

早稲田大学 スマートエスイーIoT/AIコース第6期 修了制作

大規模言語モデルを活用した ソフトウェアの障害発生箇所予測

sse06-01 須田 峻介

(CANON IMAGING SYSTEMS INC.)

背景と課題

- ソフトウェア保守開発の現場で散見される問題

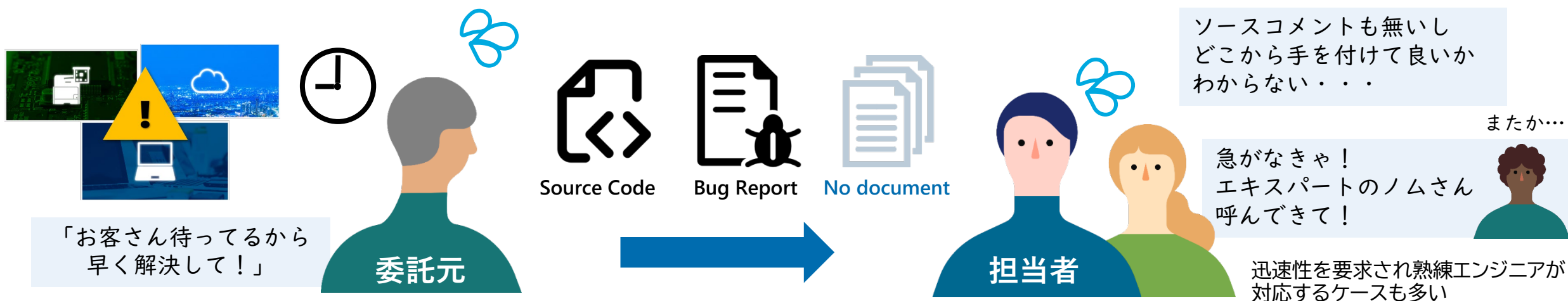
内部構造や設計理解の助けとなるドキュメントやコメントが存在しない

- コードリーディングを通じた

ソフトウェアの概要理解

バグ発生箇所の特定

に多くの時間を割かざるを得ないことが現場の課題。



手法・ツールの適用による解決



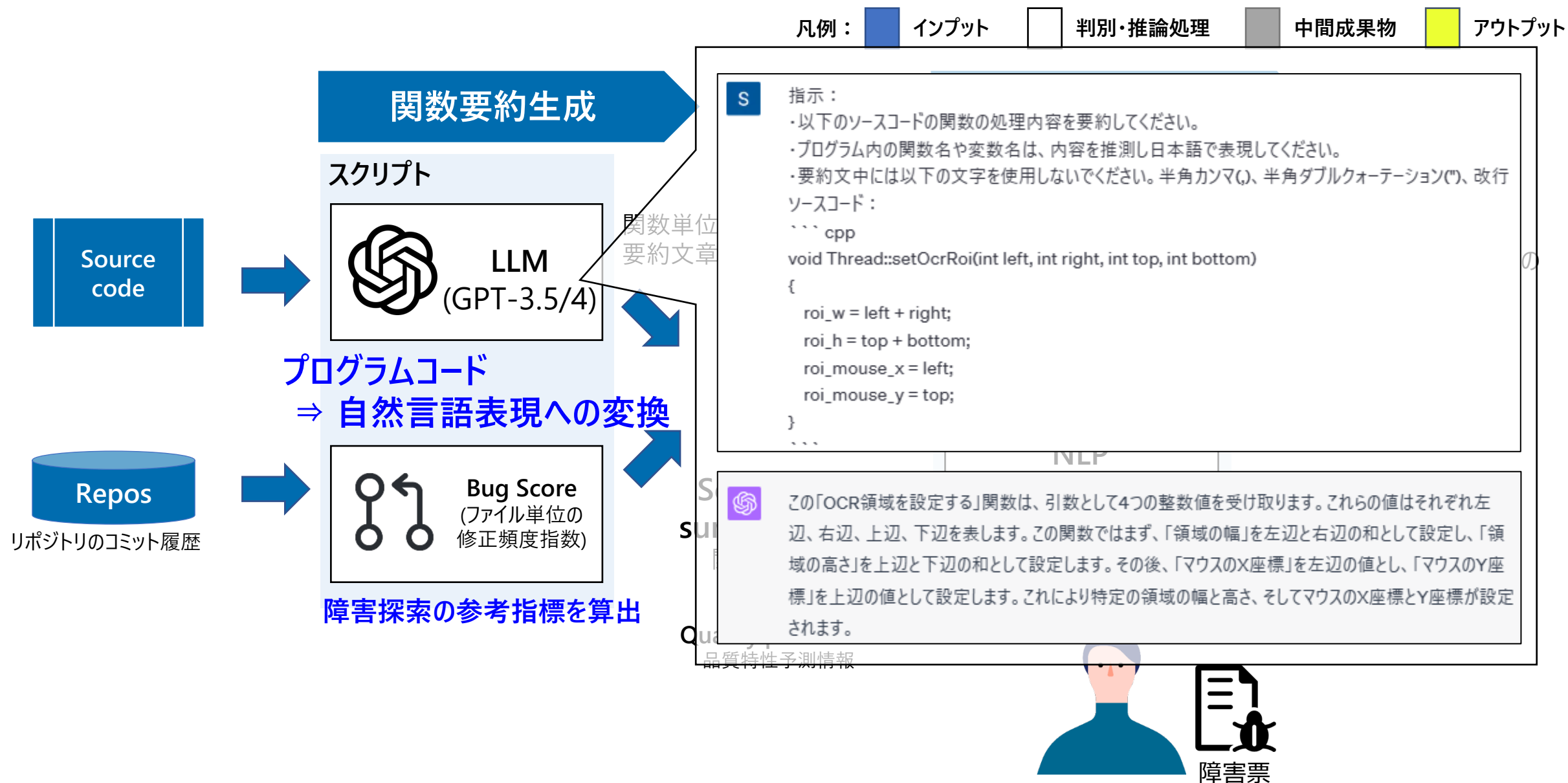
- 設計書やコメントが無くても **自然言語でプログラムの内容把握**
 - 気になる箇所と **関連がある関数をキーワード検索**
 - **障害が発生しやすいような箇所** のアタリをつけられる
- これらの情報を簡単に知ることができるガイドツールを作成

スキル依存性の減少

分析の迅速・省力化

システム概要

システム全体図



モデル化

要約情報生成の勘所/プロンプトエンジニアリング

英文プロンプトで指示



APIの使用単価=Token

- OpenAI APIを用いて生成を自動化、日本語は1文字1Tokenと割高
- 指示は英文、出力は日本語を得るようにプロンプトを調整
 - 英文プロンプトは回答精度向上の副次効果もあり^[1]

GPT-3: Code	
以下のソースコードの関数の処理内容を要約してください。プログラム内の関数名や変数名は、内容を基に日本語に翻訳すること。 要約文中で以下の文字は使用禁止。半角カンマ(,)、半角ダブルクォーテーション(“)。 改行(\n)。 ソースコード: ... cpp	
Clear	Show example
Tokens	Characters
193	141
Task: Summarize the functionality of the function in the source code and provide the answer in "Japanese". Translate the function and variable names in the program into Japanese based on context. In the summary, do not use the following characters: comma (,), double quotation marks ("), and newline (\n). Source Code: ... cpp	
Clear	Show example
Tokens	Characters
72	319
Task: Summarize the functionality of the function in the source code and provide the answer in "Japanese". Translate the function and variable names in the program into Japanese based on context. In the summary, do not use the following characters: comma (,), double quotation marks ("), and newline (\n). Source Code: ... cpp	
Clear	Show example
Tokens	Characters
72	319
Task: Summarize the functionality of the function in the source code and provide the answer in "Japanese". Translate the function and variable names in the program into Japanese based on context. In the summary, do not use the following characters: comma (,), double quotation marks ("), and newline (\n). Source Code: ... cpp	
Clear	Show example
Tokens	Characters
72	319
Task: Summarize the functionality of the function in the source code and provide the answer in "Japanese". Translate the function and variable names in the program into Japanese based on context. In the summary, do not use the following characters: comma (,), double quotation marks ("), and newline (\n). Source Code: ... cpp	
Clear	Show example
Tokens	Characters
72	319
Task: Summarize the functionality of the function in the source code and provide the answer in "Japanese". Translate the function and variable names in the program into Japanese based on context. In the summary, do not use the following characters: comma (,), double quotation marks ("), and newline (\n). Source Code: ... cpp	

図: 同様の内容のプロンプトを日本語と英語で計測した際のTokenサイズ

Few-shot promptingで誘導



回答例を追加

- 所望の様式が出力されない場合がある問題
 - Few-shot prompting^[2]で回答様式を方向づけ出力は安定化

望ましくない出力の例

- 単にプログラムの変数部分を日本語に翻訳しただけ
- 各変数の和訳を要約文章の最後に箇条書きで出力

最終的なベースプロンプト

Task: Description and summarize the functionality of the method in the source code and provide the answer in "Japanese". Translate the function and variable names in the program into Japanese based on context. Do not output the source code itself in your response. Source Code:

出力は日本語で

関数名や変数名は推測して日本語表現して

[1]:^[2308.01223] Do Multilingual Language Models Think Better in English? (arxiv.org)

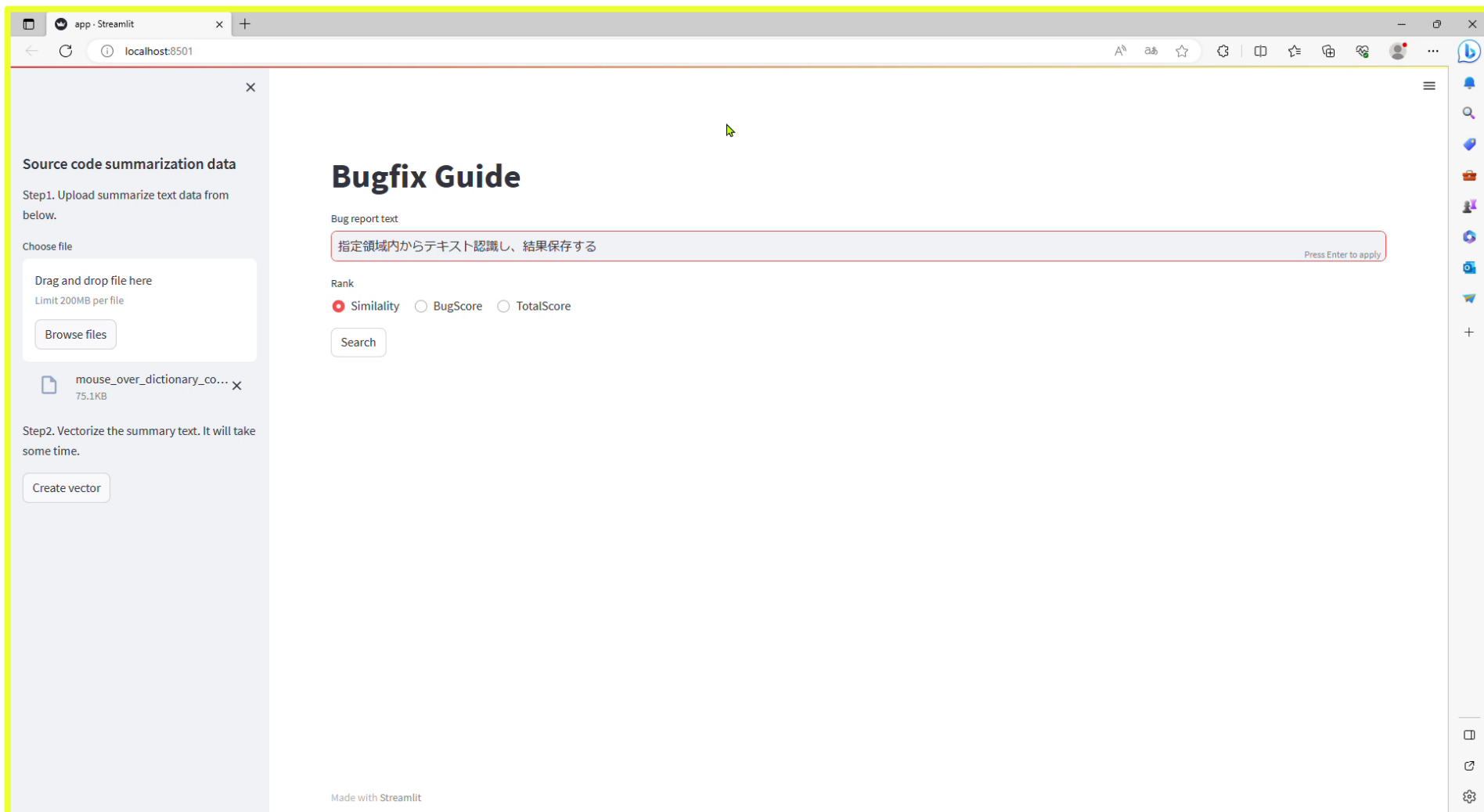
[2]:^[2202.12837] Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? (arxiv.org)

モデル化

類似度判定の使用感検証

- 関数要約情報のベクトル化・類似度判定を、Webアプリで実装して使用感を検証
→関数数が増加しても1～2秒で表示完了、**使用感に問題なし**

デモ動画



評価

GitHub公開ソースでのPoC／正解率(Average Precision)と乖離度に関する考察

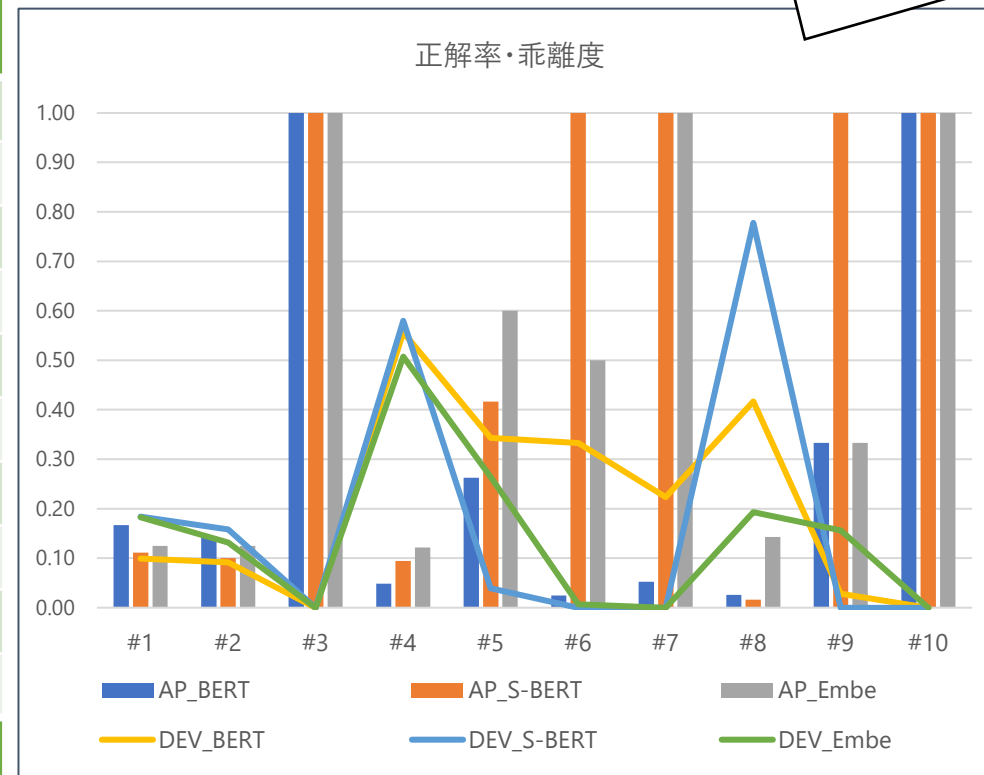
Commit log = キーワード
Change line = 正解関数

正解率大きい
→精度良

乖離度小さい
→精度良

複数のベクトル化モデルを比較
(BERT/Sentence-BERT/OpenAI Embeddings)
結果、Sentence-BERTが最もバランスが良い

#	キーワード	正解関数 No.	S-BERT 正解率	S-BERT 乖離度
1	前回終了時のウィンドウ位置の保存・復元	24	0.11	0.18
2	ウィンドウの透明度の設定	24	0.10	0.16
3	英辞郎のハイパーリンク書式	55	1.00	0.00
4	画面停止機能	24,42	0.09	0.58
5	メインウィンドウのフォントカラー設定	63,27	0.50	0.04
6	指定領域内からテキスト認識し、結果保存する	49	1.00	0.00
7	画像の上下反転	6	1.00	0.00
8	HTMLスタイル指定部分	54	0.02	0.78
9	ウィンドウを非表示にする	39	1.00	0.00
10	OCRを無効にする	58	1.00	0.00
Total Score			5.74	1.74



- ・ **正解率**：キーワードからバグの関数を正しく判定できているものもあることがわかる
 - － 関数要約情報は中央値で約500文字前後のため、文章単位でベクトル化するモデルが高い類似度が得られた
- ・ **乖離度**：要約文章とキーワード間の類似度をある程度正しく判定できていると考えられる

評価

文章類似度とソースコード品質との関係

要約情報の精度に影響するソースコードの品質特性を発見

→ 関数要約情報の精度は関数の凝集度や命名等に依存

凝集度の影響

No.24の関数要約文

このコードは `XxxxxXxxxx` クラスのコンストラクタです。このクラスはおそらくユーザーインターフェース (UI) を持つウィンドウを表現するクラスです。`XxxxxXxxxx` クラスは `QMainWindow` を継承していますので、Qtフレームワークの主要なウィンドウクラスを基本としています。コンストラクタでは、以下の操作が行われます：1. UIの設定 2. スクリーンのサイズ取得 3. 設定ファイルの読み込み 4. ウィンドウの透明度設定 5. **前回終了時のウィンドウの位置とサイズを復元** 6. 「マウス追従ウィンドウ」(カーソルの動きに追従する小さなウィンドウ)の作成と設定 7. 「画面停止ウィンドウ」の作成と設定 8. 検索履歴の設定 9. スレッドとのシグナル接続 (これによりUIとスレッドの間で情報の受け渡しが可能となります) 10. クリップボードからの設定 11. キーボードショートカットの設定 12. 文字認識スレッドの開始これらの操作を通じて、ウィンドウの設定を完了し、必要なリソースを確保し、必要に応じてユーザーとの対話を開始できます。

#1: Top 9 (類似度0.7179)

【他の処理情報を削除→高凝集な関数内容に修正】

このコードは `XxxxxXxxxx` クラスのコンストラクタです。以下の操作が行われます：
前回終了時のウィンドウの位置とサイズを復元、ウィンドウの設定を完了したら、スレッドを開始します。

#1: Top 1 (類似度0.7642)

高凝集な関数の要約情報の場合、正解率は上昇

命名の影響

変数名が適当

```
ソースコード：
... cpp
void Thread::setOcrRoi(int left, int right, int top, int bottom)
{
    w = left + right;
    h = top + bottom;
    temp_x = left;
    temp_y = top;
}
```

・w をただの「幅」と表現
・temp_x を「一時的なX座標」と表現してしまう

この「OCR領域を設定する」関数は、四つの整数値を引数として受け取ります。これらの値はそれぞれ左辺、右辺、上辺、下辺を表します。関数では、「幅」を左辺と右辺の和として、「高さ」を上辺と下辺の和として設定します。その後、「一時的なX座標」には左辺の値、「一時的なY座標」には上辺の値を設定します。

変数名が適切

```
... cpp
void Thread::setOcrRoi(int left, int right, int top, int bottom)
{
    roi_w = left + right;
    roi_h = top + bottom;
    roi_mouse_x = left;
    roi_mouse_y = top;
}
```

roi_w を「領域の幅」と表現

処理の内容も
「マウスのX座標を～」のように
自然言語表現で出力

この「OCR領域を設定する」関数は、引数として4つの整数値を受け取ります。これらの値はそれぞれ左辺、右辺、上辺、下辺を表します。この関数ではまず、「領域の幅」を左辺と右辺の和として設定し、「領域の高さ」を上辺と下辺の和として設定します。その後、「マウスのX座標」を左辺の値とし、「マウスのY座標」を上辺の値として設定します。これにより特定の領域の幅と高さ、そしてマウスのX座標とY座標が設定されます。

本修了制作での取り組み成果まとめ

①LLMを使ってソースコードの要約・キーワード探索ができることを確認



②データ生成用のスクリプトとキーワード検索用Webアプリを実装



③LLMによる関数要約文章の精度は
ソースコードの凝集度や命名等の可読性に依存するという知見が得られた

- 本修了制作で作成したツールは、**関数要約を元にしたキーワード探索は実現できており、障害分析の迅速・省力化ならびに意思決定支援に寄与できる**と考える
 - 今後は、事前知識を考慮した要約情報をLLMに生成させ要約精度を向上させたり、ユーザーテストによるツール評価を行い有用性の裏付けをしていくこと等が必要

最後に 謝辞

- 本修了制作において、御指導いただきました
九州大学 鵜林 尚靖 先生、近藤 将成 先生 へ感謝申し上げます
- 以下の皆様にも合わせて感謝申し上げます
 - 他授業で御指導いただいた、各先生方
 - 講座運営を取り仕切っていただいた、事務局の皆様
 - 日々切磋琢磨し励ましあった、第6期受講生の皆様
 - 半年間の受講環境を支えてくれた家族一同

以上