LLMを用いた自由記述データのコーディング手法の検討
―職場における生成AI活用調査を題材に―

川崎裕子・今城志保・山田香・藤村直子
（リクルートマネジメントソリューションズ　組織行動研究所）

＜あらまし＞　本研究では、職場における生成AI活用調査の自由記述データを対象に、LLMを補助的コーダーとして用いたコーディング手法の信頼性と有用性を探索的に検討した。研究者が定義したコーディングルールに基づきサンプルコーディングを実施した結果、人とAIとの不一致分析を通じたルール修正を経てCohen's K ≥0.61となる水準が確認され、全数データへ適用した。コーディングへのLLMの活用はルールの精緻化やデータ構造把握に資する可能性が示唆された。

1. 目的と背景

- 質的研究における内容分析では、信頼性確保のために複数のコーダーによる一致度の確認が求められる一方（White et al., 2006）、コーディングの不一致や再現性の問題は、しばしば課題となり（星野・荘島, 2016）、コンピュータの活用の意義が指摘されてきた（樋口, 2004）。
- ChatGPTが公開された2023年以降は、LLMのコーディング活用に関する検討が進展している。LLMがテキスト分類において人間のコーダーに匹敵する精度を示す（Gilardi *et al.*2023, Rathje *et al* 2024 ）、コードブックをLLMに与えた演繹的コーディングでのLLMと人間コーダーの高い一致率が示される（Tai *et al.* 2024 ）など。
- 一方、質的研究へのLLM利用の信頼性検証は不十分。LLMは高い分類精度と文脈理解能力を持つが、分析者がLLMを共同研究者のように錯覚し出力を十分な思考の産物として受容しやすいことや、逸脱事例への着目が損なわれることが危惧され、どの作業にどこまで用いるかは慎重を要するとされる（山口, 2026）。

→ 職場における生成AI活用調査の自由記述データを対象に、**LLMを補助的コーダーとして用いたコーディング手法の信頼性と有用性を探索的に検討**する。

2. 方法

データ：有効回答数 833件

- 「直近1カ月以内に仕事で生成AIを使ってよかった事例」がある会社員・公務員
- 週に数回以上生成AIを利用している人
- 20～40代均等割り付け、男性61.9%

調査時期：2026年3月　オンライン調査

分析の手順　・「直近1カ月以内に仕事で生成AIを使ってよかった事例」についての自由記述を使用して分類する。

研究者：カテゴリ設計・コーディングルール作成 → 研究者とLLM（Claude Sonnet4.6）：サンプルコーディング（24件 × 2回） → 研究者：不一致分析、ルール修正 → LLM：全数コーディング（833件）／研究者：判定不能ケースを中心に確認・調整

信頼性確認指標（Landis & Koch , 1977）
LLMを全数コーディングに用いる際の、実用的信頼性水準を**Cohen's $\kappa \geq 0.61$**とした。

3. コーディングカテゴリ

- 別の「職場における生成AIの活用と学習との関係の探索」を目的とした研究の一環として、生成AIへのインプットの内容と、どのような学びがあったのかの関係を検討するためのコーディングカテゴリを準備した。

インプット内容（3カテゴリ；排他）

使用設問		
A「最初のプロンプト入力内容」（判断できないとき：B「具体的な活用場面」で補完）		
カテゴリ	定義	記載例
HIGH インプット	自分の案・仮説・背景・目的を詳細に説明してアドバイスや情報を求める	「メールの添削依頼」「前提条件を記載してプログラムエラーを貼り付け」
LOW インプット	コピペ・単語のみ・極めて短い指示	「英訳して」「要約して」
判定不能	記述が曖昧で判定不能、HIGH・LOW両方にまたがる記述	「具体的に詳しく教えて」「あなたは〇〇です」

学びの型（4カテゴリ；排他）

使用設問		
C「学び実感の理由」（能力の向上や成長につながった・つながらなかった理由）（判断できないとき：D「活用してよかった点」他で補完）		
カテゴリ	定義	記載例
HOW型	プロセス・方法・コツへの気づき、未来志向の応用言及	「次からはこう書けばいいとわかった」「思考方法が身についた」「頭の中が整理された」
WHAT型	知識・情報の獲得、問題解決・疑問解消	「知識が増えた」「知らないことを知れた」「疑問点を解消できた」
学びLOW	学びの明示的否定、AI依存・代替への言及、時短のみ	「AI任せだから」「単純作業だったから」「時短になった」
判定不能	記述が曖昧で判定不能、複数のカテゴリにまたがる記述	自分になかったものがあった」「AIを使う能力が向上」

4. 信頼性確認（サンプルコーディング）

- サンプル抽出：「具体的な活用場面」の記述から分類した12の用途カテゴリ（資料作成・翻訳など）を層とし、各カテゴリ2件ずつ無作為抽出。

結果

- 信頼性確認：第1回（インプット内容κ=0.78, 学びの型κ=0.42）での不一致（インプット内容3件、学びの型10件）を検討しルール変更。

→ **第2回では両変数で、実用的な信頼性基準とした$\kappa \geq 0.61$が示された。**

	第1回（24件）	第2回（24件）
インプット内容	k=.78　（一致率 87.5%）	**k=.78　（一致率 87.5%）**
	不一致3件は全て研究者判定を採用	不一致3件は全て研究者判定を採用
学びの型	k=.42　（一致率 58.3%）	**k=.63　（一致率 83.3%）**
	不一致10件中2件はLLM、8件は研究者を採用	不一致4件中1件はLLM、残りは研究者を採用

5. 全数コーディングの結果

クロス集計の結果　（$\chi^2(6)=68.6$, $p<.001$）

- HIGHインプット ではHOW型が有意に多く、LOWインプットでは学びLOWが有意に多い結果となった。データの構造把握ができた。

		学びの型 HOW型	WHAT型	学びLOW	判定不能	計
HIGH インプット	行割合	34.5%	33.8%	26.1%	5.6%	100.0%
	度数	49	48	37	8	142
	調整済み残差	6.6	1.6	−4.5	−2.7	
LOW インプット	行割合	12.8%	28.2%	46.4%	12.6%	100.0%
	度数	81	179	294	80	634
	調整済み残差	−4.5	−0.2	3.3	0.4	
判定不能	行割合	5.3%	17.5%	50.9%	26.3%	100.0%
	度数	3	10	29	15	57
	調整済み残差	−2.3	−1.9	1.2	3.3	
計	行割合	16.0%	28.5%	43.2%	12.4%	100.0%
	度数	133	237	360	103	833

6. 考察と課題

- LLMを**補助的コーダーとするコーディング手法について、一定の実用的信頼性を確認**した。
- LLMとの不一致ケースの検討において、**ルールの精緻化への有効性**を示した。LLMの判定不能ケースが、ルールの不明箇所の可視化、洗練の手掛かりになった。

課題

- **手法の適用範囲の検証**：本研究では、短文のデータを用い、比較的シンプルなカテゴリ分類を行った。他研究への適用には、対象の特性、カテゴリ数やルール設定の影響、許容できる一致水準、LLMモデル差による頑健性など多面的な検討が必要である（O'Connor & Joffe, 2020）。
- **サンプル規模の拡大**：κはカテゴリ分布の影響を受けるため（Sim & Wright, 2005）、より安定した推定のためのサンプル規模はさらに検討の余地がある。
- **全数コーディング手順の標準化**：判定不能ケースを研究者が確認・調整したが、その手続きの妥当性や標準化には検討の余地がある。