えない」ではなく、**どこまでなら使えるのか**を定量的に測る

既存のベンチマーク

既存：単体テスト～結合テスト

- 単体テスト～結合テストのテストコード
- カバレッジ／欠陥検出力[*1]

[*1]ミューテーションスコア、実欠陥の再現

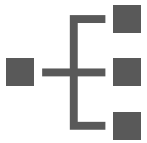
本研究：システムテスト

- システムテストにおける作成物
- テスト分析・テスト設計の能力

現時点の能力を知る、だけではなく、今後の**能力の進化も追える**ようにする

ベンチマークの設計

テスト観点導出力と欠陥識別力で測る



①テスト観点導出力（単一）

テストベースから
「何をテストするか」
を識別できるか

②テスト観点導出力（組み合わせ）

テスト観点を組み合わせて
構造化できるか

③欠陥識別力

生成したテストケースで
どの欠陥を検出し得るか

テスト観点導出のイメージ

仕様例

6.1.3.1.　キーワード検索

入力したキーワードに合致する情報を持つ社員を検索する。

① 検索方法

- 一致判定方法

  一致判定方法について 2 種類の方法から選択が可能である。それぞれの検索方法は以下の通り。

  なお、キーワードは” ”(半角スペース)をデリミタとして複数設定することが可能である。

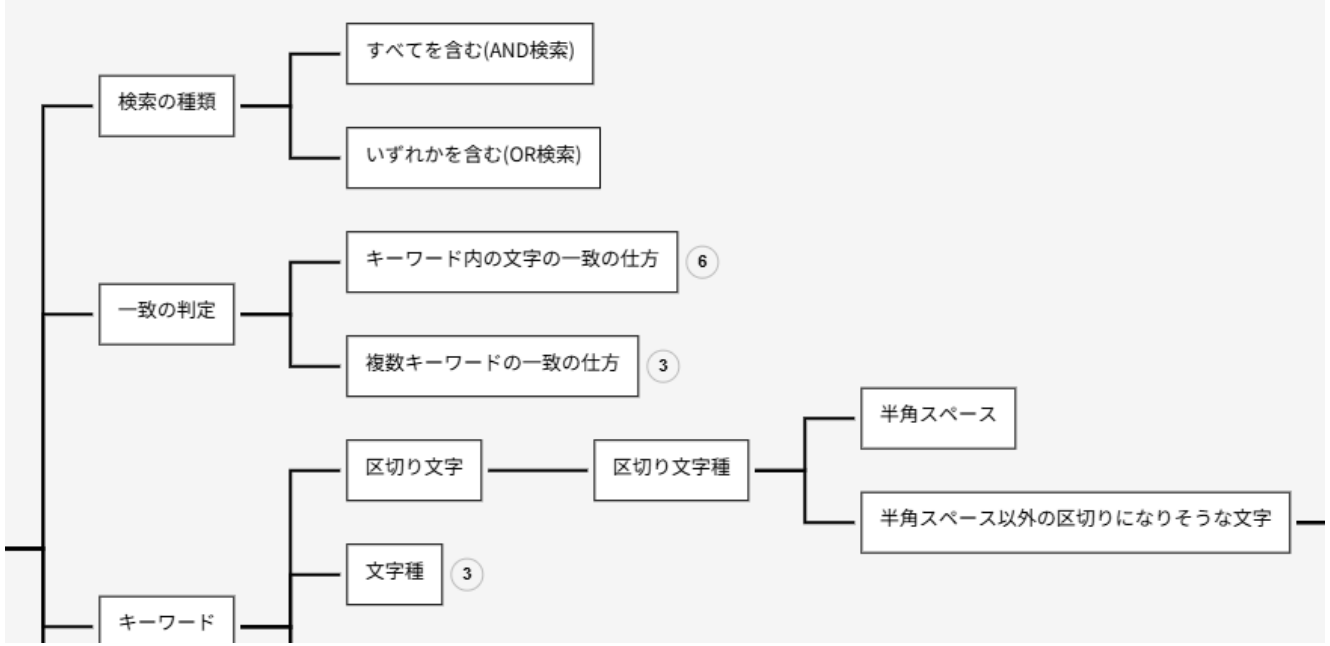
  A) すべてを含む

  設定したキーワードを検索範囲内にすべて持つ社員が検索される。

  B) いずれかを含む

  設定したキーワードを検索範囲内にいずれか持つ社員が検索される。

テスト観点例



欠陥導出のイメージ

テストケース例

テスト設計技法	カバレッジ基準	テストケースID	テストケースの意味	事前条件	入力	期待値
同値分割	同値網羅	1.1.1.1	完全一致するキーワードで検索する	検索画面表示 スキル「XXXYYY」を持つ社員が登録されている	キーワードで検索欄に「XXXYYY」を入力 「検索」押下	当該社員が検索結果画面に表示されること
同値分割	同値網羅	1.1.1.2	前方一致するキーワードで検索する	検索画面表示 スキル「XXXYYY」を持つ社員が登録されている	キーワードで検索欄に「XXX」を入力 「検索」押下	当該社員が検索結果画面に表示されること
同値分割	同値網羅	1.1.1.3	中間一致するキーワードで検索する	検索画面表示 スキル「XXXYYY」を持つ社員が登録	キーワードで検索欄に「XXYY」を入力	当該社員が検索結果画面に表示されること

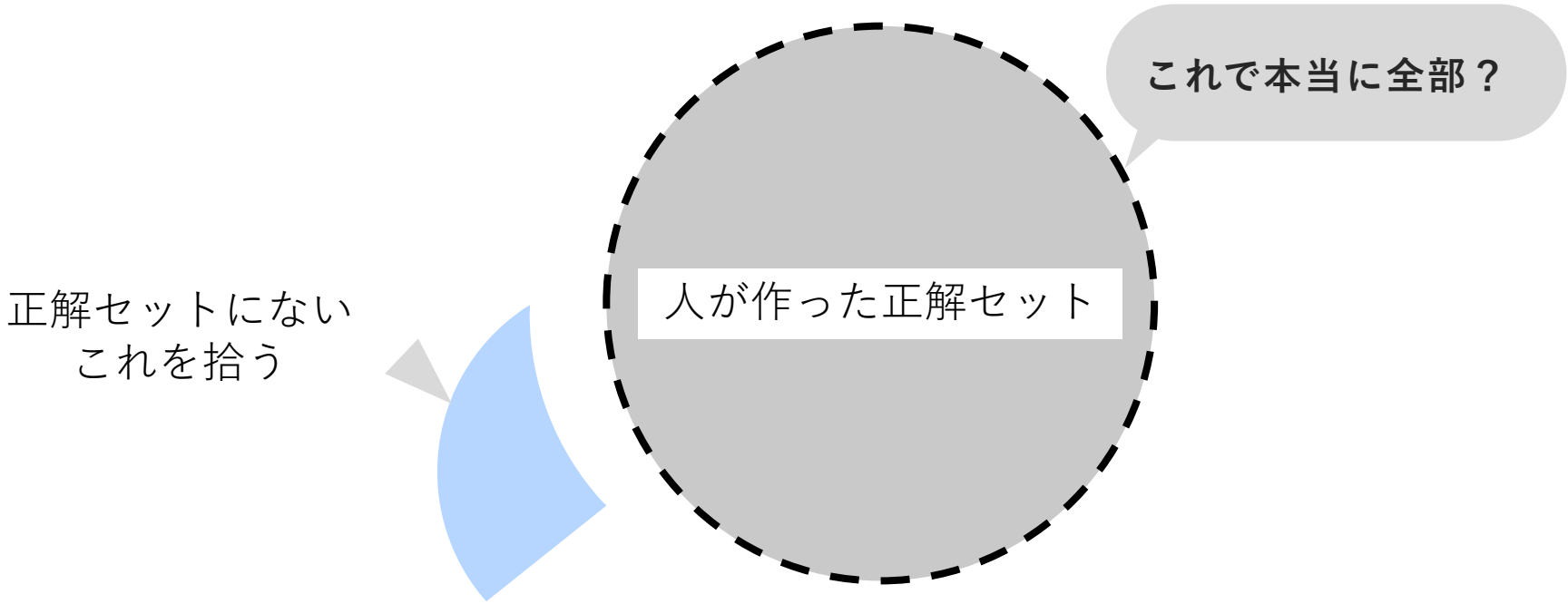
検出可能な欠陥例

検出可能な欠陥
ANDのキーワード検索を実行した際、キーワードに完全一致する社員が検索結果画面に表示されない
ANDのキーワード検索を実行した際、キーワードに前方一致する社員が検索結果画面に表示されない
ANDのキーワード検索を実行した際、キーワードに中間一致する社

正解セットにない“正解”を拾う

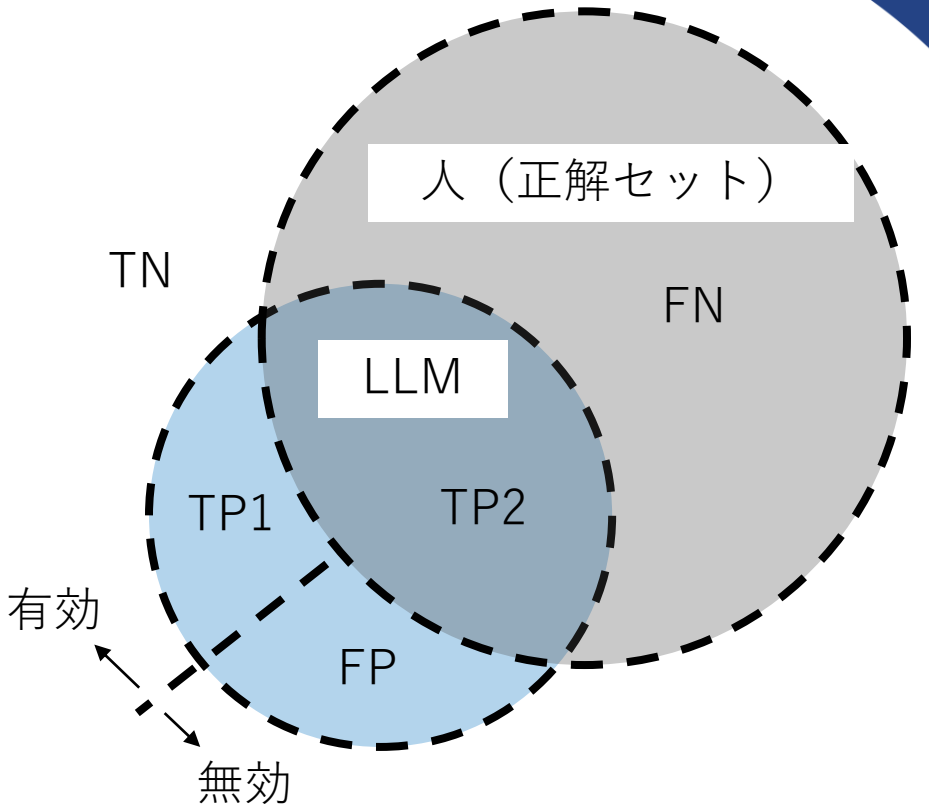
- 人のテストケースも「1つの解」に過ぎない
  - 正解セットになくても“正解”（有効）を拾う必要がある

→予測と実際を突き合わせて整理する混同行列が使える



混同行列で整理

- 出力を 有効／無効の二値分類 で判定する
- LLMの出力を“予測”、その判定を“実際”として 混同行列で集計する
  - 正解セットにない出力でも、実務上 有効なら TP（=TP1）とする
  - 重複・誤り・意味が薄いものはFP



	実際は正（有効）	実際は負（無効）
正（有効）と予測	TP（真陽性）	FP（偽陽性）
負（無効）と予測	FN（偽陰性）	TN（真陰性）

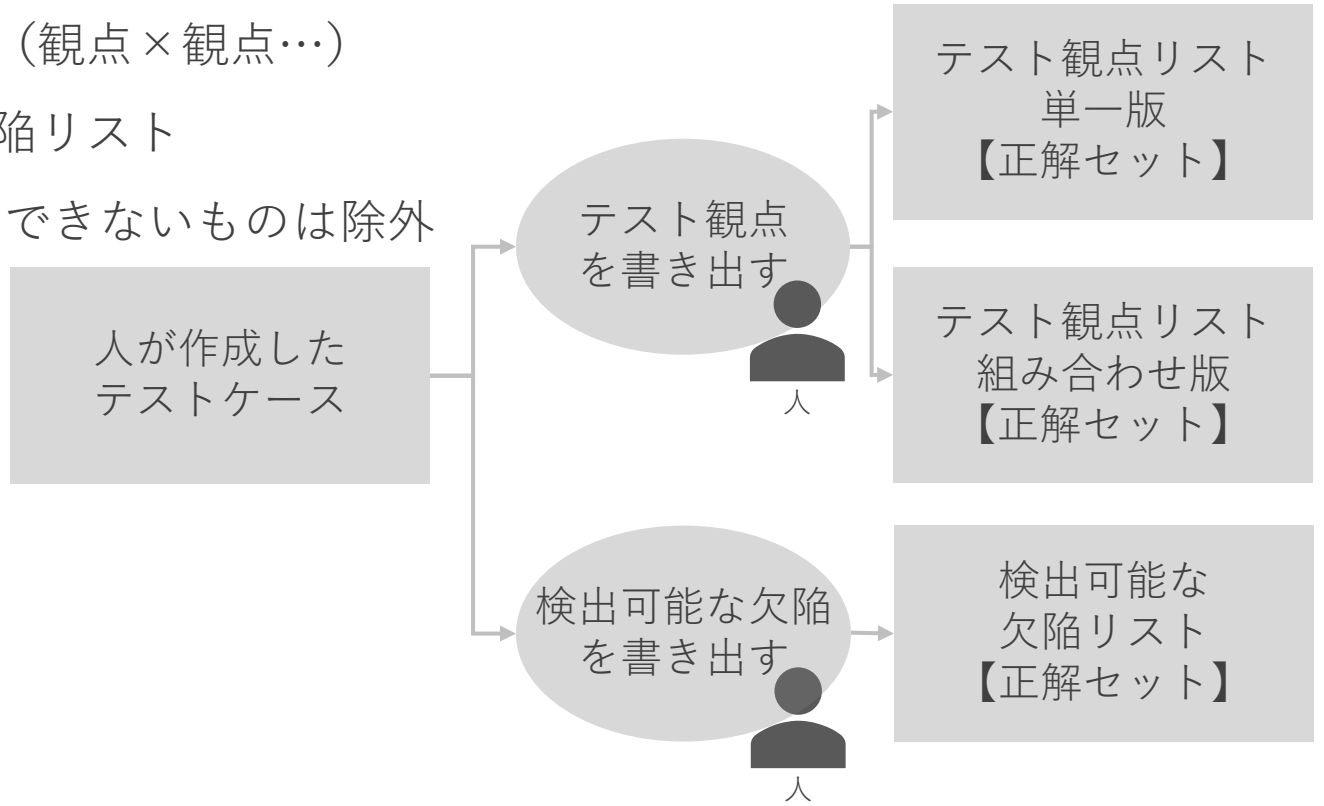
正と予測　→　LLMが出力したテストケースで検出可能な欠陥
負と予測　→　LLMが出力したテストケースで検出できない欠陥

実際は正　→　検出すべき欠陥として有効なもの
実際は負　→　検出すべき欠陥として無効なもの（重複・誤りなど）

正解セットの構築

- 人が作成したテストケース群を起点に、3種類の“正解セット”を作る
  - 抽出するもの
    - 単一のテスト観点リスト
    - 組み合わせテスト観点リスト（観点×観点…）
    - テストケースで検出可能な欠陥リスト

※ 重複やテストベースから導出できないものは除外



指標

正解率　(TP+TN) / (TP+FP+FN+TN)

全体の当たり具合

「正」も「負」もどのくらい有効なものを出せたか

適合率　TP / (TP+FP)

無駄なものを出していないか

「正」と判定したものがどのくらい信用できるか

再現率　TP / (TP+FN)

必要なものをどれだけ拾えたか

正をどれだけ出せているか

F値　2×(適合率×再現率)/(適合率+再現率)

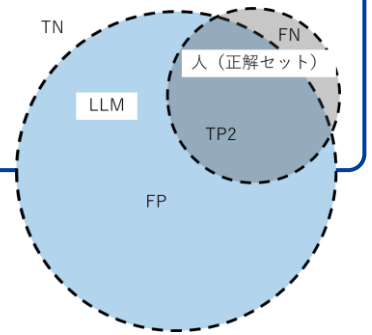
適合率×再現率のバランス（総合）

適合率と再現率のバランス

例：

大量にテストケースを作って、「正」が多く出せても、無効なものも大量に作ってしまうと

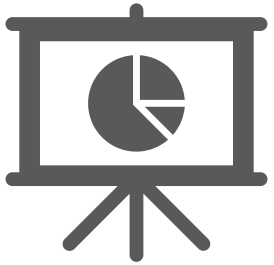
再現率は高いが、適合率は低くなってしまう



大量に作ってしまったイメージ

検証方法

適用したシステム：2つのWebアプリ



対象①テスト管理ツール

- Webアプリ
- テストベース：操作マニュアル
- 対象機能
  - テストスイートを作成する機能
  - テストスイートを設定する機能
- テストケース　50件
- 正解セット：
  - 単一テスト観点　76件
  - 組み合わせテスト観点　57件
  - 欠陥　69件

対象②社員情報検索システム

- Webアプリ
- テストベース：設計書
- 対象機能
  - キーワード検索機能

- テストケース　56件
- 正解セット：
  - 単一テスト観点　51件
  - 組み合わせテスト観点　52件
  - 欠陥　39件

対象モデルとプロンプトの方針

当時の最新

- 対象としたLLMのモデル
  - GPT-5　　社内向けアプリケーション
  - GPT-5.1　　ChatGPT
  - GPT-5.1 Thinking　　ChatGPT
  - GPT-5.1 Pro　　ChatGPT
- プロンプト方針
  - プロンプトエンジニアリングの工夫に依存させないようにシンプルに
  - テストベースを与え、システムテストのテストケースを生成

実際のプロンプト

テスト管理ツール

あなたはソフトウェア開発プロジェクトにおいて、システムテストを担当している優秀なテストエンジニアです。

このプロジェクトでは、ソフトウェアテストにおけるテストスイートを管理するテスト管理ツールを開発しています。

添付ファイルの情報からシステムテストのテストケースを表形式で作成してください。

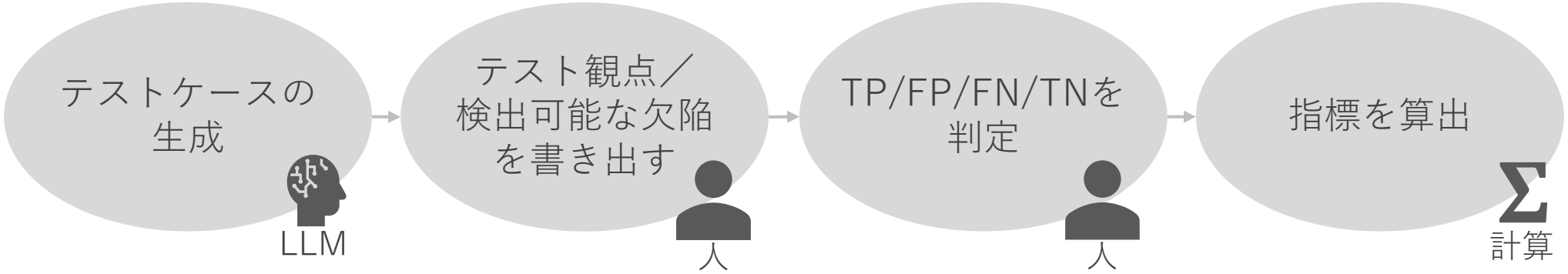
社員情報検索システム

あなたはソフトウェア開発プロジェクトにおいて、システムテストを担当している優秀なテストエンジニアです。

このプロジェクトでは、社員データベースを開発しています。
以下の情報からシステムテストのテストケースを表形式で作成してください。

仕様
（省略。ここにテキストベースで仕様を記載する）

評価手順



テストベースに対して
プロンプトを与えて
LLMにテストケースを
生成させる

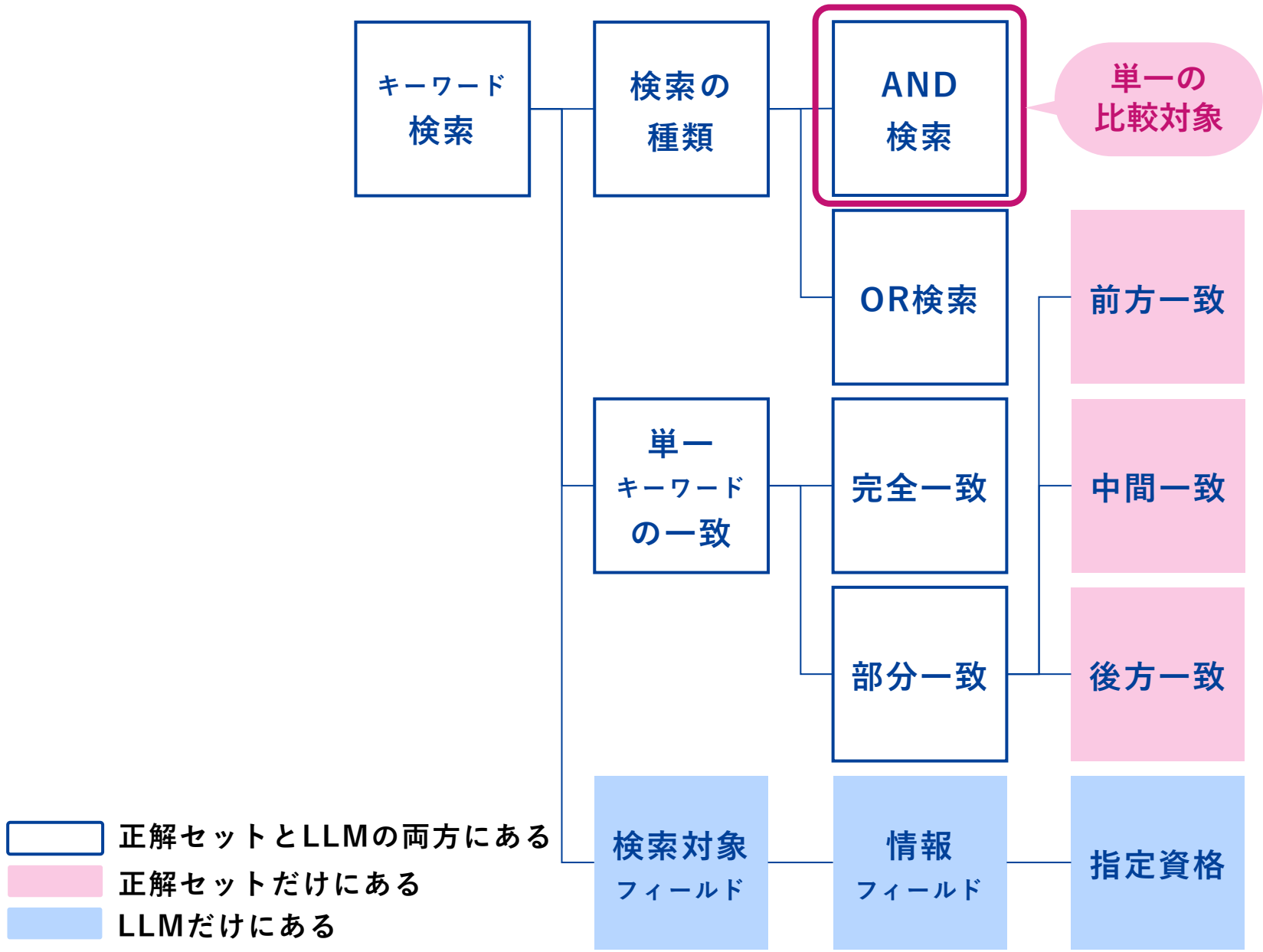
生成テストケースから、
テスト観点(単一/組み合
わせ)と検出可能欠陥を
抽出

正解セットと比較し、
TP/FP/FN/TNを判定

正解率・適合率・
再現率・F値を算出

LLMの出力（LLMのモデル・バージョンごとに実施）

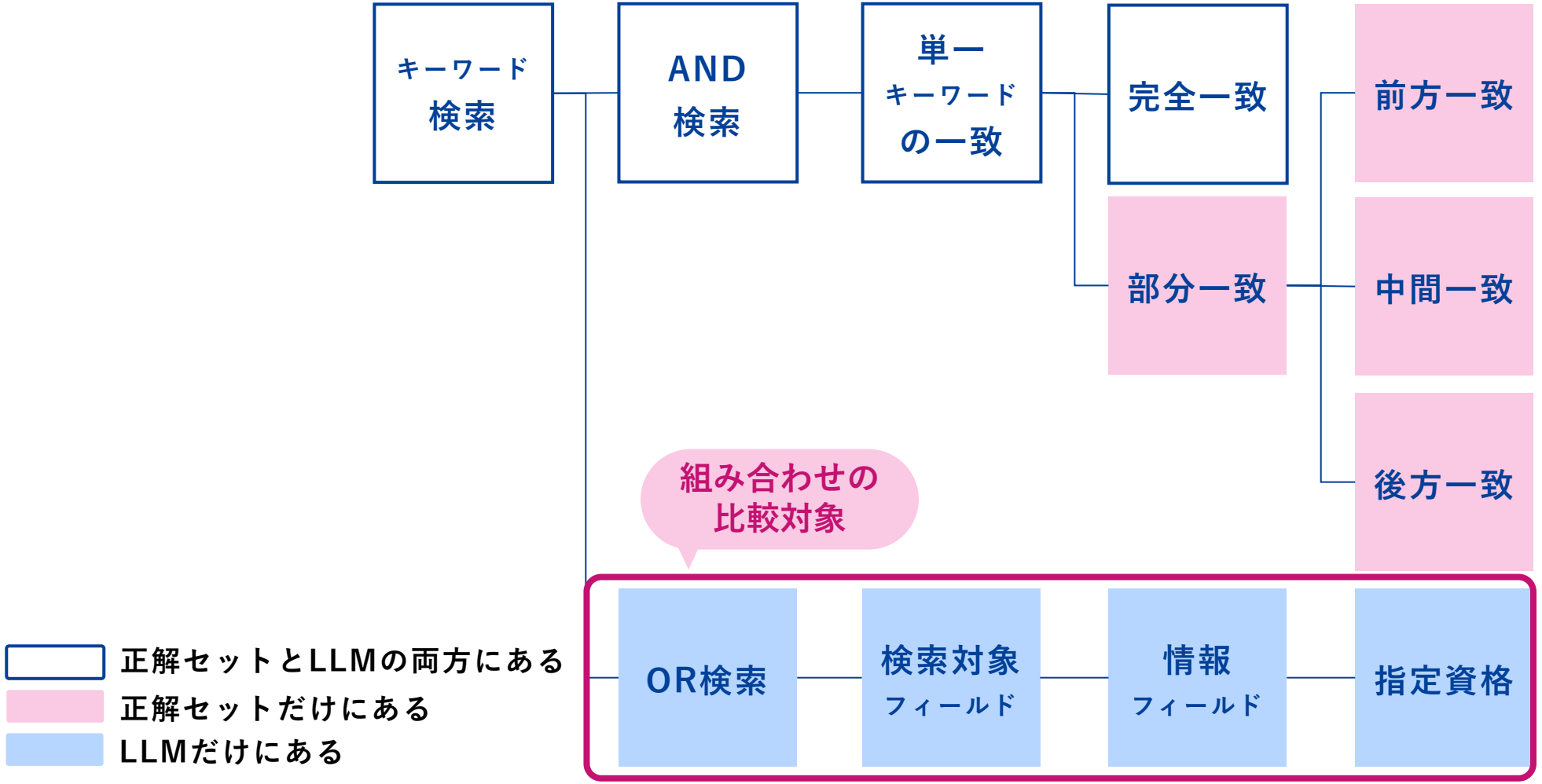
比較結果例　社員情報検索システム　テスト観点（単一）



リスト化して比較

テスト観点LLM	正解との一致
キーワード検索	○
単一キーワードの一致	○
完全一致	○
検索対象フィールド	×（有効）
情報フィールド	×（有効）
…	…

比較結果例　社員情報検索システム　テスト観点（組み合わせ）



テスト観点組合LLM	正解との一致
キーワード検索＞AND検索＞単一キーワードの一致＞完全一致	○
キーワード検索＞OR検索＞検索対象フィールド＞情報フィールド＞指定資格	×（有効）

比較結果例　検出可能な欠陥　社員情報検索システム

構成は異なるが、欠陥が一致していれば同じと判断できる

正解セット

テスト設計技法	テストケースの意味	事前条件	入力	期待値	検出可能な欠陥
同値分割	完全一致するキーワードで検索する	検索画面表示スキル「XXXYYY」を持つ社員が登録されている	キーワードで検索欄に「XXXYYY」を入力 「検索」押下	当該社員が検索結果画面に表示されること	**ANDのキーワード検索を実行した際、キーワードに完全一致する社員が検索結果画面に表示されない**
同値分割	前方一致するキーワードで検索する	検索画面表示スキル「XXXYYY」を持つ社員が登録されている	キーワードで検索欄に「XXX」を入力 「検索」押下	当該社員が検索結果画面に表示されること	ANDのキーワード検索を実行した際、キーワードに前方一致する社員が検索結果画面に表示されない

GPT-5.1Pro

テスト観点	前提条件	入力／操作	期待結果	検出可能な欠陥	一致
単一キーワードで一致する社員が検索される	社員Aの「指定ITスキル」に「Java」が登録されている。他の社員には「Java」を含む情報がない。検索方法：A)すべてを含むが選択されている。	キーワード欄に「Java」と入力して検索実行。	検索結果に社員Aのみが表示されること。	**ANDのキーワード検索を実行した際、キーワードに完全一致する社員が検索結果画面に表示されない**	○

比較結果例　検出可能な欠陥　社員情報検索システム

GPT-5.1Pro

テスト観点	前提条件	入力／操作	期待結果	検出可能な欠陥	一致
単一キーワードで一致しない場合、0件で正常終了する	いずれの社員情報・キャリア情報にも「Rust」を含む文字列が登録されていない検索方法：A)すべてを含む	キーワード欄に「Rust」と入力して検索実行	エラーにならず、検索結果件数0件として表示されること	**検索結果が0件の場合の表示が正しくない**	×(有効)
検索対象フィールド（情報フィールド）の網羅性確認	表に○がついている各フィールド（指定資格…）のそれぞれに、固有のキーワードを1つずつ設定した社員を用意。検索方法：B)いずれかを含む	各フィールドに設定したキーワードを1つずつ単独で検索実行	いずれのキーワードでも、そのキーワードを持つ社員が検索結果に表示されること（表の○が付いている全ての情報フィールドが検索対象になっていること）	・検索を実行した際、指定資格フィールドが検索対象として扱われない …	×(有効)
検索結果0件でも画面レイアウトや表示が崩れない	検索条件に一致する社員が存在しないテストデータであること	存在しない文字列（例：「ZZZZZZZZ」）を入力し検索実行	検索結果0件である旨が表示され、画面が異常表示にならないこと（ボタンや一覧ヘッダなどが正しく表示され続けること）	**検索結果が0件の場合の表示が正しくない**	×(無効)

結果と考察

LLMが生成したテストケース・テスト観点・検出可能な欠陥の件数

- 対象①：テスト管理ツール

テストケース数は全体的に少ない

	テストケース	テスト観点数	組み合わせ数	検出可能な欠陥
GPT-5	**14**	33	17	16
GPT-5.1	**17**	39	23	23
GPT-5.1 Thinking	**26**	44	27	29
GPT-5.1 Pro	**20**	48	28	32
（参考）正解セット	**50**	76	57	69

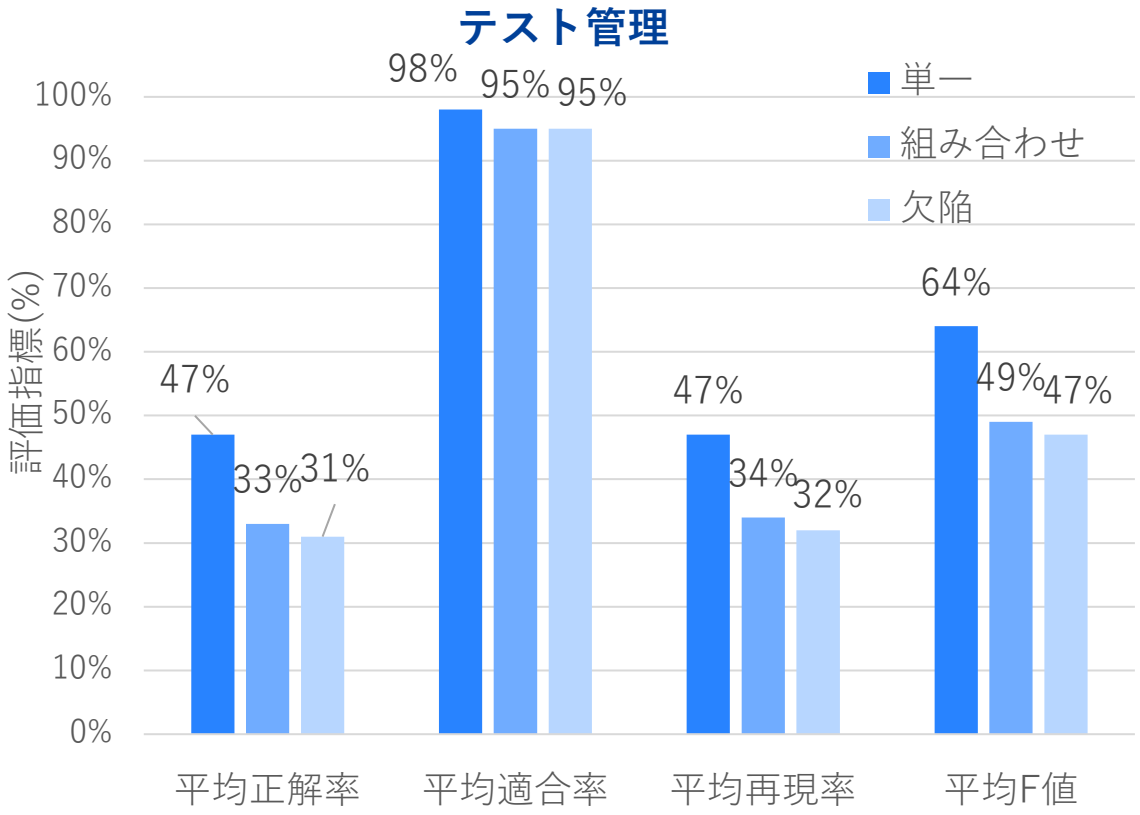
- 対象②：社員情報検索システム

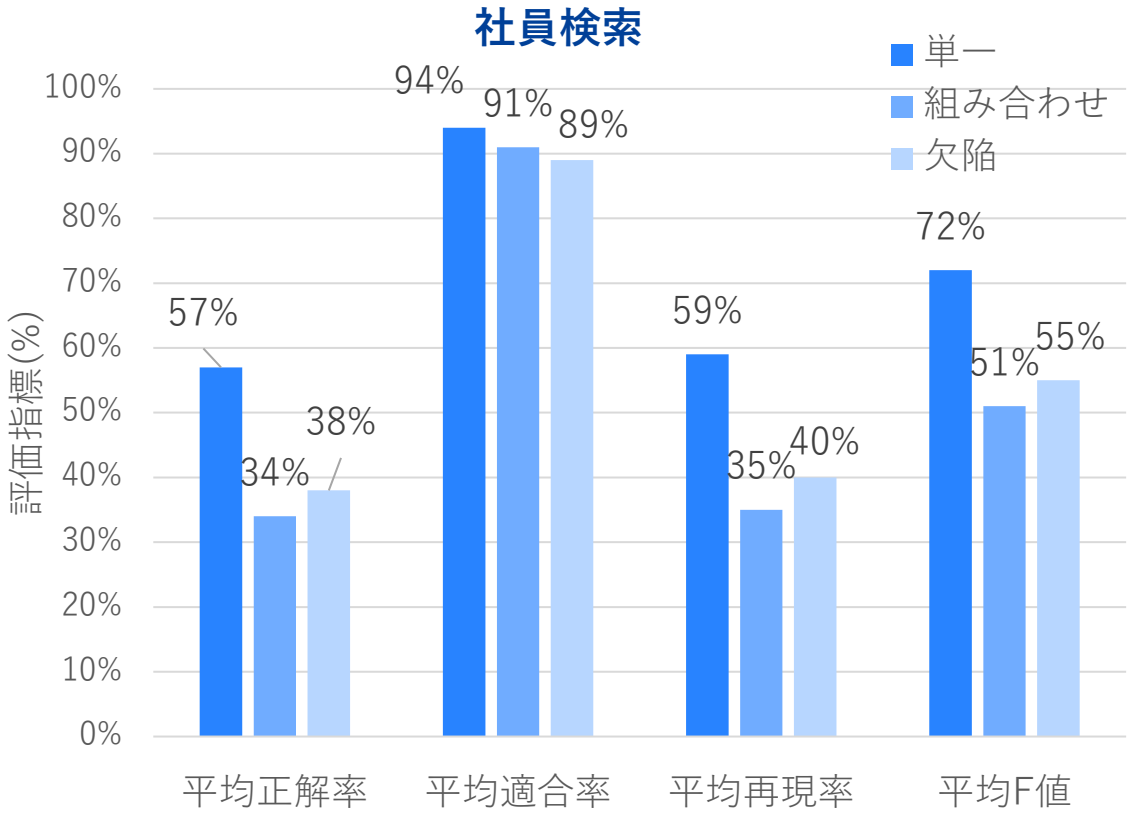
組み合わせ数が激減

	テストケース	テスト観点数	組み合わせ数	検出可能な欠陥
GPT-5	**15**	**42**	**18**	15
GPT-5.1	**25**	**52**	**25**	25
GPT-5.1 Thinking	**22**	**53**	**23**	26
GPT-5.1 Pro	**20**	**54**	**27**	31
（参考）正解セット	**56**	**51**	52	39

当たるが、漏れる

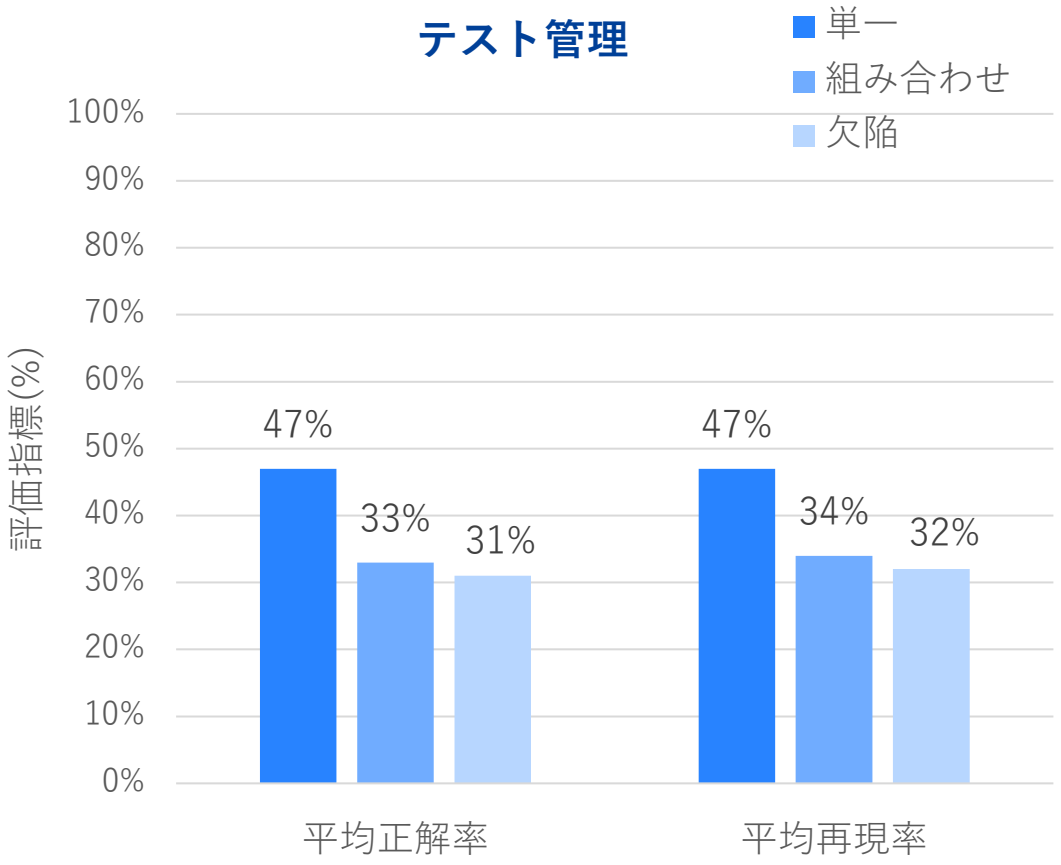
- 各モデルの平均値で見ると、適合率は89～98%と高いが、正解率や再現率は31～59%と低い

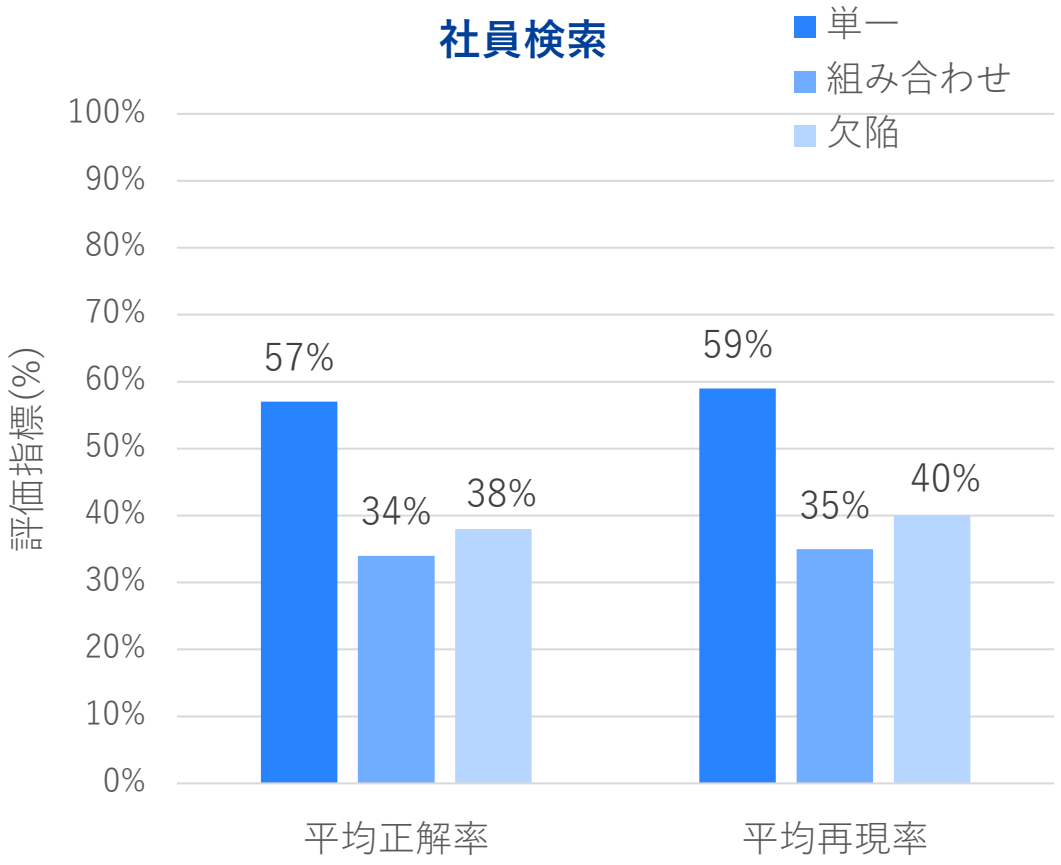




整理・列挙するのが優れている

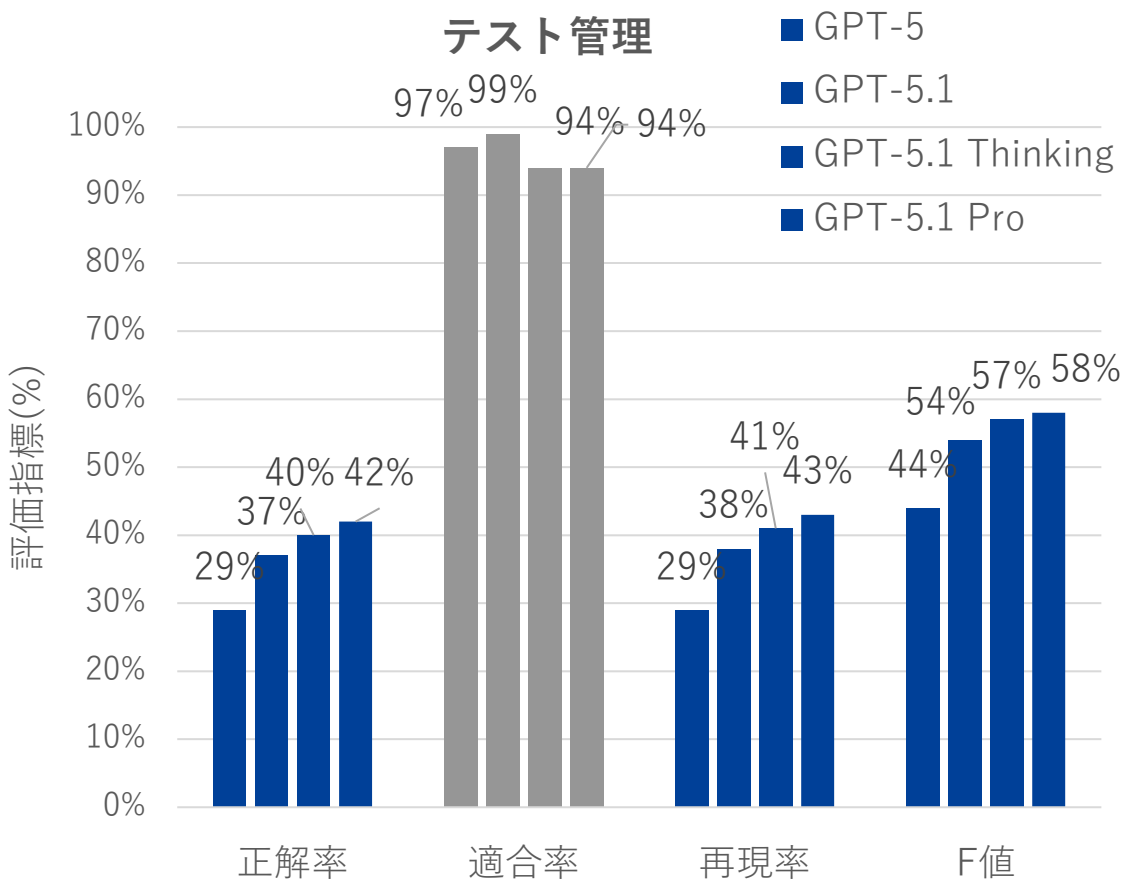
- 正解率や再現率について各モデルの平均値でみると、単一のテスト観点は47~59%と高く、組み合わせや欠陥は31~40%と低い

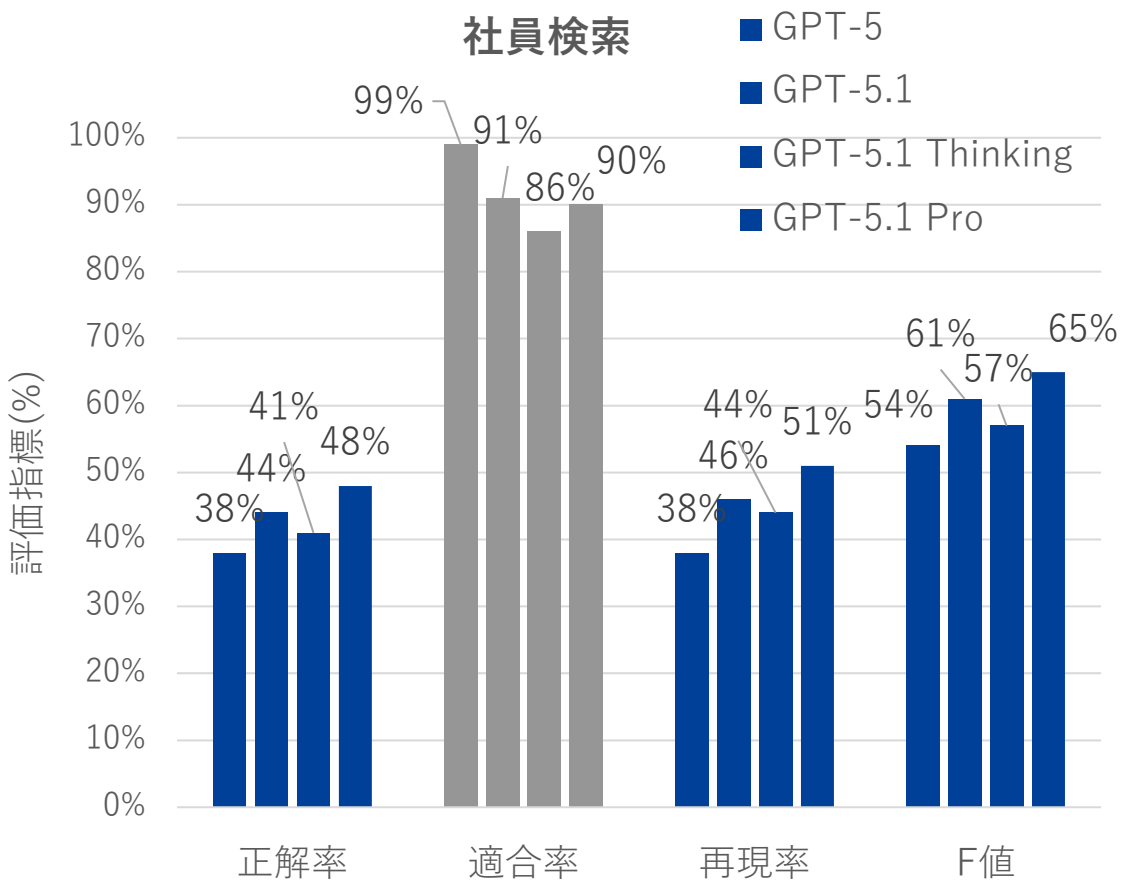




モデルによる違い

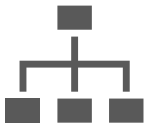
- 高い性能を発揮するモデルのほうが数値が高くなる傾向がある





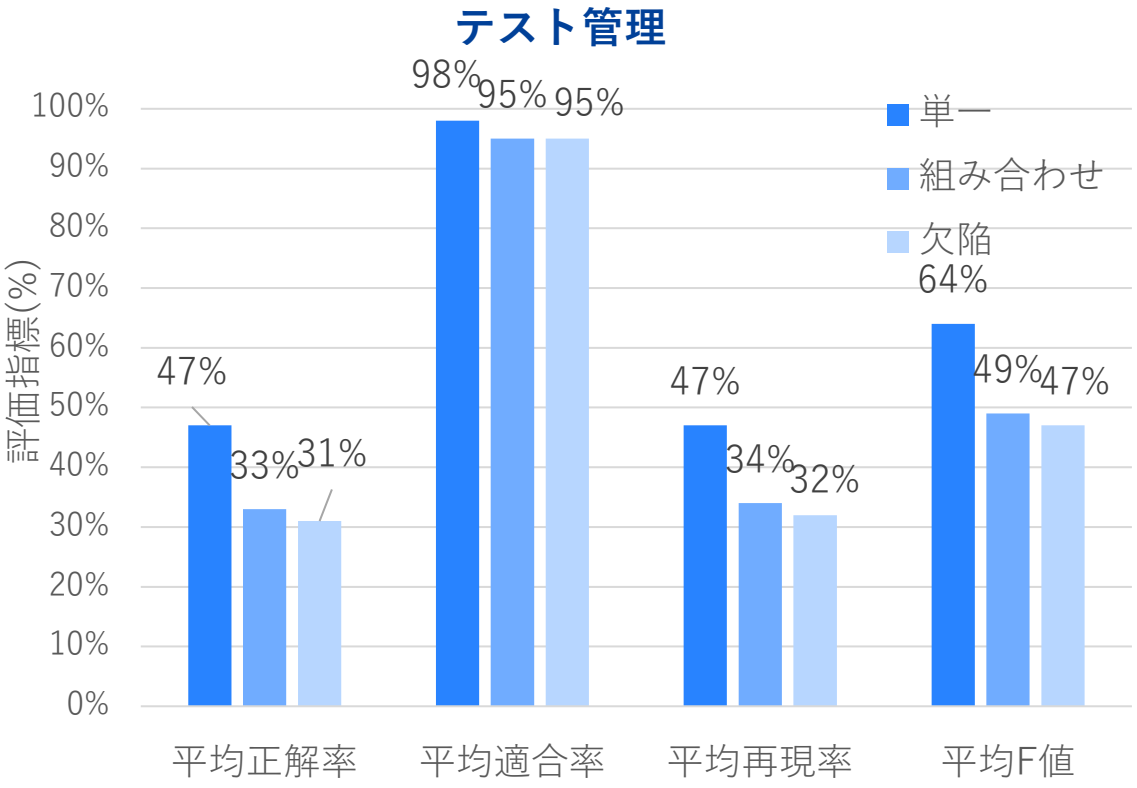
LLMの特徴

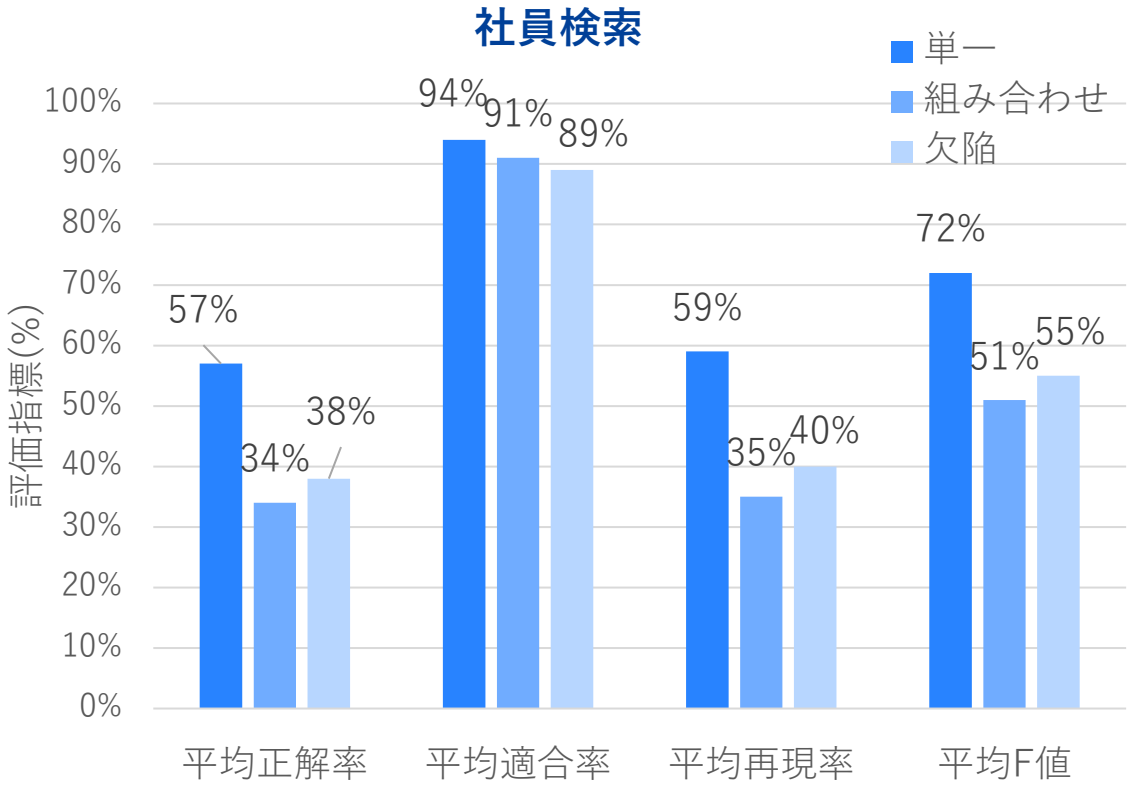
- LLMは代表的で妥当なものを慎重に抜き出してくる
- 要素の抽出・整理・列挙は比較的うまく出してくる
  - ▸ ただし、人のテスト分析能力に達しているとは言えない
- 構造化が苦手



両システム共に同じ傾向

- 適合率が高くて、正解率・再現率が低い傾向は同じ
- テスト観点単一は、組み合わせや欠陥よりもやや高い傾向は同じ



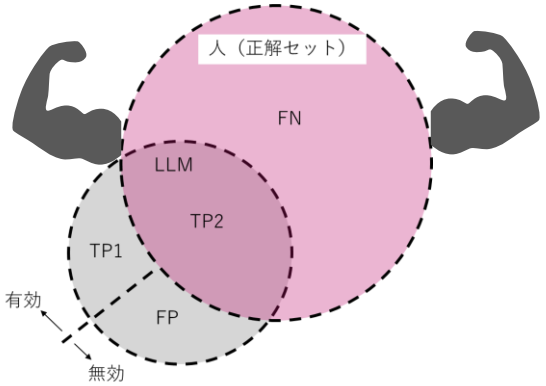


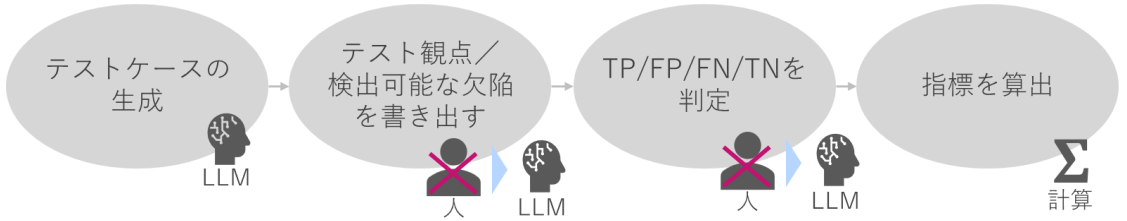
LLMとの付き合い方

- お勧めの使い方（現時点）
  - テスト分析のたたき台
    - その後の構造化は人が主導
  - 漏れのチェック
    - 人が作ったものに対する追加提案

測って、付き合う

- まとめ
  - LLMの能力はちまたの印象どおり？60点くらい
  - しかし重要なのは、日々進化するLLMを “常に計測できる” 枠組み
  - 本研究が、その基盤として位置付いてほしい
- 今後
  - モデルが出るたび自動計測できるようにする（開発中）
  - 正解セットの強化（強化中）





加速しよう、未来を。

VERISERVE

ぜひ、お試しください

株式会社ベリサーブ
広報部
TEL：050-3640-8194
MAIL：pr@veriserve.co.jp
住所：東京都千代田区神田三崎町 3-1-16
神保町北東急ビル9階

25-002

HQW!
Hello, Quality World!

イノベーションを加速させる
知恵と品質技術にアクセスする
テクノロジーライフメディア

www.veriserve.co.jp/helloqualityworld/