

AUCAD: Automated Construction of Alignment Dataset from Log-Related Issues for Enhancing LLM-based Log Generation

Hao Zhang¹, Dongjun Yu¹, Lei Zhang¹, Guoping Rong¹, Yongda Yu¹, Haifeng Shen², He Zhang¹, Dong Shao¹, Hongyu Kuang¹

¹Software Institute, Nanjing University

²Faculty of Science and Engineering,
Southern Cross University

In the 16th International Conference on Internetware 2025.

ログとは

システムが記録する動作履歴を**ログ**という。

- ログは、デバッガが使えないような本番環境での不具合原因特定に有用[2].

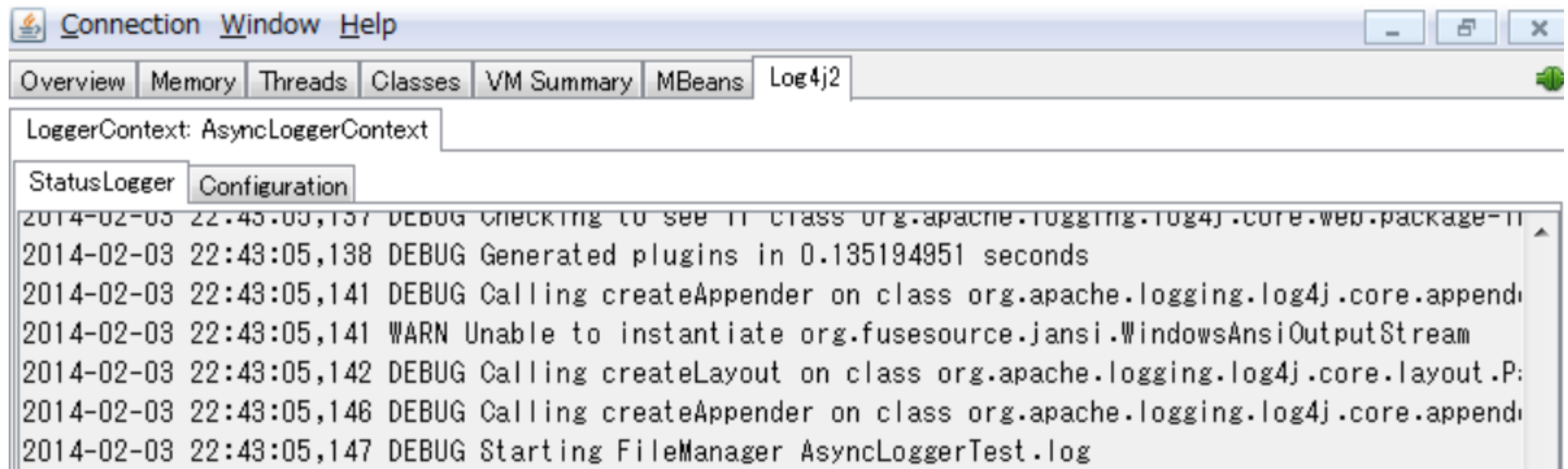


図 1 : ログの例

[2] Yuan, D., Park, S., & Zhou, Y. (2012). Characterizing logging practices in open-source software. 2012 34th International Conference on Software Engineering (ICSE), 102-112.

ロギング文

プログラムの実行中にログを出力するために、ソースコード中に記述する命令文を**ロギング文**という。

- JavaのAPIであるLog4jやSlf4jなどがロギング文を提供している。

ロギング文を含むソースコードの例

```
public void transfer(String from, String to, BigDecimal amount) {  
    logger.info("Attempting to transfer {} from [{}] to [{}].", amount, from, to);  
    try {  
        throw new SQLException("Connection timed out.");  
    } catch (SQLException e) {  
        logger.error("Failed to transfer from [{}] to [{}].", from, to, e);  
    }  
}
```

開発者は効果的なログ運用のために、適切なロギング文の記述が求められる

適切なロギング文による効果

不適切なロギング文により出力されるであろうログ

```
INFO - 処理を開始 user-123
INFO - OK product-abc
INFO - 処理を開始 null
INFO - 問題発生 user=null
INFO - 処理を開始 user-456
INFO - 例外をキャッチ: DB接続エラー
```

適切なロギング文により出力されるであろうログ

```
INFO - カートへの商品追加処理を開始。 userId: user-123, productId: product-abc
INFO - カートへの商品追加に成功。 userId: user-123, productId: product-abc
INFO - カートへの商品追加処理を開始。 userId: null, productId: product-xyz
ERROR - リクエストIDが不正なため処理を中断。 userId: null, productId: product-xyz
INFO - カートへの商品追加処理を開始。 userId: user-456, productId: product-def
ERROR - カートへの商品追加中に予期せぬエラーが発生。 userId: user-456, productId: product-def
java.lang.RuntimeException: データベースへの接続に失敗しました。
at GoodExampleWithoutMdc.addProductToCart(GoodExampleWithoutMdc.java:23)
at GoodExampleWithoutMdc.main(GoodExampleWithoutMdc.java:41)
```

ロギング文を適切に作成することで、ログからより有意義な情報が得られるようになる

適切なロギング文による効果

不適切なロギング文により出力されるであろうログ

INFO 処理を開始 user:123

しかし、適切なロギング文についての基準は明文化
されていないため、

適切なロギング文を記述するのは難しい



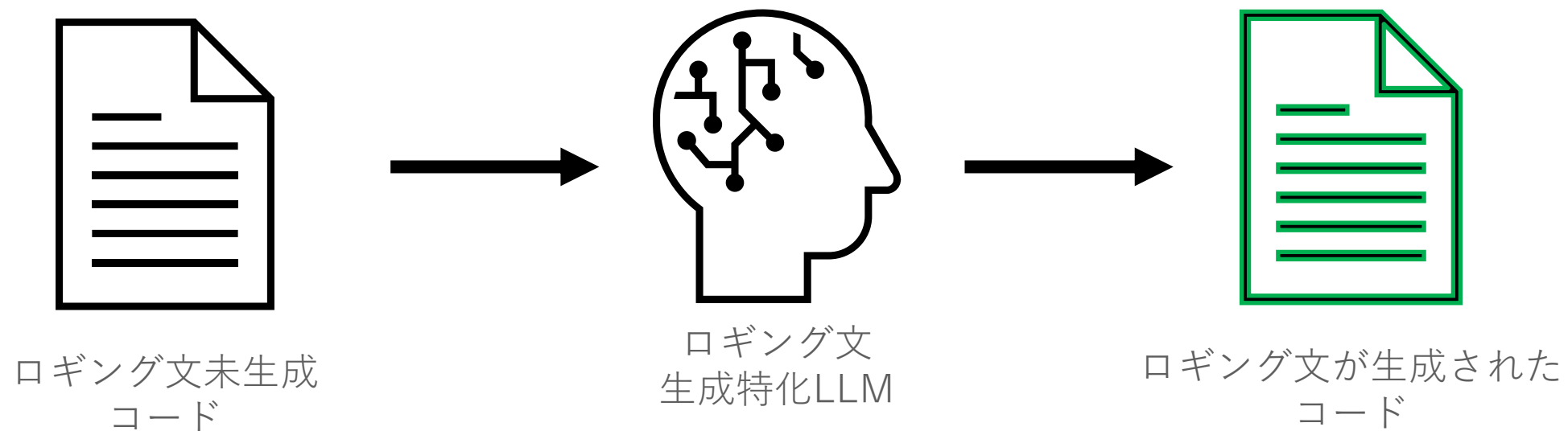
ロギング文記述の自動化が求められている

るようになる

ロギング文生成特化LLM

ロギング文生成特化のLLMを作成する研究が行われている[4].

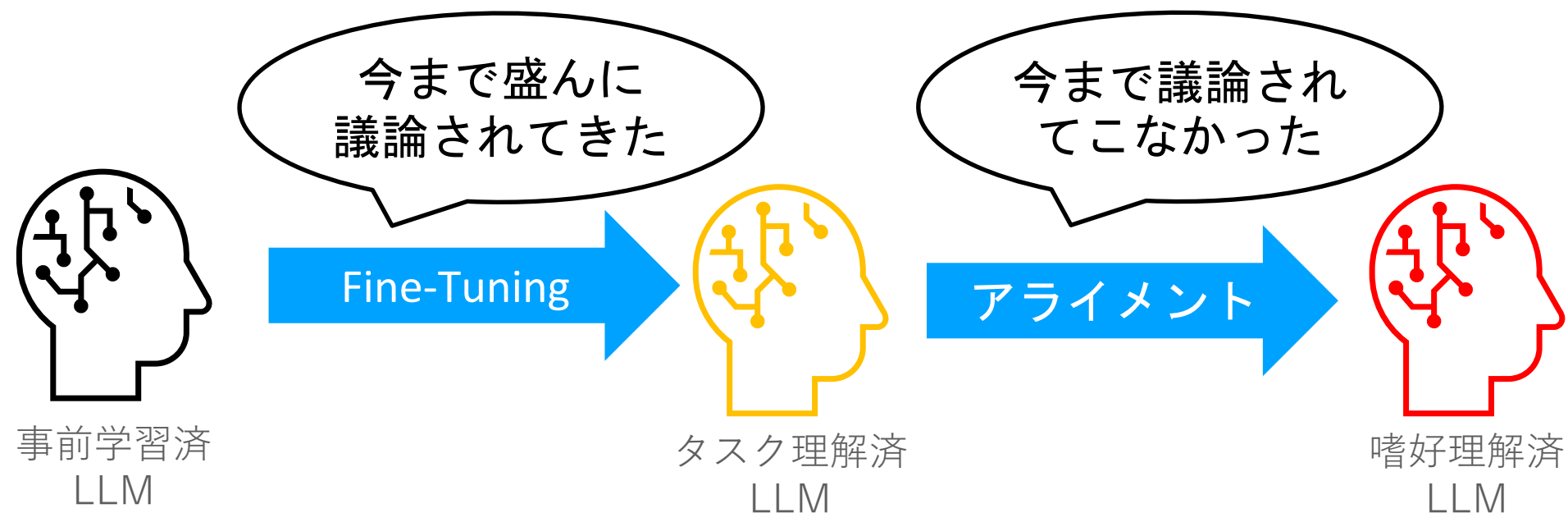
- LLMに人間が書いたロギング文を学習させることにより, 適切なロギング文の**暗黙的基準**を習得させる



[4]Mastropaolo, A., Pascarella, L., & Bavota, G. (2022, May). Using deep learning to generate complete log statements. In *Proceedings of the 44th international conference on software engineering* (pp. 2279-2290).

LLMにロギング文を学習させる手法

- ロギング文をLLMに学習させる手法はFine-Tuningが一般的
- アライメント(※)によるロギング文の学習はあまり議論されていない
 - アライメントに必要な正例と負例のラベル付けがされているロギング文データセットが存在していない



※アライメント：人間の好みや倫理観に合わせ、回答を調整する工程。LLMの回答を報酬モデルで評価することで好みを学習させる手法であるRLHF等がこれにあたる。

LLMにロギング文を学習させる手法

ロギング文をLLMに学習させる手法はFine-Tuningが一般

仮説：Fine-Tuningに加えて、アライメントも行えば、
ロギング文生成LLMの性能はさらに向上するのではないか？

※アライメント：人間の好みや倫理観に合わせ、回答を調整する工程。LLMの回答を報酬モデルで評価する手法であるRLHF等がこれにあたる。

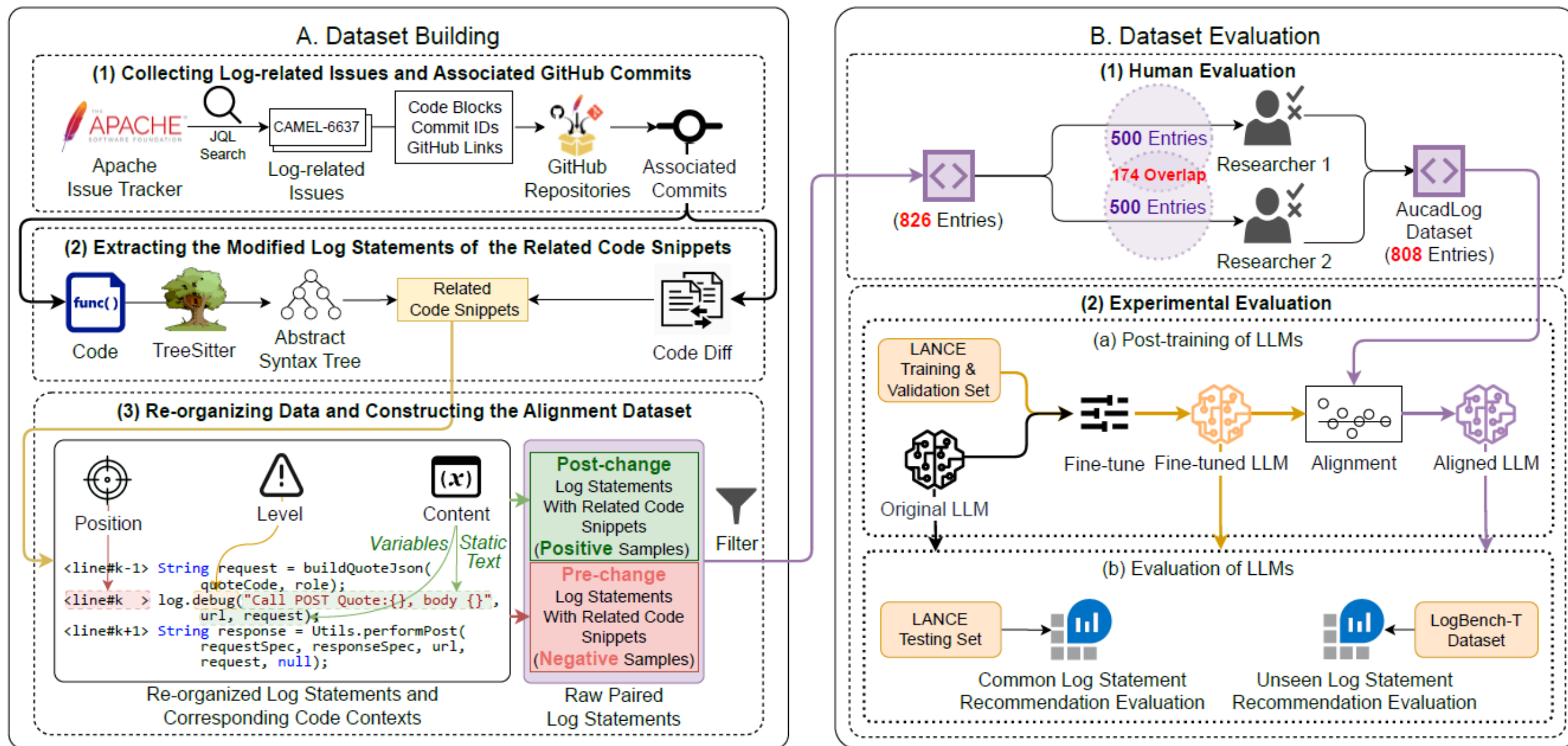
研究概要

目的：ロギング文生成における事後学習LLMの性能向上

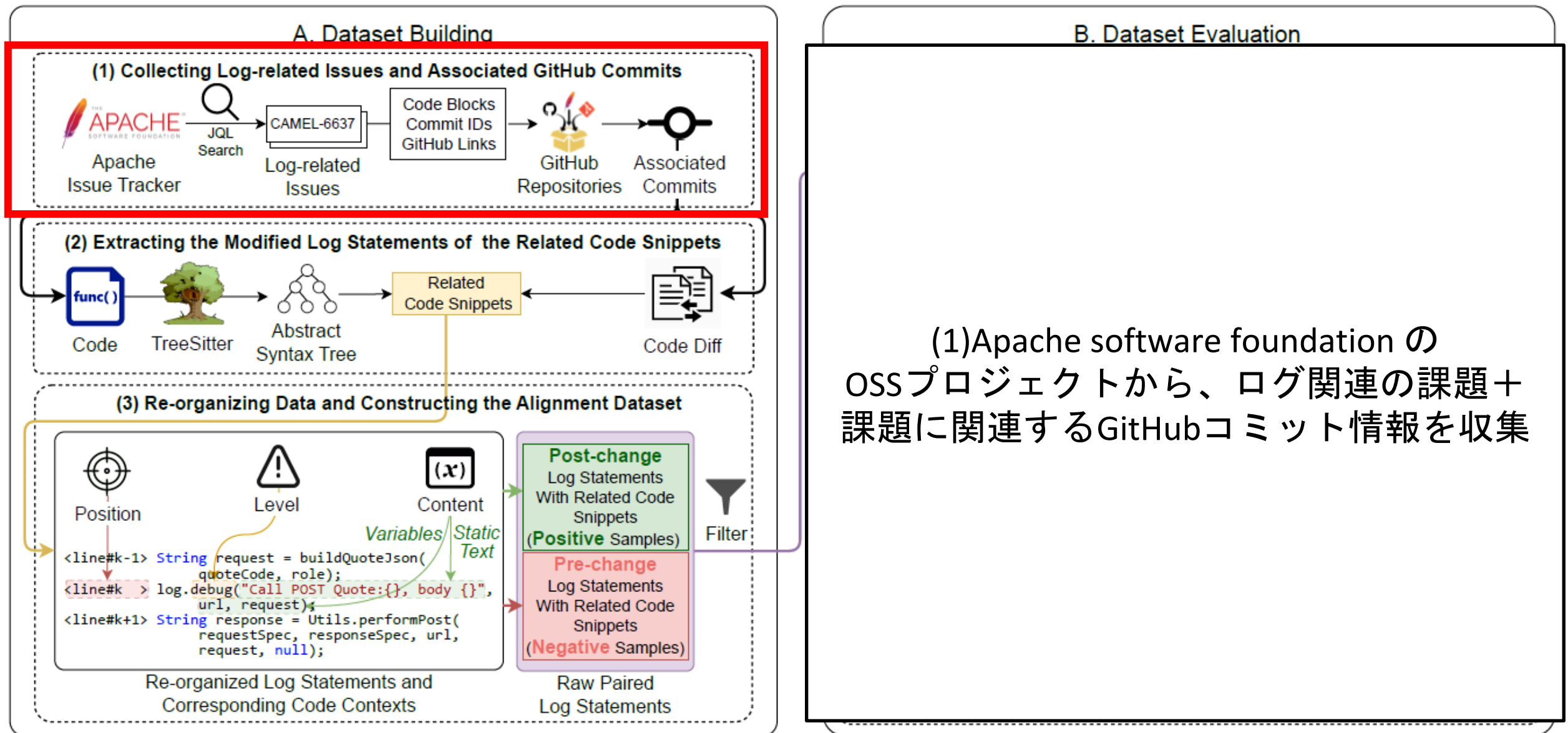
アプローチ：

- 高品質なログ文で構成される、構造化されたデータセットの構築手法(AUCAD)の提案
 - 高品質である理由：
 - ログ文に関連する文脈的なコードスニペットがコミュニティから大きな注目を集めている
 - ログ文と文脈的なコードスニペットの最終形態が、シニア開発者の承認を反映している
- AUCADにより構築したデータセットでアライメントを施したLLMのロギング文生成能力の評価

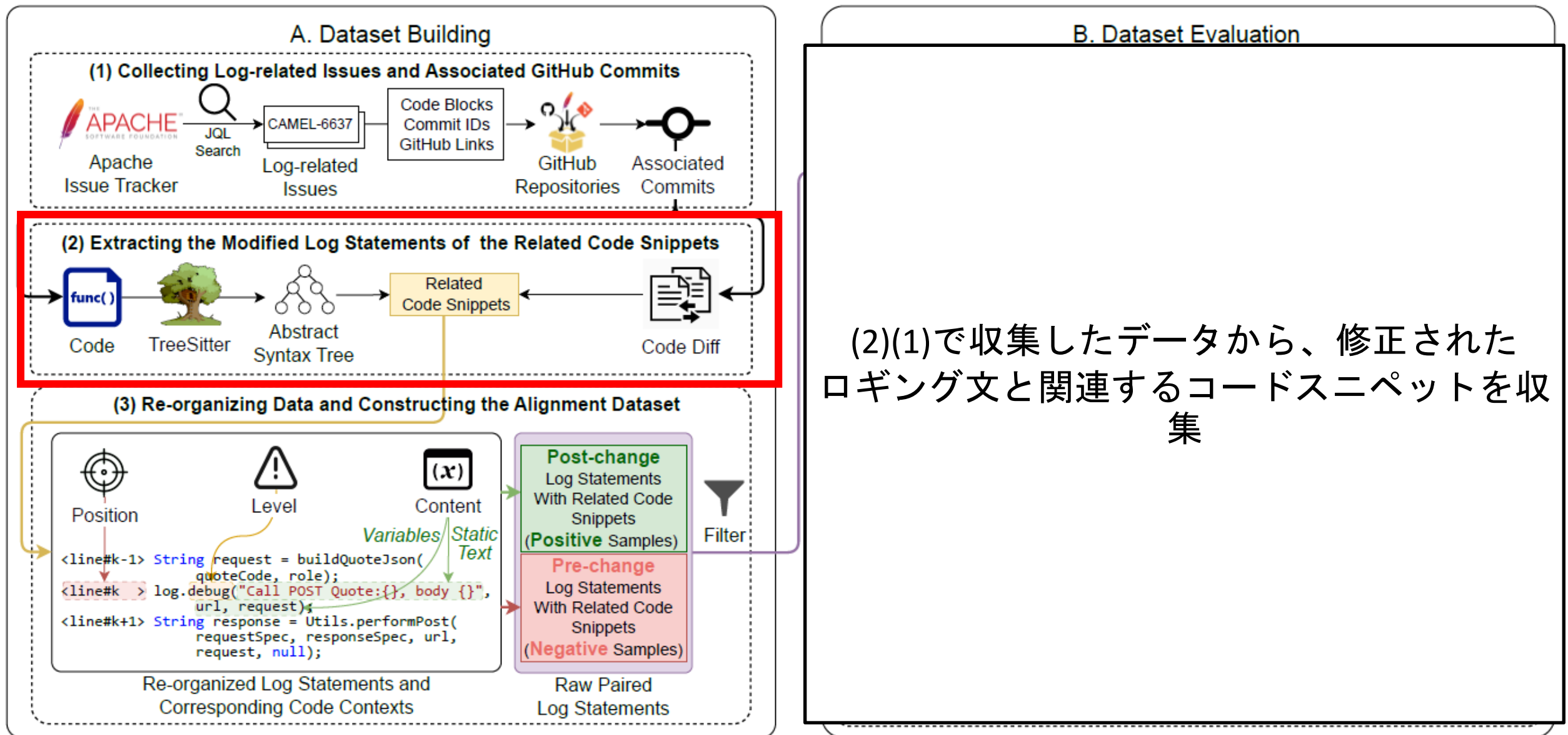
アプローチ：AUCAD概要



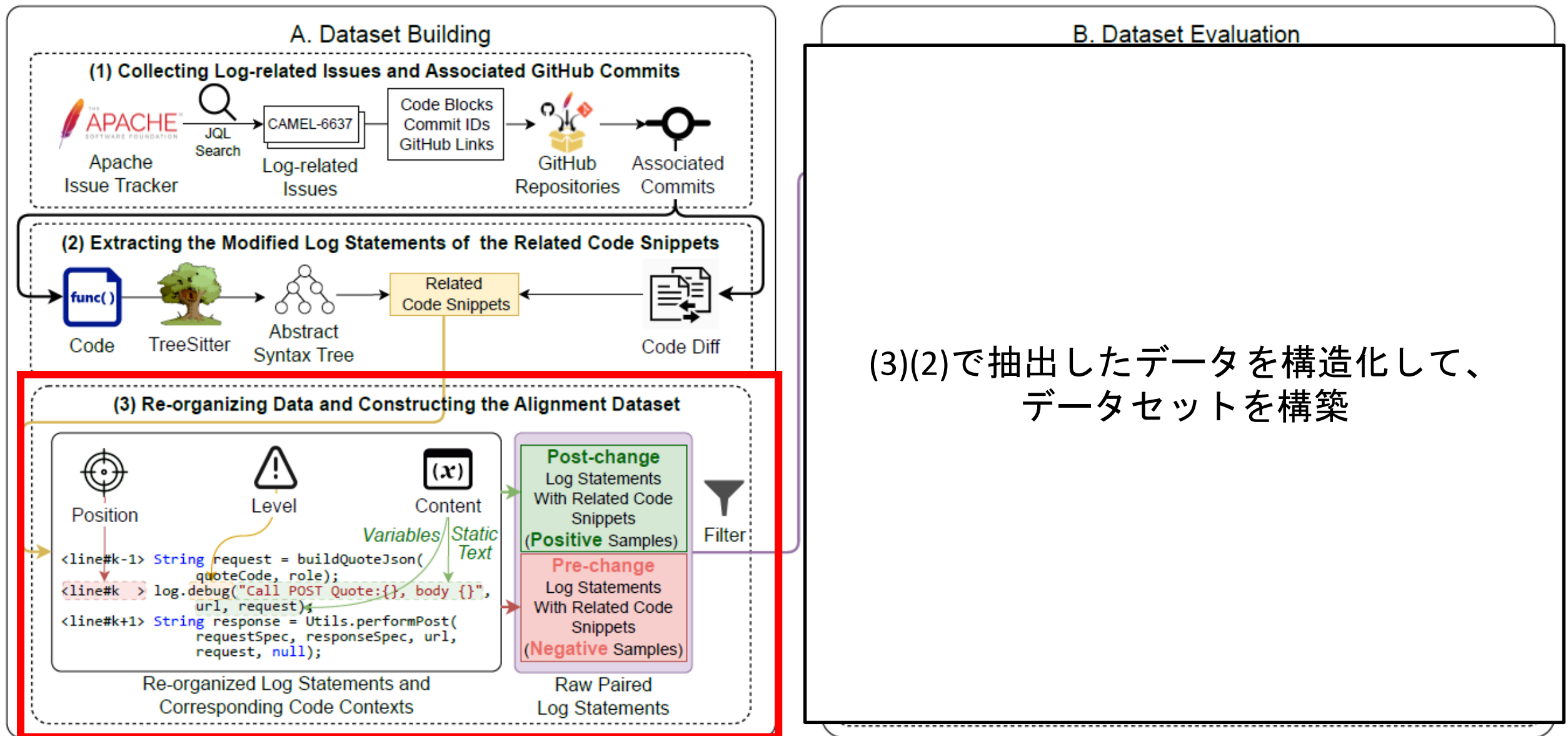
アプローチ：AUCAD概要



アプローチ：AUCAD概要



アプローチ：AUCAD概要



実験：ベースLLMとベースライン

本研究のデータセットで事後学習を行うLLMとして、Llama 3.1とMagicoderを選択

- 大規模なコードデータセットで事前学習が行われているため

評価におけるベースラインとして、以下3つをLlama3.1とMagicoderそれぞれで用意

- 事前学習のみ(=そのまま)
- 事前学習＋Fine-Tuning
- 事前学習＋Fine-Tuning＋アライメント
 - アライメント手法として今回はDPO, cDPO, SimPOを使用
 - ただし、後述する結果においてはSimPOの結果のみ記載(3手法すべての結果は参考資料に記載)

評価：実験設定

- モデルのFTとアライメントにはNVIDIA A100 80GBを使用
- 各メソッドとデータセットの対応関係は以下の通りである

Method	Training & Validation Set	Testing Set
Original Model	N/A	LANCE testing set, LogBench-T
UniLog-w/ warmup	LANCE training set (as examples) and validation set (for warmup)	
FT	Fine-tuning (FT for short): LANCE training & validation set Alignment: AuccadLog	
FT+DPO [ours]		
FT+cDPO [ours]		
FT+SimPO [ours]		

評価：評価指標(1/2)

評価指標として、以下の指標を使用

- PA(位置精度)：ロギング文の生成位置が一致している場合1, そうでない場合0
- LA(レベル精度)：ログレベルが正しく予想された場合1、 そうでない場合0
- Adj PA(調整済みレベル精度): ログレベルが以下の表の白い枠に対応する場合1、 そうでない場合0

Level in response \ Level in ground truth	PARAL	ERROR	WARN	INFO	DEBUG	TRACE
PARAL						
ERROR						
WARN						
INFO						
DEBUG						
TRACE						

 Level must be adjusted

評価：評価指標(2/2)

MA(メッセージ精度):ログテキストに含まれる単語のうち、グラウンドトゥールースと一致する単語の割合

BLEU-DM:以下の式で評価される

$$\text{BLEU-DM} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

VP,VR,VF1(変数精度):それぞれログ変数に限定した精度、再現率、F1スコア

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718 (+9.2%)	0.264 (+40.2%)	0.281 (+24.4%)	0.474 (+30.1%)	0.449 (+25.8%)	0.451 (+28.9%)
Llama3.1+FT+ アライメント	0.635 (+67.4%)	0.591 (+25.0%)	0.745 (+13.0%)	0.275 (+46.1%)	0.293 (+29.9%)	0.488 (+34.2%)	0.468 (+31.1%)	0.468 (+33.8%)
	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Magicode	0.303	0.432	0.559	0.148	0.180	0.285	0.275	0.272
Magicode+FT	0.493 (+62.4%)	0.429 (-0.5%)	0.534 (-4.4%)	0.182 (+23.3%)	0.202 (+11.8%)	0.343 (+20.0%)	0.326 (+18.3%)	0.325 (+19.6%)
Magicode+FT+ アライメント	0.511 (+68.3%)	0.444 (+2.9%)	0.554 (-0.9%)	0.188 (+27.6%)	0.211 (+16.8%)	0.354 (+23.9%)	0.337 (+22.5%)	0.336 (+23.7%)

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718 (+9.2%)	0.264 (+40.4%)	0.281 (+5.3%)	0.474 (+30.2%)	0.449 (+25.2%)	0.451 (+28.9%)
Llama3.1+FT+ アライメント	0.635 (+67.4%)	0.591 (+25.0%)	0.745 (+13.0%)	0.271 (+46.3%)	0.281 (+5.3%)	0.474 (+30.2%)	0.449 (+25.2%)	0.468 (+33.8%)
	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Magocoder	0.303	0.432	0.559	0.141	0.211	0.354	0.337	0.272
Magocoder+FT	0.493 (+62.4%)	0.429 (-0.5%)	0.534 (-4.4%)	0.188 (+33.3%)	0.211 (+11.8%)	0.354 (+20.0%)	0.337 (+18.3%)	0.325 (+19.6%)
Magocoder+FT+ アライメント	0.511 (+68.3%)	0.444 (+2.9%)	0.554 (-0.9%)	0.188 (+33.3%)	0.211 (+16.8%)	0.354 (+23.9%)	0.337 (+22.5%)	0.336 (+23.7%)

ロギング文の位置に関しての
Llamaの性能がFTとアライメント
により67.4%向上している

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718 (+9.2%)	0.264 (+40.0%)	0.281	0.474	0.449	0.451 (+28.9%)
Llama3.1+FT+ アライメント	0.635 (+67.4%)	0.591 (+25.0%)	0.745 (+13.0%)	0.27 (+46.0%)				0.468 (+33.8%)
	PA	LA	Adj. LA					VF1
Magicode	0.303	0.432	0.559	0.14				0.272
Magicode+FT	0.493 (+62.4%)	0.429 (-0.5%)	0.534 (-4.4%)	0.18 (+23.3%)	(+11.8%)	(+20.0%)	(+18.3%)	0.325 (+19.6%)
Magicode+FT+ アライメント	0.511 (+68.3%)	0.444 (+2.9%)	0.554 (-0.9%)	0.188 (+27.6%)	0.211 (+16.8%)	0.354 (+23.9%)	0.337 (+22.5%)	0.336 (+23.7%)

FTのみだとLlamaの性能向上は
50%程度にとどまっている

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718 (+9.2%)	0.264 (+40.2%)	0.281 (+24.4%)	0.474 (+30.1%)	0.449 (+25.8%)	0.451 (+28.9%)
Llama3.1+FT+ アライメント	0.635 (+67.4%)	0.591 (+25.0%)	0.745 (+13.0%)	0.275 (+46.1%)	0.293 (+29.9%)	0.488 (+34.2%)	0.468 (+31.1%)	0.468 (+33.8%)
	PA	LA	Adj. LA					VF1
Magicode	0.303	0.432	0.559	0.1				0.272
Magicode+FT	0.493 (+62.4%)	0.429 (-0.5%)	0.534 (-4.4%)	0.1				0.325 (+19.6%)
Magicode+FT+ アライメント	0.511 (+68.3%)	0.444 (+2.9%)	0.554 (-0.9%)	0.1				0.336 (+23.7%)

Magicodeでも性能が向上しているが、FTとFT+アライメントの差はあまりない

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718 (+9.2%)	0.264 (+40.2%)	0.281 (+24.4%)	0.474	0.449	0.451
Llama3.1+FT+ アライメント	0.635 (+67.4%)	0.591 (+25.0%)	0.745 (+13.0%)	0.275 (+46.1%)	0.293 (+29.9%)			
	PA	LA	Adj. LA	MA	BLEU			
Magicode	0.303	0.432	0.559	0.148	0.180			
Magicode+FT	0.493 (+62.4%)	0.429 (-0.5%)	0.534 (-4.4%)	0.182 (+23.3%)	0.202 (+11.8%)	(+20.0%)	(+18.3%)	(+19.6%)
Magicode+FT+ アライメント	0.511 (+68.3%)	0.444 (+2.9%)	0.554 (-0.9%)	0.188 (+27.6%)	0.211 (+16.8%)	0.354 (+23.9%)	0.337 (+22.5%)	0.336 (+23.7%)

ログレベルに関してのLlamaの
性能はそれぞれ25%、13.3%向上
している

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718 (+9.2%)	0.264 (+40.2%)	0.281 (+24.4%)	0.474	0.449	0.451
Llama3.1+FT+ アライメント	0.635 (+67.4%)	0.591 (+25.0%)	0.745 (+13.0%)	0.275 (+46.1%)	0.293 (+29.9%)			
	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Magicode	0.303	0.432	0.559	0.148	0.180			
Magicode+FT	0.493 (+62.4%)	0.429 (-0.5%)	0.534 (-4.4%)	0.182 (+23.3%)	0.202 (+11.8%)	(+20.0%)	(+18.3%)	(+19.6%)
Magicode+FT+ アライメント	0.511 (+68.3%)	0.444 (+2.9%)	0.554 (-0.9%)	0.188 (+27.6%)	0.211 (+16.8%)	0.354 (+23.9%)	0.337 (+22.5%)	0.336 (+23.7%)

FTのみだとLlamaの性能向上は
それぞれ20.5%、9.2%に
とどまる

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718 (+9.2%)	0.264 (+40.2%)	0.281 (+24.4%)	0.474 (+30.1%)	0.449 (+25.8%)	0.451 (+28.9%)
Llama3.1+FT+ アライメント	0.635 (+67.4%)	0.591 (+25.0%)	0.745 (+13.0%)	0.275 (+46.1%)	0.293 (+29.9%)	0.488 (+34.2%)	0.468 (+31.1%)	0.468 (+33.8%)
	PA	LA	Adj. LA	MA	BLEU	Magicoderではあまり効果が なかった		
Magicoder	0.303	0.432	0.559	0.148	0.180			
Magicoder+FT	0.493 (+62.4%)	0.429 (-0.5%)	0.534 (-4.4%)	0.182 (+23.3%)	0.202 (+11.8%)			
Magicoder+FT+ アライメント	0.511 (+68.3%)	0.444 (+2.9%)	0.554 (-0.9%)	0.188 (+27.6%)	0.211 (+16.8%)			

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718 (+9.2%)	0.264 (+40.2%)	0.281 (+24.4%)	0.474 (+30.1%)	0.449 (+25.8%)	0.451 (+28.9%)
Llama3.1+FT+ アライメント	0.635	0.591	0.745	0.275 (+46.1%)	0.293 (+29.9%)	0.488 (+34.2%)	0.468 (+31.1%)	0.468 (+33.8%)
				MA	BLEU-DM	VP	VR	VF1
Magicode				0.148	0.180	0.285	0.275	0.272
Magicode+FT				0.182 (+23.3%)	0.202 (+11.8%)	0.343 (+20.0%)	0.326 (+18.3%)	0.325 (+19.6%)
Magicode+FT+ アライメント				0.188 (+27.6%)	0.211 (+16.8%)	0.354 (+23.9%)	0.337 (+22.5%)	0.336 (+23.7%)

ログテキストに関して、
FT+アライメントにより
大幅に性能が向上した

結果分析

以下に示す表は、各LLMのロギング文生成能力に関するスコアの一覧表である

- 括弧の内容はベースラインとなる事前学習済みLLMに対してのスコアの増加率

	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1	
Llama3.1	0.380	0.473	0.657	0.188	0.266	0.364	0.357	0.350	
Llama3.1+FT	0.571 (+50.5%)	0.570 (+20.5%)	0.718	0.264	0.281	0.474 (+30.1%)	0.449 (+25.8%)	0.451 (+28.9%)	
Llama3.1+FT+ アライメント	0.635 (+67.4%)	0.591 (+25.0%)	ログ変数も同様に、 FT+アライメント(SimPO) により大幅に性能が向上 した			0.488 (+34.2%)	0.468 (+31.1%)	0.468 (+33.8%)	
	PA	LA				M	VP	VR	VF1
Magicoder	0.303	0.432					0.285	0.275	0.272
Magicoder+FT	0.493 (+62.4%)	0.429 (-0.5%)					0.343 (+20.0%)	0.326 (+18.3%)	0.325 (+19.6%)
Magicoder+FT+ アライメント	0.511 (+68.3%)	0.444 (+2.9%)	0.554 (-0.9%)	0.188 (+27.6%)	0.211 (+16.8%)	0.354 (+23.9%)	0.337 (+22.5%)	0.336 (+23.7%)	

結論

アライメントにより、ファインチューニングされたモデルはログ文生成において大幅な改善を達成した

今後はより広範なシナリオにわたって嗜好アライメントデータセットを構築することを検討すべき

- 広範なシナリオ・・・様々なLLM, Java以外のプログラミング言語

また、アライメントの際に与えるコンテキストの範囲に関しても更なる議論の余地がある

余談

疑問：今回の形式のアライメントデータセットを作るなら， Issueを探さずとも， 直接GitHubのリポジトリをマイニングすればいいのでは？

解答：この研究は元々アライメントをする研究ではなく， 高品質なデータセットを作る研究だったためこのような回りくどいことをしている．

- 元々の研究はこのデータセットでFine-Tuningを行うという研究だった
- それが論文採択にあたり， Fine-Tuningからアライメントに代わっていた

所感

アライメント＋高品質データセットの組み合わせだったからこそ性能が伸びた、とは言えないのでは？

- この結果だけだと、適当に拾ってきたロギング文差分データを使ってアライメントするだけでも性能が向上する可能性は全然あり得る
- その逆で十分な量の高品質データセットでFTしても性能が向上する可能性もある

上記の謎が解けないと、今回のデータセットが高品質であるかどうかすらも議論ができない

參考資料

ロギング文の構成要素

ロギング文の構成要素として、**ログレベル**、**ログテキスト**、**ログ変数**の3つが挙げられる。

- **ログレベル(Level)** : そのログの重要度を示すもの。
 - TRACE, INFO, ERROR, ...といったレベルが与えられる。
- **ログテキスト(Message)** : そのログの内容を説明する部分。
- **ログ変数(Variable)** : ログテキスト中で使われる変数。

The diagram shows the following code snippet with annotations:

```
logger.debug("Debugging user: username=" + username + ", id=" + userId);
```

Annotations and brackets:

- ログレベル** (red text) points to `debug`.
- ログテキスト** (blue text) points to the entire string `"Debugging user: username=" + username + ", id=" + userId`.
- ログ変数** (green text) points to `username` and `userId` via separate brackets.

結果分析：RQ2

以下に今回の実験の結果を示す

Method (Backbone)	PA	LA	Adj. LA	MA	BLEU-DM	VP	VR	VF1
<i>Results of the LANCE testing set:</i>								
LANCE [42]	0.546	0.378	0.546	0.100	0.110	0.271	0.253	0.254
FastLog [59]	0.527	0.410	0.505	0.156	0.174	0.228	0.234	0.224
UniLog [61] (Llama 3.1)	0.523	0.522	0.674	0.239	0.255	0.356	0.348	0.345
UniLog [61] (Magicoder)	0.116	0.213	0.315	0.077	0.081	0.121	0.119	0.119
Llama 3.1	0.380	0.473	0.657	0.188	0.226	0.364	0.357	0.350
FT (Llama 3.1)	0.571 (+ 50.5%)	0.570 (+20.5%)	0.718 (+ 9.2%)	0.264 (+40.2%)	0.281 (+24.4%)	0.474 (+30.1%)	0.449 (+25.8%)	0.451 (+28.9%)
FT+DPO (Llama 3.1) [ours]	0.634 (+67.2%)	0.581 (+22.8%)	0.743 (+13.0%)	0.273 (+45.0%)	0.292 (+29.4%)	0.484 (+33.1%)	0.467 (+30.7%)	0.465 (+32.9%)
FT+cDPO (Llama 3.1) [ours]	0.635 (+67.2%)	0.586 (+23.8%)	0.738 (+12.4%)	0.274 (+45.6%)	0.291 (+29.1%)	0.486 (+33.7%)	0.466 (+30.6%)	0.466 (+33.2%)
FT+SimPO (Llama 3.1) [ours]	0.635 (+67.4%)	0.591 (+25.0%)	0.745 (+13.3%)	0.275 (+46.1%)	0.293 (+29.9%)	0.488 (+34.2%)	0.468 (+31.1%)	0.468 (+33.8%)
Magicoder	0.303	0.432	0.559	0.148	0.180	0.285	0.275	0.272
FT (Magicoder)	0.493 (+62.4%)	0.429 (- 0.5%)	0.534 (- 4.4%)	0.182 (+23.3%)	0.202 (+11.8%)	0.343 (+20.0%)	0.326 (+18.3%)	0.325 (+19.6%)
FT+DPO (Magicoder) [ours]	0.507 (+67.3%)	0.444 (+ 2.8%)	0.552 (- 1.2%)	0.189 (+27.7%)	0.210 (+16.5%)	0.351 (+23.0%)	0.334 (+21.5%)	0.334 (+22.7%)
FT+cDPO (Magicoder) [ours]	0.510 (+68.0%)	0.443 (+ 2.6%)	0.554 (- 0.9%)	0.190 (+28.8%)	0.211 (+17.1%)	0.355 (+24.5%)	0.339 (+23.1%)	0.338 (+24.3%)
FT+SimPO (Magicoder) [ours]	0.511 (+68.3%)	0.444 (+ 2.9%)	0.554 (- 0.9%)	0.188 (+27.6%)	0.211 (+16.8%)	0.354 (+23.9%)	0.337 (+22.5%)	0.336 (+23.7%)
<i>Results of the LogBench-T dataset:</i>								
LANCE [42]	0.506	0.356	0.539	0.059	0.071	0.341	0.333	0.332
FastLog [59]	0.493	0.429	0.571	0.095	0.118	0.330	0.339	0.329
UniLog [61] (Llama 3.1)	0.347	0.442	0.671	0.106	0.132	0.334	0.333	0.330
UniLog [61] (Magicoder)	0.158	0.210	0.345	0.041	0.049	0.141	0.140	0.139
Llama 3.1	0.411	0.503	0.708	0.129	0.154	0.396	0.393	0.388
FT (Llama 3.1)	0.592 (+44.1%)	0.529 (+ 5.3%)	0.708 (- 0.9%)	0.129 (+28.9%)	0.154 (+ 7.2%)	0.441 (+11.4%)	0.426 (+ 8.4%)	0.427 (+10.0%)
FT+DPO (Llama 3.1) [ours]	0.639 (+55.4%)	0.528 (+ 5.0%)	0.763 (+ 6.9%)	0.138 (+37.6%)	0.167 (+16.4%)	0.450 (+13.7%)	0.438 (+11.3%)	0.437 (+12.6%)
FT+cDPO (Llama 3.1) [ours]	0.640 (+55.7%)	0.525 (+ 4.3%)	0.767 (+ 7.4%)	0.140 (+39.9%)	0.168 (+16.8%)	0.459 (+15.8%)	0.445 (+13.3%)	0.445 (+14.6%)
FT+SimPO (Llama 3.1) [ours]	0.642 (+56.1%)	0.536 (+ 6.5%)	0.769 (+ 7.7%)	0.140 (+39.8%)	0.168 (+16.6%)	0.464 (+17.1%)	0.451 (+14.7%)	0.450 (+16.0%)
Magicoder	0.250	0.409	0.492	0.072	0.107	0.282	0.276	0.274
FT (Magicoder)	0.475 (+90.0%)	0.432 (+ 5.8%)	0.573 (+16.5%)	0.098 (+35.4%)	0.124 (+16.2%)	0.338 (+20.1%)	0.328 (+19.1%)	0.327 (+19.5%)
FT+DPO (Magicoder) [ours]	0.495 (+97.8%)	0.448 (+ 9.6%)	0.595 (+20.8%)	0.108 (+49.5%)	0.131 (+22.8%)	0.347 (+23.1%)	0.337 (+22.1%)	0.335 (+22.4%)
FT+cDPO (Magicoder) [ours]	0.493 (+97.0%)	0.442 (+ 8.3%)	0.593 (+20.5%)	0.106 (+47.0%)	0.132 (+23.3%)	0.350 (+24.2%)	0.340 (+23.2%)	0.338 (+23.4%)
FT+SimPO (Magicoder) [ours]	0.493 (+97.2%)	0.449 (+ 9.9%)	0.595 (+20.8%)	0.105 (+45.2%)	0.130 (+21.5%)	0.351 (+24.4%)	0.340 (+23.3%)	0.338 (+23.6%)