

Harnessing Large Language Models for Curated Code Reviews

Oussama Ben Sghaier

Université de Montréal

Montréal, Canada

oussama.ben.sghaier@umontreal.ca

Martin Weyssow

Singapore Management University

Singapore

mweyssow@smu.edu.sg

Houari Sahraoui

Université de Montréal

Montréal, Canada

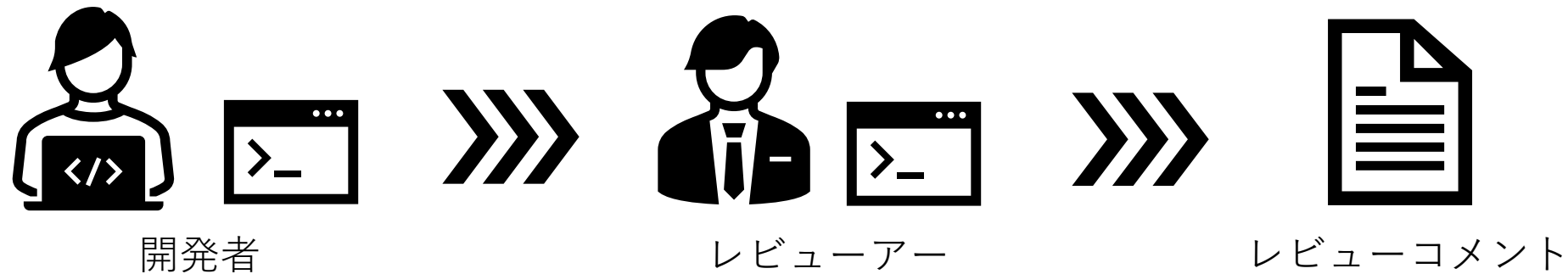
sahraouh@iro.umontreal.ca

Conference: 2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)

発表者：野口

コードレビュー

コード品質維持, バグの発見などを目的



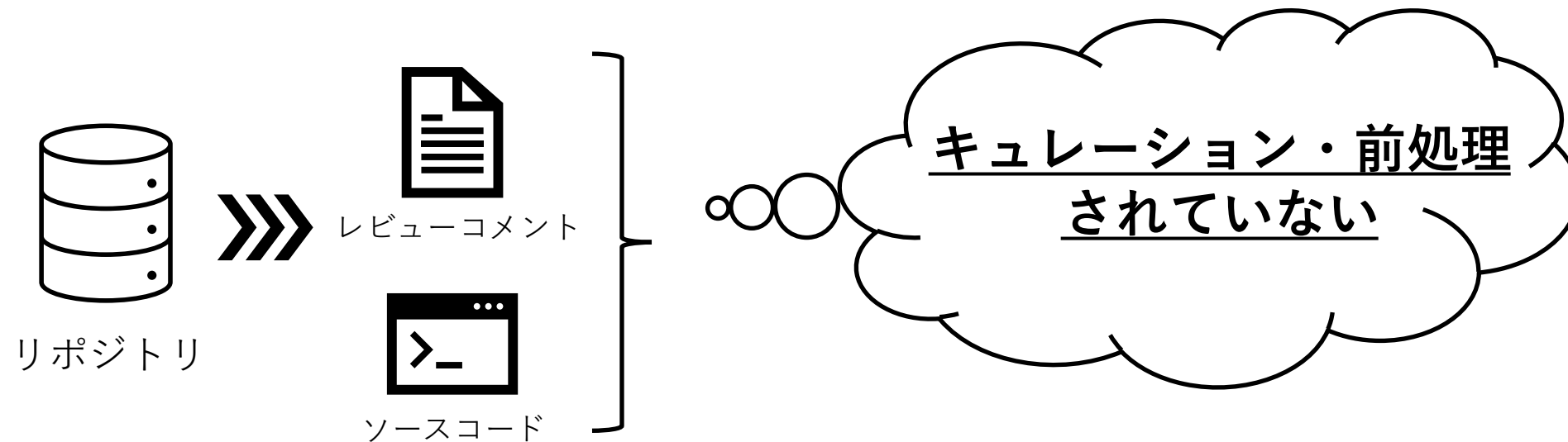
レビューコメント自動生成

- 機械学習と自然言語処理からコードを分析し, コメント生成
- 問題点
 - 正確かつ実用的なコメントを生成するのが難しい
 - 人間が生成するコメントには匹敵しない

問題点の改善

トレーニングデータの品質

- データ不十分な場合，モデルは効果的に機能しない



➤ 様々な形式のノイズが含まれている可能性

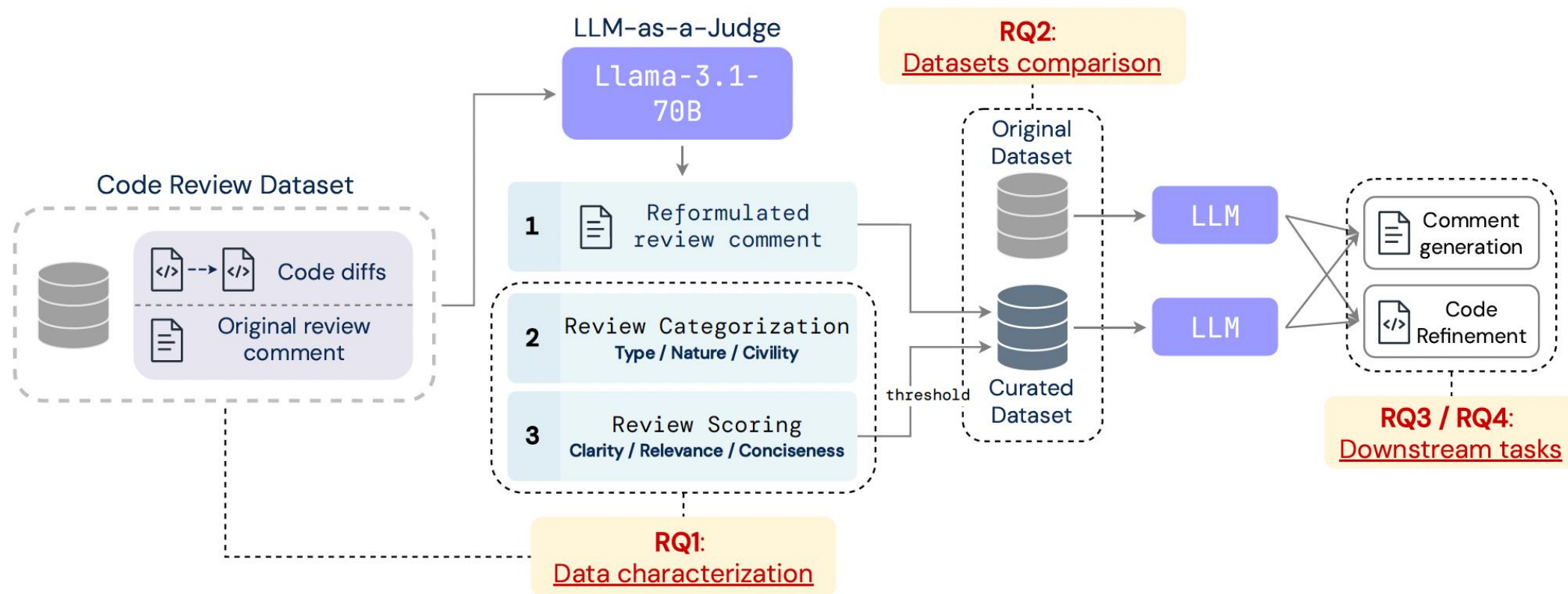
モデルが望ましくないパターンを学習し，支離滅裂なコメントを生成してしまう

目的とアプローチ

目的

- コードレビューデータセットの品質を向上させる

アプローチ

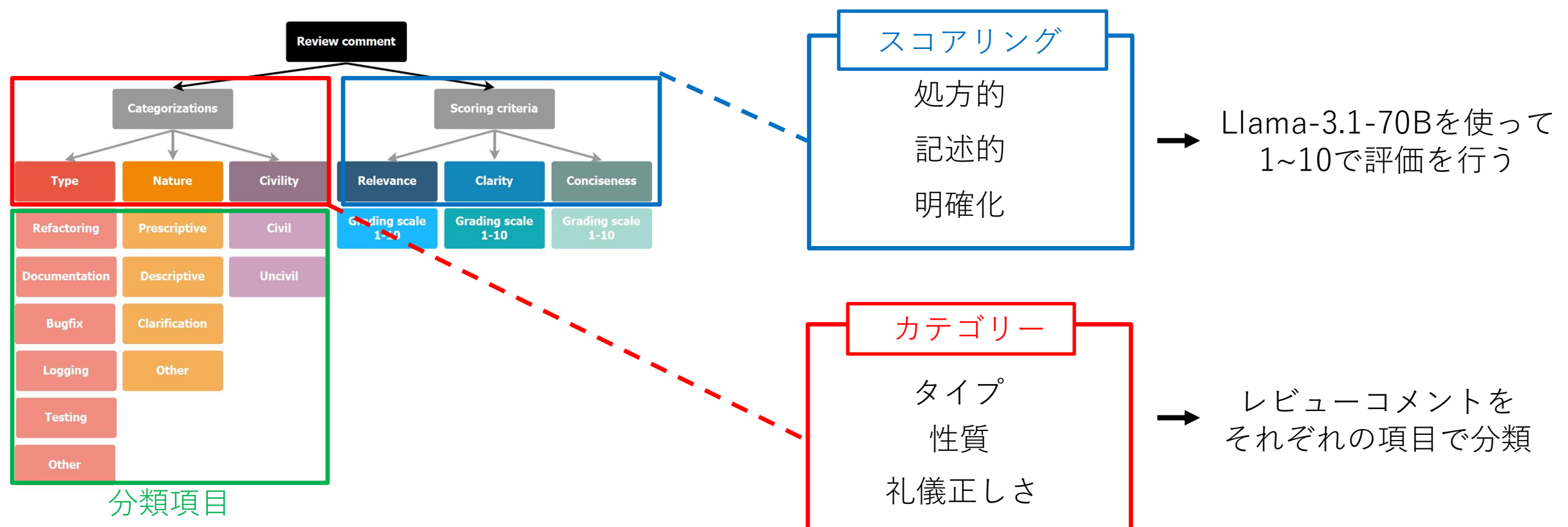


RQ1：データセットの品質評価

データセットの選択

- コードレビュー用の最大のデータセット^[11]を使用
 - 9つのプログラミング言語からなる176,613のサンプルで構成

評価フレームワーク



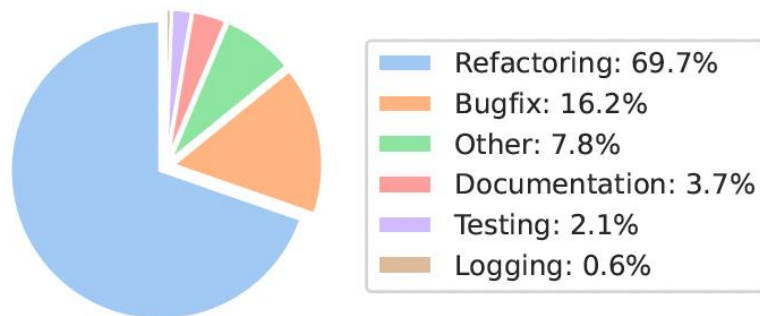
[11] Z. Li, S. Lu, and D. e. a. Guo, "Automating code review activities by large-scale pre-training," in Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 1035–1047.

RQ1：結果

品質の分類結果

Type

タイプ

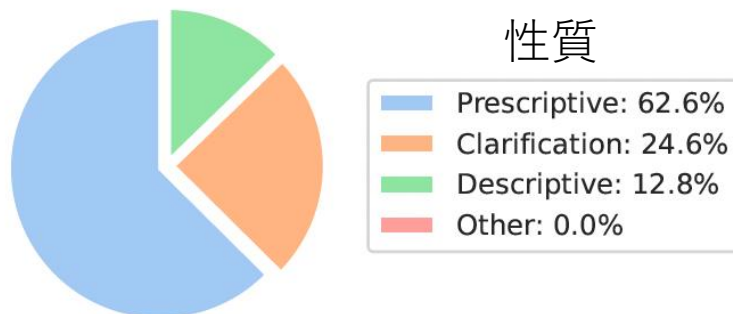


- ・ 大部分がリファクタリングカテゴリーに分類
→ コードの品質とパフォーマンスを向上させることに焦点

コードの品質維持とバグ防止といったコードレビューの目的に一致

Nature

性質

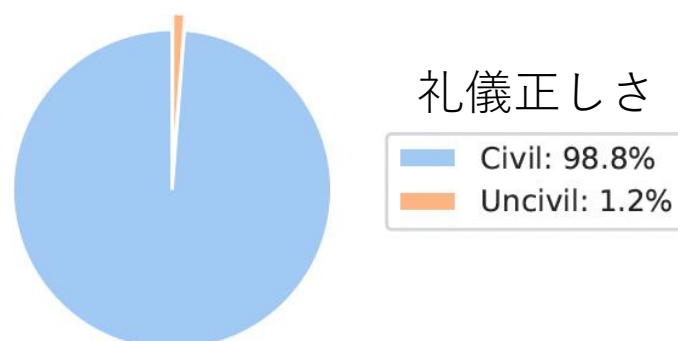


- ・ 処方的なレビューコメントが62.6%を占めている
→ 正確なコード変更を行うのを助ける実用的なコメントに焦点

直接的な解決策と明示的なフィードバックを提供したレビューコメント

Civility

礼儀正しさ

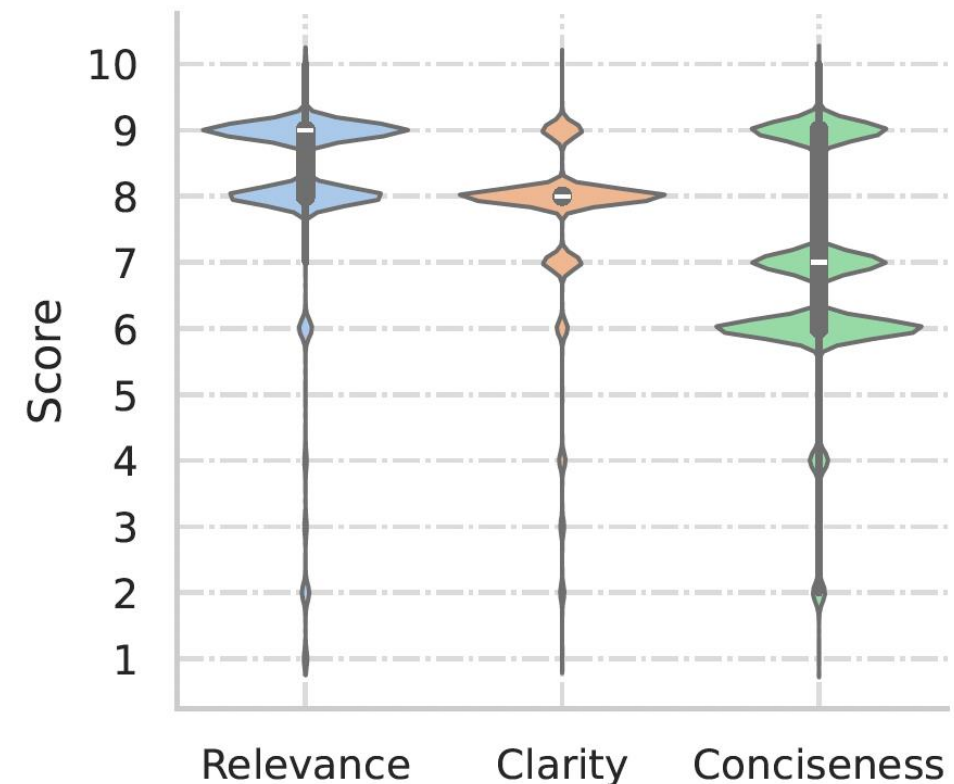


- ・ ほとんどのコメントが礼儀正しいとして分類
→ コードレビュープロセスが専門的に行われていることを示唆
- ・ 無礼なコメントも少ない（1.2%）が存在している
→ 不適切なコメント生成を防ぐために、完全に削除する必要がある

RQ1：結果

スコアリング基準の結果

		関連性	明瞭性	簡潔さ
Category	Subcategory	Relevance	Clarity	Conciseness
Type	Refactoring	8.32	7.79	6.99
	Bugfix	8.53	7.74	6.84
	Testing	8.42	7.92	6.97
	Logging	8.43	7.84	6.85
	Documentation	8.33	7.62	6.72
	Other	7.02	6.86	5.90
Nature	Descriptive	7.14	6.63	5.61
	Prescriptive	8.52	7.95	7.19
	Clarification	8.29	7.57	6.66
	Other	4.24	4.40	4.12
Civility	Civil	8.26	7.75	6.93
	Uncivil	5.60	4.34	4.34
Average	–	8.23	6.89	7.71



- リファクタリング・テストのスコアが高い
 - 明確で簡潔に書かれていることから、コメントがよく理解されている
- その他・無礼なコメントはスコアが低い
 - データセットの品質を低下させてしまっている

RQ2：データセットキュレーション

レビューコメントのフィルタリング

- 関連性の指標について，閾値を4としフィルタリング
 - 5,895のレビューコメントを除去

レビューコメントの再構築

- 礼儀正しさ，明瞭さ，簡潔さの強化を目的
- プロンプトを元にレビューコメントを生成
 - 生成されたコメントをRQ1と同様に評価

RQ2：結果

明瞭さと簡潔さの両方で改善

Category	Subcategory	Clarity	Conciseness
Type	Refactoring	8.95 (↑0.63)	8.06 (↑1.07)
	Bugfix	8.98 (↑0.45)	8.03 (↑1.19)
	Testing	8.98 (↑0.56)	8.03 (↑1.06)
	Logging	8.97 (↑0.54)	8.02 (↑1.17)
	Documentation	8.98 (↑0.65)	8.02 (↑1.30)
	Other	8.95 (↑1.93)	8.05 (↑2.15)
Nature	Descriptive	8.76 (↑1.62)	8.02 (↑2.41)
	Prescriptive	8.96 (↑0.44)	8.05 (↑0.86)
	Clarification	8.96 (↑0.67)	8.03 (↑1.37)
	Other	9.00 (↑4.76)	8.00 (↑3.88)
Civility	Civil	8.96 (↑0.70)	8.05 (↑1.12)
	Uncivil	–	–
Average	–	8.96 (↑2.07)	8.05 (↑0.34)

・ 明瞭さ

コメントがより構造化されていた

→文法的に正しい語彙と適切に形成された文になっている

・ 簡潔さ

LLMが凝ったテキストを生成する傾向がある

コメント分類結果

Category	Subcategory	Count	Percentage
Nature	Descriptive	1,674	0.95↓
	Prescriptive	159,306	90.20↑
	Clarification	29,586	16.75↓
	Other	17,491	9.90↑
Civility	Civil	176,613	100↑
	Uncivil	0	0↓

・ 処方的コメントが90.2%まで上がった

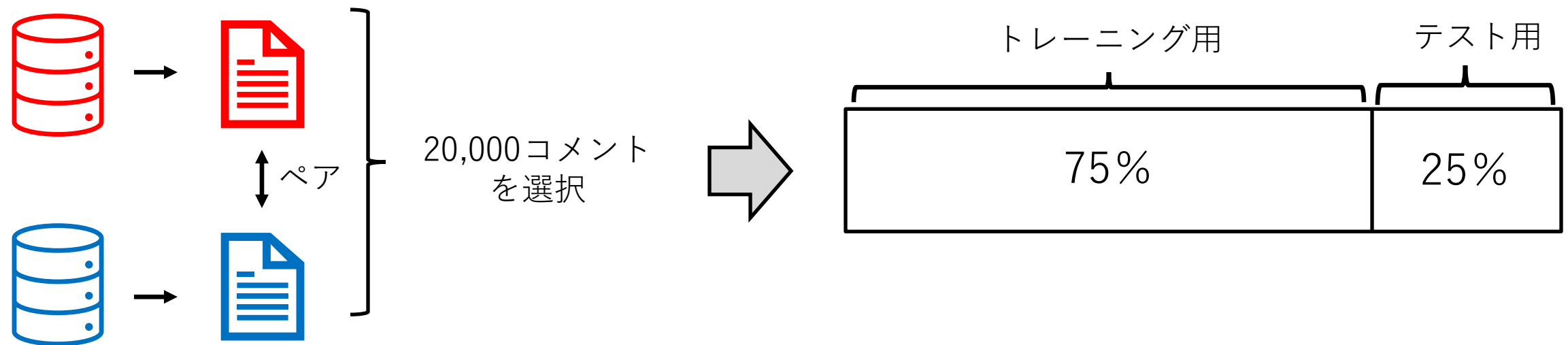
開発者へより直接的・明示的・実用的なコメントになった

・ 全てのコメントが礼儀正しいとなった

より専門的で建設的なデータセットが作成できた

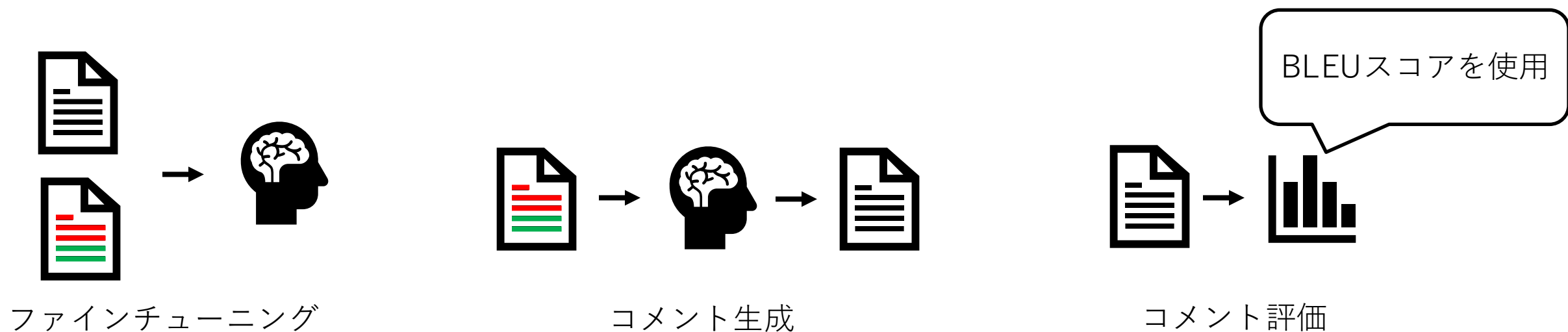
RQ3：自動コメント生成の影響

モデルとデータを選択



実験

- DeepSeek-6.7B-Instructを使用



RQ3：結果

データセットを再構築することでBLEUスコアが向上

	Original Comments	Curated Comments
BLEU	7.71	11.26

- 強化された明瞭さと構造からモデルが学習しやすくなっている

生成されたコメント例

Code change

```
@@ -75,6 +75,8 @@  
  
import com.google.common.annotations.VisibleForTesting;  
  
+import javax.servlet.http.Part;
```

Original dataset

Real review comment why this import is required ?

Generated review comment I think we should use `javax.servlet.http.Part` instead of `org.apache.commons.fileupload.FileItem`.

Curated dataset

Reformulated review comment This import is not immediately clear in its necessity. Could you provide context or explain why `javax.servlet.http.Part` is required for this functionality?

Generated review comment Consider removing the import statement for `javax.servlet.http.Part` as it's not used in the code.

RQ4：自動コードリファインメントの影響

データと実験

- RQ3と同様
- 入力をコード差分ではなく，コメントにし，コード差分を生成
- 評価メトリクス
 - Code-BLEU：生成されたdiffの類似性を測定
 - Exact Match（EM）：正確に一致するコードdiffの数を測定

実験結果

Dataset Version	CodeBLEU	Exact Match
Original Comments	0.36	408
Curated Comments	0.44	445

キュレーションされたコメントがレビュー指示のモデル解釈における曖昧さを減らすことを示唆

まとめ

コードレビューデータセットの品質を向上させるための方法を揭示

- データセットの品質を向上させるため、レビューコメントを再構築
- 元のデータセットと比較して、より良いコメントを生成できた

今後の研究

- 複数のLLMを審査員として組み込み、多様なレビューコメントを生成
- レビューコメントの優先順位付けをより探求