

**実験 1 (RQ1) : LLMに障害情報を含めることで、出力結果に対し、どのように影響するか**

**検証内容** : LLMに以下2条件で、原因と対策の提示が可能かを検証する。

- ①障害情報を与えない場合      ②障害情報を与えた場合

**<検証結果>**

①障害情報を与えない場合

**■プロンプト**  
要件確定が遅く、要件変更も多く、要件の設計への反映も正確に確認できていなかった原因を教えてください。  
そして、その原因に対する対策を教えてください。

**■出力結果**  
要件確定が遅く、要件変更が多く、要件の設計への反映が正確に確認できていなかった原因は、いくつか考えられます。  
以下に一般的な原因とその対策を示します。

**原因**

1. コミュニケーション不足  
⇒ステークホルダー間のコミュニケーションが不足していると、要件の理解にズレが生じることがあります。
2. 不十分な要件定義  
⇒初期の要件定義が曖昧だったり、不完全だったりすることで、後から変更が必要になることがあります。
3. 変化するビジネス環境  
⇒ビジネス環境や市場の変化により、要件が変更されることがあります。
4. 意思決定の遅延  
⇒意思決定に時間がかかると、その間に要件が確定しないことがあります。
5. 確認プロセスの欠如  
⇒要件が設計に正確に反映されているかを確認するためのプロセスやツールが不足していると、ミスが発生しやすくなります。

**対策**

1. 効果的なコミュニケーションの確立  
⇒定期的なミーティングを開催し、ステークホルダー間のコミュニケーションを促進します。  
ツールを活用してリアルタイムで情報共有を行うことも効果的です。
2. 要件定義の強化  
⇒要件定義の段階で詳細なドキュメントを作成し、ステークホルダー全員の合意を得るようにします。  
プロトタイプやワークショップを活用して、要件を具体化することも有効です。
3. アジャイル開発の導入  
⇒アジャイル開発手法を導入することで、変化に迅速に対応しやすくなります。  
スプリントレビューやバックログの見直しを定期的に行うことで、要件の変更に柔軟に対応します。
4. 意思決定プロセスの明確化  
⇒意思決定の責任者を明確にし、迅速な決定が行えるようにプロセスを整備します。
5. 確認とテストの強化  
⇒要件が設計に正しく反映されていることを確認するためのレビューやテストプロセスを強化します。  
テストケースやチェックリストを用意し、システムティックに確認を行います。  
これらの対策を組み合わせることで、要件確定の遅延や変更の頻度を減らし、  
設計への正確な反映を実現できる可能性が高まります。

## ②障害情報を与えた場合

### ■プロンプト

以下の事例をもとに、要件確定が遅く、要件変更も多く、要件の設計への反映も正確に確認できていなかった原因を教えてください。そして、その原因に対する対策を教えてください。

### ■事例

#### 問題

Windows95の登場、PCの廉価化、インターネットの普及等をきっかけに、90年代後半から企業の活動にITが不可欠となるだけでなく、競争の激化にともない、商品の複雑化が進み、すべての商品がシステム開発をともなう形で出されることとなり、従来とはレベルの違う量のシステム開発が行われるようになってきた。しかしながら、急激に増大したシステム開発を成功させるためのポイントも十分掴めないまま実施したことにより、システムトラブルが多発し、企業の本来活動に支障をきたし、経営問題となった。

#### 原因

コンサルタントも入れた分析の結果、システムトラブルの8割は、上流の要件定義局面でのコミュニケーション・ギャップから問題が生じていることが判明した。コンサルタントからは、以下のような問題事象も指摘された。

① ビジネス側の要件の確定が遅い（期日までに決めなくてはいけないというマインドが乏しい）。

② 要件の変更が多い。

③ 要件を最終的に文書で確認していない。

④ その要件が他システムにどのような影響を与えるかの分析が甘い。

⑤ 要件が設計に正しく反映されたかを複数の眼でチェックしていない。

④、⑤については、情報システム部門内での仕事の進め方に関する問題であり、

これらについても抜本的な改善が必要であるが、①～③については、

事業部門が「システム開発を情報システム部門の仕事」として任せきりにするのではなく、

事業部門も情報システム開発において一定の役割と責任を果たすようにしないと、本質的に解決できない問題であると考えた。

#### 対策

この問題を解決するために、システム開発におけるビジネスサイドの役割と責任を明確化し、

コミュニケーションの質を高めることとした。

そのためには、お互いの理解を補正し合い、「言いたいこと」と「聞いて理解したこと」が同じになるまで、

対等な立場でオープンなダイアログを繰り返すことができる「態勢」づくりが必要である。

この態勢を「アプリケーション・オーナー制度」と呼ぶ。アプリケーション・オーナー制度のポイントは以下のとおり。

① システム開発は、情報システム部門に任せきりにすべき仕事ではなく、

自分の考えた商品や施策を具体化するために行う自分自身の仕事であるという「オーナーシップ」の考え方を持たせる。

② 事業部門に、要件の詳細が固まるまで、情報システム部門と対話を繰り返す責任を持たせ、要件定義の最終責任を負わせる。

③ 事業部門に、要件定義どおりにシステムが出来たかどうか受入れテストを実施する責任を負わせる。

→「事業部門は要件定義に責任を持つこと」とするだけでは、表面的な責任に留まり、要件の揺戻し等の問題は解消しないので、要件を定義した以上、その要件どおりシステムが出来たかどうかを確認するUAT（User Acceptance Test、ユーザ受入れテスト）のテストケースを作ってテストすることが重要である。実際に手を動かす（ハンズオン）の責任にしない限り、本当に責任を取ることにしない。

因みに、この制度に基づく、東京海上日動火災保険株式会社の情報システム部門とソフトベンダとの契約関係は以下のとおりである。

この責任態勢は、アプリケーション・オーナーが要件定義～外部設計の工程、

及びシステムテスト～運用テストの工程に責任を持つことで裏打ちされている。

### ■出力結果

この事例における要件確定の遅れや要件変更の多さ、要件の設計への反映の不備は、以下のような原因が考えられます。

#### 原因

1. コミュニケーション不足:

⇒ビジネス側と情報システム部門の間でのコミュニケーションが不十分で、要件の理解にギャップが生じていた。

ビジネス部門が情報システム部門に要件を丸投げし、相互の理解が不足していた。

2. ビジネス側の関与不足:

⇒ビジネス部門がシステム開発における自分たちの役割や責任を十分に認識しておらず、積極的に関与していなかった。

3. 要件管理の不備:

⇒要件が文書で明確に確認されておらず、変更管理も適切に行われていなかった。

他システムへの影響分析が不十分で、全体最適が図られていなかった。

4. 品質管理の不備:

⇒要件が設計に正しく反映されているかの確認が十分に行われていなかった。

#### 対策

1. アプリケーション・オーナー制度の導入:

⇒ビジネス部門にオーナーシップを持たせ、要件定義から受入れテストまでの責任を明確にする。

ビジネス側が情報システム部門と継続的に対話し、要件が詳細に固まるまで確認を行う。

2. コミュニケーションの強化:

⇒対等な立場でのオープンなダイアログを促進し、言いたいことと理解したことが一致するまで繰り返し確認する。

定期的なミーティングやワークショップを通じて、相互理解を深める。

3. 要件管理プロセスの整備:

⇒要件を文書化し、変更があればトレーサビリティを持たせる。他システムへの影響を考慮した分析を行い、全体最適を図る。

4. 品質管理の強化:

⇒要件が設計に正しく反映されているかを複数の視点からチェックするプロセスを導入する。

ユーザ受入れテスト（UAT）のテストケースを作成し、実際にビジネス部門が確認する。

これらの対策を講じることで、要件確定の遅れや変更の多さを改善し、システム開発プロジェクトの成功率を向上させることが期待できます

**実験 2 (RQ2) : RAGにおける検索性能によって、出力結果に対しどのように影響するか**

**検証内容** : 以下2種のデータ使用時、検索Queryのキーワードの出現率を検証する。

- ①PDFのデータ使用時
- ②テキストのデータ使用時

**<検証結果>**

■プロンプト  
 保守作業での設定ミスで待機系側に必要ファイルが存在せず、障害時に切替失敗の原因を教えてください。  
 その原因に対する対策を教えてください。

①PDFのデータ使用時

表 1. PDFデータ使用時、検索Queryのキーワード出力結果

■検索Queryのキーワード出力結果														
試行回数	保守作業	設定ミス	必要ファイル	不存在	障害時	切替失敗	原因	対策	待機系側	事例	待機系ファイル不足	※1	※2	※3
1	○	○	○	○	○	○	○	○	○	○	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	○	○	○
3	○	○	○	○	○	○	○	○	○	○	-	-	-	-
4	○	○	○	○	○	○	○	○	-	○	○	-	-	-
5	○	○	○	○	○	○	○	○	○	○	○	-	-	-
合計	4	4	4	4	4	4	4	4	3	4	2	1	1	1

- ※1 : 設定ミスによる切替失敗の原因と対策
- ※2 : 待機系に必要なファイルがない場合の切替失敗事例
- ※3 : 保守作業での切替失敗の原因と対策

②テキストのデータ使用時

表 2. テキストデータ使用時、検索Queryのキーワード出力結果

■検索Queryのキーワード出力結果														
試行回数	保守作業	設定ミス	必要ファイル	不存在	障害時	切替失敗	原因	対策	待機系側	事例	待機系ファイル不足	※1	※2	※3
1	-	-	-	-	-	-	-	-	-	-	-	○	○	○
2	-	-	-	-	-	-	-	-	-	-	-	○	○	○
3	○	○	○	○	○	○	○	○	○	○	-	-	-	-
4	-	-	-	-	-	-	-	-	-	-	-	○	○	○
5	-	-	-	-	-	-	-	-	-	-	-	○	○	○
合計	1	1	1	1	1	1	1	1	1	1	0	4	4	4

- ※1 : 設定ミスによる切替失敗の原因と対策
- ※2 : 待機系に必要なファイルがない場合の切替失敗事例
- ※3 : 保守作業での切替失敗の原因と対策

表 3. 各データ使用時、検索Queryのキーワード出力結果まとめ

プロンプトへ入力した事例	検索Query	検索Query発生回数		考察		
		テキスト	PDF	検索Query	事例との一致性	
保守作業での設定ミスで待機系側に必要ファイルが存在せず 障害時に切替失敗した事例の原因と対策を教えてください	保守作業	1/5回	4/5回	特徴的なキーワード	具体的事例と一致し易い	
	設定ミス	1/5回	4/5回			
	必要ファイル	1/5回	4/5回			
	不存在	1/5回	4/5回			
	障害時	1/5回	4/5回			
	切替失敗	1/5回	4/5回			
	原因	1/5回	4/5回			
	対策	1/5回	4/5回			
		待機系側	1/5回	3/5回	一般的キーワード	一般的事例と一致し易い
	事例	1/5回	4/5回			
	待機軽ファイル不足	1/5回	2/5回			
		設定ミスによる切替失敗	4/5回	1/5回	文章的なキーワード	「原因」とは一致し易い 「対策」とは一致し難い
	待機系に必要なファイルがない	4/5回	1/5回			
	保守作業での切替失敗	4/5回	1/5回			

**実驗3 (RQ3) : 障害情報内に複数の現象が含まれる場合、出力結果に対しどのように影響するか**

**検証内容** : 現象①～③を含むプロンプトを使った試行を、各5回実施し、その出力結果から、原因と対策内容を比較し、データ形式や検索Queryの相違点を確認する。

**<検証結果>**

事例G11を使用し、現象①～③を含むプロンプトを使った試行を、各5回実施したスコア一覧(表4. 現象①～③別のスコア)

表4. 現象①～③別のスコア

現象	プロンプト内容	実行回数	PDF				TXT			
			原因		対策		原因		対策	
			評価	合計	評価	合計	評価	合計	評価	合計
①	保守作業での設定ミスで待機系側に必要ファイルが存在せず、障害時に切替失敗した原因と対策を教えてください	1回目	×		×		○		×	
		2回目	×		×		×		×	
		3回目	○	0.4	×	0	×	0.8	×	0
		4回目	×		×		○		×	
		5回目	×		×		×		×	
②	ディスクモジュールの自己診断機能で、異常検知のみで機能停止する仕様となっていた事例の原因と対策を教えてください	1回目	○		○		○		○	
		2回目	○		○		○		○	
		3回目	○	2	○	2	○	2	○	2
		4回目	○		○		○		○	
		5回目	○		○		○		○	
③	保守作業チェックや修正プログラムの適用がシステムによらず一律となっており、重要なミスを見落とした事例の原因と対策を教えてください	1回目	○		○		○		○	
		2回目	△		△		△		△	
		3回目	○	1.4	○	1.6	△	1.6	△	1.2
		4回目	△		△		○		△	
		5回目	△		△		○		△	

○ : G11の障害データのみ検出 (2点)  
 △ : G11を含む複数の障害データを検出 (1点)  
 × : G11未検出 (0点)

現象①～③のプロンプトを各5回ずつ、2種(PDF/テキスト)のデータ使用時の出力結果

現象①

■プロンプト	
保守作業での設定ミスで待機系側に必要ファイルが存在せず、障害時に切替失敗の原因を教えてください。	
その原因に対する対策を教えてください。	

**<PDFのデータ使用時> ※各表にて、想定事例の出力結果は黄色網掛け**

■原因

表5. 現象①のPDFデータ使用時、原因内容一覧

原因内容	事例	試行回数(単位: 回目)				
		1	2	3	4	5
1. 独立設定の不備	T7	○	○	-	○	○
2. パラメータ更新の失敗	T7	-	-	-	○	-
3. 必要なファイルの不備	T7	-	○	-	-	○
4. ミラーリングの誤設定	G11	-	-	○	-	-

■対策

表6. 現象①のPDFデータ使用時、対策内容一覧

原因内容	事例	試行回数(単位: 回目)				
		1	2	3	4	5
1. 定期的な訓練と確認	T7	○	-	-	-	○
2. 手順書の整備と確認	T7	○	○	○	○	○
3. 原因分析と再発防止策の実施	T7	○	-	○	-	-
4. ソフトウェアパラメータの一元管理	T7	-	○	-	○	-
5. 事前の障害訓練	T7	-	-	○	○	-
6. 同期処理の強化	T7	-	-	-	-	○

**<テキストのデータ使用時> ※各表にて、想定事例の出力結果は黄色網掛け**

■原因

表7. 現象①のテキストデータ使用時、原因内容一覧

原因内容	事例	試行回数(単位: 回目)				
		1	2	3	4	5
1. ミラーリングの誤設定	G11	○	-	-	○	-
2. 入出力制御機能の不具合	G11	○	-	-	-	-
3. ソフトウェアパラメータ設定の怠慢	T7	-	○	○	-	○
4. 独自設定の不備	T7	-	○	○	-	○
5. 保守作業指示書の不備	G11	-	-	-	○	-

■対策

表8. 現象①のテキストデータ使用時、対策内容一覧

原因内容	事例	試行回数(単位: 回目)				
		1	2	3	4	5
1. 待機系装置からの停止制御機能の追加	T2	○	-	-	-	-
2. 上位制御装置からの停止機能の強化	T2	○	-	-	-	-
3. 監視機能の強化	T2	○	-	-	-	○
4. ソフトウェアパラメータの一元管理	T7	-	○	○	-	-
5. 事前の障害訓練	T7	-	○	○	○	○
6. 監視機能の強化	T7	-	-	○	-	-
7. 設定の厳密化	T9	-	-	-	○	○
8. 体制強化	T9	-	-	-	○	-

**<検索Queryのキーワードごとの出力結果>**

表9. 各データ使用時、検索Queryのキーワード出現回数一覧

プロンプト内容	データ形式	試行回数	出現回数														
			作業	保守	ミス	設定	必要ファイル	不存在	障害時	失敗	切替	原因	対策	系待機	事例	ファイル不足	待機系
保守作業での設定ミスで待機系側に必要ファイルが存在せず、障害時に切替失敗した事例の原因と対策を教えてください。	PDF	5回	4	4	4	4	4	4	4	4	4	3	4	2	1	1	1
	TXT		1	1	1	1	1	1	1	1	1	1	1	0	4	4	4

※1 設定ミスによる切替失敗の原因と対策  
 ※2 待機系に必要なファイルがない場合の切替失敗事例  
 ※3 保守作業での切替失敗の原因と対策



現象③

■プロンプト  
 保守作業チェックや修正プログラムの適用がシステムによらず一律となっており、  
 重要なミスを見落としした事例の原因と対策を教えてください

<PDFのデータ使用時> ※各表にて、想定事例の出力結果は黄色網掛け

■原因

表 1 5. 現象③のPDFデータ使用時、原因内容一覧

原因内容	事例	試行回数(単位:回目)				
		1	2	3	4	5
1. システムの重要度に応じた適用範囲の見直し不足	G11	○	-	○	-	○
2. チェック体制の不備	G11	○	○	○	○	-
3. 情報共有の不足	G11	○	○	○	-	-
4. テスト環境と本番環境の差異	T6	-	○	-	-	-
5. 作業環境の問題	T21	-	-	-	○	○

<テキストのデータ使用時> ※各表にて、想定事例の出力結果は黄色網掛け

■原因

表 1 7. 現象③のテキストデータ使用時、原因内容一覧

原因内容	事例	試行回数(単位:回目)				
		1	2	3	4	5
1. チェック体制の不備	G11	○	○	○	○	-
2. システムの重要度に応じた適用範囲の見直し不足	G11	○	-	-	-	-
3. テスト環境と本番環境の差異	T6	-	○	○	-	-
4. 情報共有の不足	G11	-	-	-	○	○
5. 修正プログラムの情報不足	G11	-	-	-	-	○
6. ミラーリングの誤設定	G11	-	-	-	-	○

■対策

表 1 6. 現象③のPDFデータ使用時、対策内容一覧

原因内容	事例	試行回数(単位:回目)				
		1	2	3	4	5
1. システムの重要度に応じた適用範囲の見直し	G11	○	○	○	-	○
2. 二重チェック体制の導入	G11	○	○	○	○	-
3. チェック体制の強化	T6	-	○	-	-	-
4. 情報共有の不足	G11	-	-	○	-	-
5. ルールの明文化と適用範囲の拡大	G11	-	-	-	○	-
6. システム障害訓練の実施	G11	-	-	-	○	-
7. テスト環境と本番環境の差異の把握	T21	-	-	-	-	○

■対策

表 1 8. 現象③のテキストデータ使用時、対策内容一覧

原因内容	事例	試行回数(単位:回目)				
		1	2	3	4	5
1. 二重チェック体制の導入	G11	○	-	○	-	-
2. システムの重要度に応じた適用範囲の見直し	G11	○	-	-	-	-
3. 環境差異の明確化と分析	T6	-	○	○	-	-
4. リスク分析と対策の共有	T6	-	○	○	-	-
5. 作業者の確認手順の強化	T27	-	○	-	-	-
6. チェック体制の強化	T6	-	-	○	○	-
7. 事前の影響範囲調査の強化	T5	-	-	-	○	-
8. 情報共有の改善	G11	-	-	-	○	○
9. 共通資料の準備	G11	-	-	-	-	○
10. システム障害訓練の実施	G11	-	-	-	-	○

<検索Queryのキーワードごとの出力結果>

表 1 9. 各データ使用時、検索Queryのキーワード出現回数一覧

プロンプト内容	データ形式	試行回数	出現回数												
			対 策	原 因 と ミ ス	重 要 な 見 落 と し	事 例	チ ェ ッ ク	保 守 作 業	グ ラ ム	修 正 プ ロ グ ラ ム	適 用	一 律	保 守 シ ス テ ム	※1	※2
保守作業チェックや修正プログラムの適用がシステムによらず一律となっており、重要なミスを見落としした事例の、原因と対策を教えてください	PDF	5回	4	4	1	4	4	4	3	4	3	1	1	1	0
	TXT		3	2	0	3	3	3	2	0	0	3	2	0	3

※1 保守作業チェックのミス事例と原因

※2 修正プログラムの適用ミスの原因と対策

※3 システムメンテナンスにおける重要ミスの防止策

※4 システム保守作業の見落とし原因と対策

**実験4 (RQ4) : RAGに蓄積したデータの形式と構造によって、出力結果に対しどのように影響するか**

**検証内容** : RAGに蓄積したデータ形式と構造の違いが出力結果の精度に影響すると仮説を立てて検証する

**対象** : RAGパイプラインの評価フレームワークであるRAGAS (Retrieval Augmented Generation Assessment)に、障害情報に関するドキュメントを読み込ませ、ドキュメント情報を利用して生成したテスト (User Input (質問), Reference (正解の回答)) を、RAGシステムに渡して返ってくる回答 (Response) と参照ドキュメント (Retrieved Context) を合わせた情報をテストデータセットとして、測定対象とした。

**測定方法** : RAGの回答と参照ドキュメントを、入力に使用したテストデータと合わせて、RAGASに渡して、各テストケースにおけるRAGを評価する指標のScore (Context Precision, Context Entity Recall, Context Recall, Answer Relevancy, Faithfulness) を算出した。

**統計 :**

- ・ 全般的事項  
解析対象とする全てのデータについて、記述的な解析を行った。連続変数の要約統計量は平均値と標準偏差・標準誤差、およびデータ数とした。検定時の有意水準は  $\alpha=0.05$  (両側) とし  $P<0.05$  にて統計学的に有意とした。区間推定は両側信頼区間とし、信頼係数は95%を用いた。検定の多重性に伴う P 値補正は Dunnett 法とTukey法を用いた。
- ・ 主要評価項目  
データを可視化したグラフを確認し、RAGAS Scoreを正規分布に従うと推定し、パラメトリックな手法を第一選択とした。4群間でのRAGAS Scoreの差を検定する目的で、一元配置分散分析を使用した。また、各群間でのRAGAS Scoreの差を検定する目的で、多重比較検定を使用した。データ分析には、Microsoft Excel for Microsoft 365 MSO (Ver. 2202) 64bitを使用した。

**<検証結果>**

- ① 一元配置分散分析により、Context Precision区分によるRAGAS Score の平均値に統計学的有意な差があると認められた ( $P<0.05$ ) 。  
⇒ 「表2.1. Context Precision」 参照
- ② 一元配置分散分析により、Context Entity Recall区分によるRAGAS Score の平均値に統計学的有意な差があると認められた ( $P<0.05$ ) 。  
⇒ 「表2.2. Context Entity Recall」 参照
- ③ 多重比較検定により、PDFとL2のContext Precisionの平均値に統計学的に有意な差が認められた ( $P<0.05$ ) 。  
⇒ 「表2.6. PDFをコントロール群としたDunnett検定」 参照
- ④ 多重比較検定により、PDFとL3のContext Precisionの平均値に統計学的に有意な差が認められた ( $P<0.05$ ) 。  
⇒ 「表2.6. PDFをコントロール群としたDunnett検定」 参照
- ⑤ 多重比較検定により、Context Precision におけるL2とL3の平均値に統計学的に有意な差は認められなかった ( $P>0.05$ ) 。  
⇒ 「表2.6. PDFをコントロール群としたDunnett検定」 参照
- ⑥ 多重比較検定により、L1とPDFのContext Entity Recallの平均値に統計学的に有意な差が認められた ( $P<0.05$ ) 。  
⇒ 「表2.8. L1をコントロール群としたDunnett検定」 参照
- ⑦ 多重比較検定により、L1とL2のContext Entity Recallの平均値に統計学的に有意な差が認められた ( $P<0.05$ ) 。  
⇒ 「表2.8. L1をコントロール群としたDunnett検定」 参照
- ⑧ 多重比較検定により、L1とL3のContext Entity Recallの平均値に統計学的に有意な差が認められた ( $P<0.05$ ) 。  
⇒ 「表2.8. L1をコントロール群としたDunnett検定」 参照

Context Precisionにおいて、Dunnett法の数値表の該当値、2.364と比較して、L2統計量2.568とL3統計量2.520が大きいため、 $P<0.05$ であり統計学的な有意差が認められた。Context Entity Recallにおいて、Dunnett法の数値表の該当値、2.364と比較して、PDF統計量2.850とL2統計量2.675とL3統計量2.511が大きいため、 $P<0.05$ であり統計学的な有意差が認められた。データは、Mean±SDで示した。

表2.0. RAGAS Scoreの一元配置分散分析結果: Mean (SD)

区分	Context Precision <sup>①</sup> (N = 83) <sup>②</sup>	Context Entity Recall <sup>②</sup> (N = 83) <sup>②</sup>	Context Recall <sup>③</sup> (N = 83) <sup>②</sup>	Answers <sup>④</sup> Relevancy <sup>④</sup> (N = 83) <sup>②</sup>	Faithfulness <sup>⑤</sup> (N = 83) <sup>②</sup>
PDF <sup>③</sup>	0.715 (0.104) <sup>③</sup>	0.347 (0.055) <sup>③</sup>	0.803 (0.117) <sup>③</sup>	0.678 (0.020) <sup>③</sup>	0.887 (0.036) <sup>③</sup>
L1 <sup>③</sup>	0.635 (0.110) <sup>③</sup>	0.245 (0.050) <sup>③</sup>	0.840 (0.103) <sup>③</sup>	0.673 (0.022) <sup>③</sup>	0.886 (0.037) <sup>③</sup>
L2 <sup>③</sup>	0.580 (0.116) <sup>③</sup>	0.341 (0.052) <sup>③</sup>	0.826 (0.108) <sup>③</sup>	0.685 (0.019) <sup>③</sup>	0.865 (0.038) <sup>③</sup>
L3 <sup>③</sup>	0.582 (0.127) <sup>③</sup>	0.335 (0.059) <sup>③</sup>	0.833 (0.117) <sup>③</sup>	0.685 (0.017) <sup>③</sup>	0.883 (0.034) <sup>③</sup>

■ 項目別の分散分析表

表2.1. Context Precision

変動要因 <sup>①</sup>	変動 <sup>②</sup>	自由度 <sup>③</sup>	分散 <sup>④</sup>	観測された分散比 <sup>⑤</sup>	P-値 <sup>⑥</sup>	F 境界値 <sup>⑦</sup>
グループ間 <sup>⑧</sup>	0.996 <sup>⑧</sup>	3 <sup>⑧</sup>	0.332 <sup>⑧</sup>	2.896 <sup>⑧</sup>	0.035 <sup>⑧</sup>	2.632 <sup>⑧</sup>
グループ内 <sup>⑧</sup>	37.594 <sup>⑧</sup>	328 <sup>⑧</sup>	0.115 <sup>⑧</sup>			
合計 <sup>⑧</sup>	38.590 <sup>⑧</sup>	331 <sup>⑧</sup>				

表2.2. Context Entity Recall

変動要因 <sup>①</sup>	変動 <sup>②</sup>	自由度 <sup>③</sup>	分散 <sup>④</sup>	観測された分散比 <sup>⑤</sup>	P-値 <sup>⑥</sup>	F 境界値 <sup>⑦</sup>
グループ間 <sup>⑧</sup>	0.585 <sup>⑧</sup>	3 <sup>⑧</sup>	0.195 <sup>⑧</sup>	3.626 <sup>⑧</sup>	0.013 <sup>⑧</sup>	2.632 <sup>⑧</sup>
グループ内 <sup>⑧</sup>	17.639 <sup>⑧</sup>	328 <sup>⑧</sup>	0.054 <sup>⑧</sup>			
合計 <sup>⑧</sup>	18.224 <sup>⑧</sup>	331 <sup>⑧</sup>				

表2.3. Context Recall

変動要因 <sup>①</sup>	変動 <sup>②</sup>	自由度 <sup>③</sup>	分散 <sup>④</sup>	観測された分散比 <sup>⑤</sup>	P-値 <sup>⑥</sup>	F 境界値 <sup>⑦</sup>
グループ間 <sup>⑧</sup>	0.0700 <sup>⑧</sup>	3 <sup>⑧</sup>	0.023 <sup>⑧</sup>	0.210 <sup>⑧</sup>	0.889 <sup>⑧</sup>	2.632 <sup>⑧</sup>
グループ内 <sup>⑧</sup>	36.439 <sup>⑧</sup>	328 <sup>⑧</sup>	0.111 <sup>⑧</sup>			
合計 <sup>⑧</sup>	36.509 <sup>⑧</sup>	331 <sup>⑧</sup>				

表2.4. Answer Relevancy

変動要因 <sup>①</sup>	変動 <sup>②</sup>	自由度 <sup>③</sup>	分散 <sup>④</sup>	観測された分散比 <sup>⑤</sup>	P-値 <sup>⑥</sup>	F 境界値 <sup>⑦</sup>
グループ間 <sup>⑧</sup>	0.008 <sup>⑧</sup>	3 <sup>⑧</sup>	0.003 <sup>⑧</sup>	0.140 <sup>⑧</sup>	0.936 <sup>⑧</sup>	2.632 <sup>⑧</sup>
グループ内 <sup>⑧</sup>	6.390 <sup>⑧</sup>	328 <sup>⑧</sup>	0.019 <sup>⑧</sup>			
合計 <sup>⑧</sup>	6.398 <sup>⑧</sup>	331 <sup>⑧</sup>				

表2.5. Faithfulness

変動要因 <sup>①</sup>	変動 <sup>②</sup>	自由度 <sup>③</sup>	分散 <sup>④</sup>	観測された分散比 <sup>⑤</sup>	P-値 <sup>⑥</sup>	F 境界値 <sup>⑦</sup>
グループ間 <sup>⑧</sup>	0.026 <sup>⑧</sup>	3 <sup>⑧</sup>	0.009 <sup>⑧</sup>	0.242 <sup>⑧</sup>	0.867 <sup>⑧</sup>	2.632 <sup>⑧</sup>
グループ内 <sup>⑧</sup>	11.801 <sup>⑧</sup>	328 <sup>⑧</sup>	0.036 <sup>⑧</sup>			
合計 <sup>⑧</sup>	11.828 <sup>⑧</sup>	331 <sup>⑧</sup>				

■ 多重比較法

・Context Precision

表 2.6. PDFをコントロール群としたDunnett検定

誤差分散 <sup>e3</sup>	0.115 <sup>e3</sup>
統計量 L0 <sup>e3</sup>	1.507 <sup>e3</sup>
統計量 L2 <sup>e3</sup>	2.568 <sup>e3</sup>
統計量 L3 <sup>e3</sup>	2.520 <sup>e3</sup>
相関係数 <sup>e3</sup>	0.5 <sup>e3</sup>
棄却限界値 <sup>e3</sup>	2.364 <sup>e3</sup>

表 2.7. Tukey検定

PDF-L0 <sup>e3</sup>	2.131 <sup>e3</sup>
PDF-L2 <sup>e3</sup>	3.632 <sup>e3</sup>
PDF-L3 <sup>e3</sup>	3.564 <sup>e3</sup>
L0-L2 <sup>e3</sup>	1.501 <sup>e3</sup>
L0-L3 <sup>e3</sup>	1.433 <sup>e3</sup>
L2-L3 <sup>e3</sup>	0.067 <sup>e3</sup>
棄却限界値 <sup>e3</sup>	3.74 <sup>e3</sup>

・Context Entity Recall

表 2.8. L1をコントロール群としたDunnett検定

誤差分散 <sup>e3</sup>	0.0538 <sup>e3</sup>
統計量 PDF <sup>e3</sup>	2.850 <sup>e3</sup>
統計量 L2 <sup>e3</sup>	2.675 <sup>e3</sup>
統計量 L3 <sup>e3</sup>	2.511 <sup>e3</sup>
相関係数 <sup>e3</sup>	0.5 <sup>e3</sup>
棄却限界値 <sup>e3</sup>	2.364 <sup>e3</sup>

表 2.9. Tukey検定

PDF-L0 <sup>e3</sup>	4.031 <sup>e3</sup>
PDF-L2 <sup>e3</sup>	0.248 <sup>e3</sup>
PDF-L3 <sup>e3</sup>	0.479 <sup>e3</sup>
L0-L2 <sup>e3</sup>	3.783 <sup>e3</sup>
L0-L3 <sup>e3</sup>	3.552 <sup>e3</sup>
L2-L3 <sup>e3</sup>	0.231 <sup>e3</sup>
棄却限界値 <sup>e3</sup>	3.74 <sup>e3</sup>

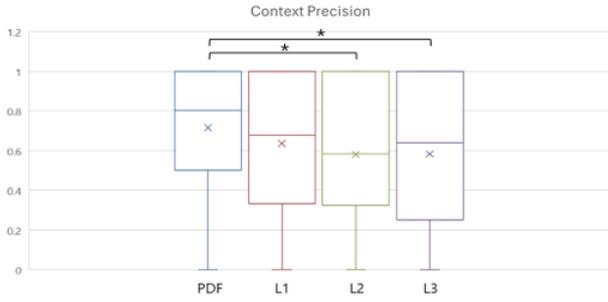


図 1. Context Precisionの分布図

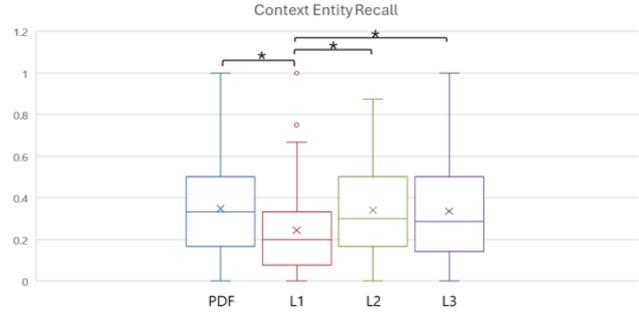


図 2. Context Entity Recallの分布図