

セキュア・ローカルLLMアーキテクチャにもとづく

説明可能な経営分析支援の設計と評価

－7B/14B経営助言ベンチマークと

Self-hosted Lecture-to-Quiz (L2Q) によるQC参照実装－

Design and Evaluation of Explainable Management-Analysis Support Based on a Secure Local LLM Architecture: A 7B/14B Management-Advisory Benchmark and a Self-Hosted Lecture-to-Quiz (L2Q) QC Reference Implementation

新 谷 正 嶺*

Seine A. SHINTANI

要旨

機密性の高い経営データや学内資料を外部のLLMサービスに送信せずに活用したいという要請は強い。しかしローカルLLMは、(i) 構造化出力の崩れ、(ii) 統計信号の読み違い、(iii) 非有意結果の因果断定など、実装上の失敗が意思決定を誤らせる。本研究はAI2L（ブラックボックス最小化と人間の検証を重視する指針）を設計理念とし、(A) 統計要約から経営助言を生成するタスク（合成ベンチマーク）と、(B) 講義PDFから5択問題を生成し、決定論的品质管理（QC）で受理・警告・再試行を制御するSelf-hosted Lecture-to-Quiz (L2Q) を比較した。ケース(A)では11ケースで7B/14Bを評価し、厳密JSONとして解析可能な成功率が7Bで0/11、14Bでも1/11と低く、モデル精度以前に構造化の安定化がボトルネックであることを確認した。Tier2では7Bで方向ミス2/44、14Bで欠落8/44が残り、文章面でも非有意の言い切り事故が観察された。一方ケース(B)では、文法制約（GBNF）と決定論的QCを外側に置くことで、15 runsで120/120件を受理しつつ、致命的エラーはリトライ2/122で自動回復、軽微なリスクは警告8/120（6.7%）としてレビューキューに分離できた。以上より、ローカルLLMを実務に接続する鍵は、モデルを“判断主体”として信頼するのではなく、(1) 出力の文法・スキーマ拘束、(2) 決定論的QC、(3) 再試行、(4) レビューキュー、(5) 静的成果物化、で包み込む設計にあると結論づける。

キーワード：ローカルLLM, セキュア・ローカルLLMアーキテクチャ, 経営分析支援, 説明可能性 (XAI), AI2L, 構造化出力, JSONスキーマ, 文法制約 (GBNF), 決定論的QC

*生命健康科学部生命医科学科－AI数理データサイエンスセンター（兼任）

名古屋大学高等研究院（兼任）

Department of Biomedical Sciences－AI to Learn

1. はじめに

会計・販売・在庫といった経営データは、意思決定に直結する一方で秘匿性が高い。クラウドLLMに投げれば「それらしい説明」は得られるが、データを外部に出せない現場は多い。したがって、ローカル環境で動くLLMを用い、社内（学内）に閉じたまま“読める説明”を得る技術が重要になる。

しかし、実装の観点では「自然な文章」より前に越えるべき壁がある。第一に、下流システム（ダッシュボード、LMS、ログ監査）で扱うには構造化出力が必要であるが、LLMは長いJSONを途中で崩しやすい。第二に、統計的に非有意な結果であっても因果を断定するなど、説明責任上の事故が起きやすい。第三に、ローカル上の事故が起きやすい。第三に、ローカルLLMはモデル規模の制約があり、品質のばらつきを前提にした設計が必要である。

本研究は2025年度産業経済研究所学内プロジェクト「ローカル小規模LLMを活用した説明可能経営分析支援ツールの構築と評価」の一環として、まず“壊れ方”を特定し、次年度以降の実装方針を定めることを目的とする。

具体的には、(A) 統計要約（平均差・p値等）から経営助言を生成する評価（合成ベンチマーク）と、(B) 講義PDFから問題集を生成しGoogle Formに変換する自己完結パイプライン（Self-hosted Lecture-to-Quiz）を並置し、両者に共通する設計課題－構造化・監査・品質管理－を整理する。後者は教育支援を入口とするが、社内文書・仕様書・報告書など、外部送信できない文書を扱う業務支援一般に同型の課題を含むため、経営分析支援に対する参照実装として扱う。

本稿の貢献は次の3点に整理できる。

- Tier1.3（構造／信号／文章）の分解により、ローカルLLM運用の失敗様式を“実装要件”として可視化した。
- 経営助言タスクで、7B/14Bのトレードオフ（方向ミス vs 欠落）を定量化し、モデル選定より先に解くべき課題を明確化した。
- L2QをQC参照実装として位置づけ、文法制約＋決定論的QC＋レビューキューにより“壊れない成果物”へ落とす一般レシピを提示した。

本稿の貢献を以下にまとめる。

経営助言生成について、合成ベンチマークとTier1.3（構造／信号／文章）評価で失敗様式を可視化した。

講義PDF→問題集パイプラインにおける文法制約と決定論的QCの運用統計を整理し、構造化出力の安定化手段を示した。

AI2Lの観点から、ローカルLLMを“判断主体”ではなく“ドラフト生成器＋検査

対象”として扱う設計指針を提案した。

2. 背景：ローカル運用・説明責任・構造化

ローカルLLMの利点は、(i) データ主権（外部送信の回避）と、(ii) コスト・レイテンシの予見性にある。特に中小企業や研究教育現場では、資料・データの二次利用や情報漏えいへの懸念が導入の阻害要因になりやすい。

一方で、意思決定支援では「正しく見える誤り」が致命的になり得る。統計の符号規約を取り違えて増減を逆に読む、非有意にもかかわらず因果を断定する、あるいは監査のための必須項目が欠落する、といった失敗は、誤った施策の実行に直結する。

AI2L (AI to Learn) は、AIを“学習・理解の補助”に限定し、ブラックボックスを最小化する指針として提案されている（新谷）。本研究ではAI2Lを、(1) ローカル運用（プライバシー）、(2) 監査可能な出力（根拠とログ）、(3) 人間の最終確認、(4) 必要最小限のAI利用（静的成果物化）、として具体化し、設計と評価を行う。

3. 方法

3. 1 ケース（A）：統計要約からの経営助言生成（合成ベンチマーク）

月次の売上指標（販売数量units_sold・売上revenue_yen）と在庫指標（在庫日数DIO・保管費用holding_cost_yen）を企業ごとに合成し、競合ショック（pre/post）と在庫イベント（baseline/over）を注入した。合成により、イベント注入の有無（ground truth）を保持したまま、統計的には偽陽性・偽陰性が起こり得る“現実的な紛らわしさ”を再現できる。

評価では時系列そのものはLLMに渡さず、平均・平均差・p値などの統計要約のみを入力とし、LLMの役割を「読み取り・説明・助言」に限定した。出力は厳密JSON（signals/findings/recommendations/caveats/glossary等）とし、後段で自動処理できることを要件とした。

評価は3段階で行った。Tier1は構造（JSONとして解析できるか、必須キーが揃うか）、Tier2は信号（有意性・方向の正しさ）、Tier3は文章（非有意で断定しない、根拠を示す、現実的な助言である）である。

合成データは以下の生成モデル（概要）で作成した。

$$\text{販売数量} : U_{i,t} = \mu_i + \beta_i t + s_{i,m(t)} + \varepsilon_{i,t}, \quad \varepsilon_{i,t} \sim \mathcal{N}(0, 60^2)$$

$$\text{売上} : R_{i,t} = p_i U_{i,t} + \eta_{i,t}, \quad \eta_{i,t} \sim \mathcal{N}(0, (30 p_i)^2)$$

$$\text{DIO} : D_{i,t} = \max\{1, d_i^{(0)} + \xi_{i,t}\}, \quad \xi_{i,t} \sim \mathcal{N}(0, 6^2)$$

$$\text{保管費} : H_{i,t} = \max\{0, h_i^{(0)} + \zeta_{i,t}\}, \quad \zeta_{i,t} \sim \mathcal{N}(0, (2 \times 10^5)^2)$$

3. 2 ケース (B) : Self-hosted Lecture-to-Quiz (講義PDF→問題集)

講義資料PDF (テキスト抽出可能な範囲) を入力し, ローカルLLMで5択問題 (MCQ) を生成し, Google Formに取り込めるCSV/JSONLへ変換するパイプラインを対象とした。

ここでは, 生成物をそのまま配布するのではなく, 文法制約 (llama.cppのGBNF) でJSONスキーマに固定し, さらに決定論的QC (スキーマ適合, 選択肢重複, 数値・定数同値, 単一正解検証) を通過したものだけを受理する。

QCは2層構造とし, 致命的な違反は再試行 (bounded retries) で排除し, 軽微だが誤解を招く恐れがある事項 (例: 丸め指示欠落) は警告としてログに残し, 人間のレビュー対象として可視化する。

4. 結果 (根拠データ)

4. 1 ケース (A) の結果: 経営助言生成 (7B/14B)

11ケース×4指標 (計44 signals) で評価した。主要結果を表1と図1, 2に示す。

表1 ケース (A) の主要結果 (n=11)

指標	7B	14B
Tier1: JSON 解析成功 (件)	0/11	1/11
Tier2: signals 正解 (件)	42/44	36/44
Tier2: 方向ミス (件)	2	0
Tier2: 欠落 (件)	0	8
Tier2: 正解率 (欠落含む)	0.955	0.818

Tier1 (構造) では, 7Bは0/11, 14Bでも1/11しか厳密JSONとして解析できず, 構造化出力の安定化が最大のボトルネックであった。

Tier2 (信号) では, 7Bはsignalsを全ケースから抽出できたが方向ミスが2/44発生した。14Bは抽出できた範囲では方向ミスが無い一方, 2ケースで出力欠落 (8/44) が生じた。

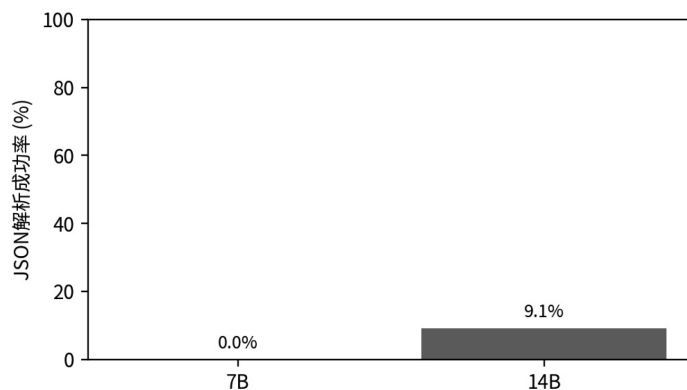


図1 Tier1：厳密JSON解析成功率 (n=11)

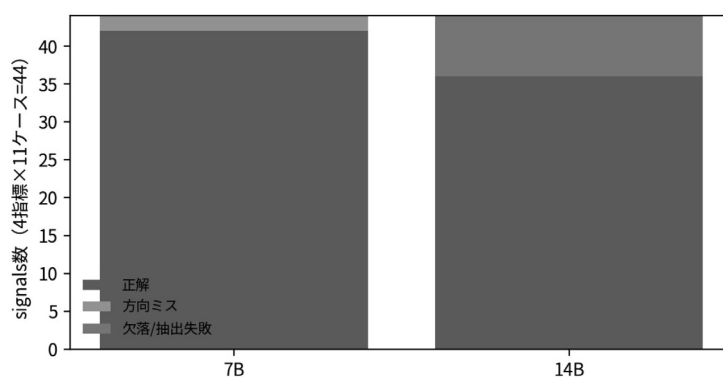


図2 Tier2：signals内訳（正解／方向ミス／欠落）

4. 2 ケース（B）の結果：講義PDF→問題集パイプライン（QC統計）

Self-hosted Lecture-to-Quiz (L2Q) のseed sweep（3講義×5 seed，計15 runs）では，8問/Runを目標として120問を生成した。QC統計を表2および図3に示す。

表2 ケース（B）のQC統計（L2Q seed sweep）

項目	値
Runs	15 (3 lectures × 5 seeds)
受理数 (Acceptance)	120/120
総試行 (Attempts)	122
リトライ率	2/122 (1.6%)
警告数 (Warnings)	8
警告率	8/120 (6.7%)
実行時間/Run (平均±SD)	58.55 ± 2.34 s
最終配布セット	24 問 (警告 0)

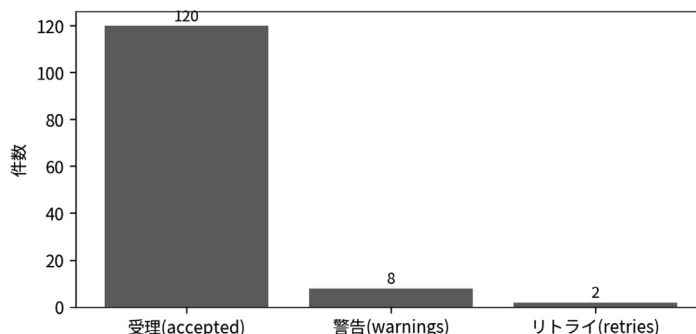


図3 L2Q：QC統計 (15 runs, 120 items, 122 attempts)

致命的な形式崩れ（スキーマ不適合，複数正解，同値選択肢など）をハードQCで排除しつつ，丸め指示欠落などの軽微なリスクは警告として残すことで，“自動化できる範囲”と“人が見るべき範囲”を分離できた。

4. 3 補助結果：ローカルLLMの生成速度と品質（卒研評価）

卒業研究では，情報量（エントロピー）に関する講義範囲から5択8問を生成し，ループリック評価（問題生成30点，翻訳支援25点）と生成時間を比較した。表3と図4に示す。

表3 モデル別のループリック得点と生成時間（5択8問）

Model	MCQ_rubric_30	Translation_rubric_25	Gen_time_sec
ChatGPT 5.2 Pro	28	25	
Qwen2.5-32B	24	22	55.8
Qwen2.5-14B	25	22	31.3
Phi-4	23	21	35.6
Qwen2.5-7B	19	20	17.4

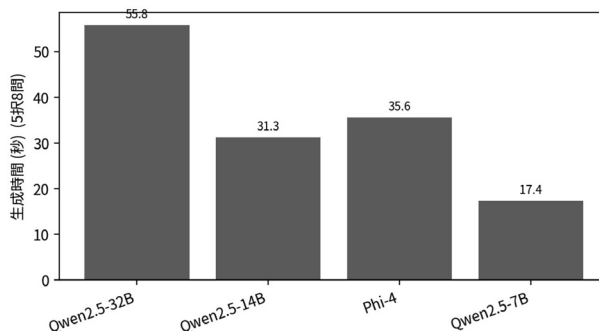


図4 ローカルLLMの生成時間（5択8問，Colab A100）

品質面では14B以上がChatGPTに近い得点を示し、生成時間は32Bでも約1分以内であった。これらは、ローカル運用が“速度面で完全に非現実的”ではないこと、ただし品質確保には設計と検証が必要であることを示唆する。

5. 考察：経営分析支援ツールに必要な設計条件

- (1) Tier1はモデル性能ではなく“設計問題”である。経営助言生成では、厳密JSONの成立率が極端に低かった。これは、推論の正しさ以前に、下流で扱える形式を保証できないことを意味する。対策としては、文法制約やスキーマ制約による生成（constrained decoding）、生成長の抑制、JSON修復と再試行、生成の分割（signals→文章）など、外側のシステムで“壊れた出力を出さない”設計が不可欠である。
- (2) “計算で確定できるものはLLMにやらせない”。有意性や方向のように機械的に確定できる信号は、コード側で確定してLLMには固定値として渡し、LLMは説明・注意点・施策案の文章化に限定する方が、誤読の事故コストを下げる。これはAI2Lのブラックボックス最小化とも整合する。
- (3) QCの二層化が有効である。ケース（B）では、ハードQCで破綻を排除し、警告をレビューキューとして残す運用が安定した。経営助言でも同様に、（a）形式と必須キー、（b）断定表現（非有意の因果断定）、（c）根拠参照の欠落、を自動検査し、通過したもののみを配布・保存する設計が現実的である。
- (4) “静的成果物化”が運用を変える。L2Qは最終的にJSONL/CSVとして出力し、学習時にはLLMを呼ばない。経営支援でも、レポート生成を一度行い、監査ログ付きの静的レポートとして保存・共有する設計は、継続課金・再現性・説明責任の観点で有利である。
- (5) 参照アーキテクチャ：LLMの外側に“品質保証”を置く。ケース（B）の設計を一般化した参照アーキテクチャを図5に示す。LLM出力は文法制約で崩れを防ぎ、決定論的QCで落とし、致命的違反は自動再試行で回復する。警告はレビューキューに集約し、人が見るべき箇所だけを可視化する。

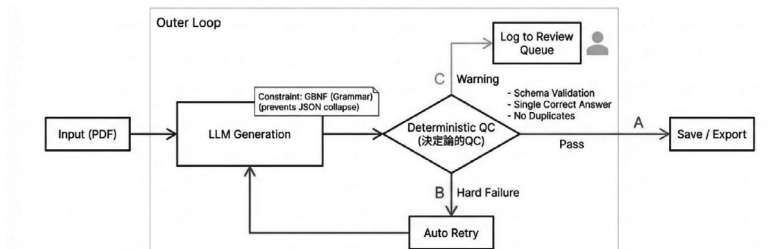


図5 セキュア・ローカルLLM参照アーキテクチャ（文法制約＋決定論的QC＋レビューキュー）

6. 結論

本稿は、ローカルLLMを用いた業務支援を“安全に動くシステム”として成立させるために、失敗様式をTier1.3（構造／信号／文章）で整理し、経営助言生成（ケースA）と講義PDF→問題集（ケースB）で具体的に検証した。ケースAでは、厳密JSONの成立率が7Bで0/11, 14Bで1/11と低く、モデルの賢さ以前に構造化出力の保証が最大の障壁であることを確認した。

これに対しケースBでは、文法制約（GBNF）と決定論的QCを外側に置くことで、120/120件を受理しつつ、致命的違反は自動再試行で回復し、軽微なリスクだけを警告としてレビューキューに分離できた。両者を踏まえると、ローカルLLM運用の要点は次の3点である。

- （1）計算で確定できる信号はコード側で確定し、LLMには“説明と提案”を担当させる。
- （2）出力は文法・スキーマで拘束し、決定論的QCで落とす（壊れた出力を下流に流さない）。
- （3）警告ログと静的成果物化により、運用時の監査・再現・責任分界を明確にする。

次段階では、この参照アーキテクチャを経営支援タスクへ移植し、(a) signalsの自動確定、(b) 構造化出力の強制、(c) 断定表現や根拠参照の自動検査、(d) レビューキュー運用、を組み合わせたプロトタイプを構築する。さらに実データを用いたユーザ評価（理解度、説明責任、意思決定支援の有用性）を行い、AI2Lの評価軸（透明性・人間の検証可能性・プライバシー・持続可能性）に沿って、教育・実務の双方へ接続する。

謝辞

本研究は、2025年度中部大学産業経済研究所学内プロジェクト研究「ローカル小規模LLMを活用した説明可能経営分析支援ツールの構築と評価」の支援を受けた。また、講義PDF→問題集パイプラインに関する卒業研究発表（天野優里）および初期実装の試行は、本研究の設計検討に有益な参照例となった。

参考文献

- [1] 新谷正嶺: AI2L: Human-Centered Guidelines for Black-Box-Free AI and Empirical Law Discovery via Symbolic Regression, jxiv preprint, DOI: 10.51094/jxiv.1435 (Version 2).
- [2] 新谷正嶺: AI2L: 学習支援ツールとしてのAI活用指針の定義と応用可能性, 生命健康科学研究所紀要 Vol.21 (2024).
- [3] Welch, B. L.: The generalization of “Student's” problem when several different population variances are involved, *Biometrika* 34(1-2), 28-35 (1947).
- [4] Doshi-Velez, F. and Kim, B.: Towards a rigorous science of interpretable machine learning, arXiv:1702.08608 (2017).
- [5] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* 1, 206-215 (2019).
- [6] Qwen Team: Qwen2.5 Technical Report, arXiv:2412.15115 (2024).
- [7] Gerganov, G. and contributors: llama.cpp: LLM inference in C/C++, GitHub repository, <https://github.com/ggml-org/llama.cpp> (accessed 2026-02-16).
- [8] Post, M. and Vilar, D.: Fast lexically constrained decoding with dynamic beam allocation for neural machine translation, *NAACL-HLT* (2018).
- [9] JSON Schema: Specification (2020-12), json-schema.org.
- [10] Shintani, S. A.: Self-hosted Lecture-to-Quiz: Local LLM MCQ Generation with Deterministic Quality Control, manuscript (February 2026).
- [11] 天野優里 (指導教員: 新谷正嶺): 外部LLMに送らない学習支援AI: 講義資料PDFからの問題生成・翻訳の品質評価とGoogle Form化, 卒業研究発表資料 (2025年度).
- [12] Wasserstein, R. L. and Lazar, N. A.: The ASA's Statement on p-Values: Context, Process, and Purpose, *The American Statistician* 70(2), 129-133 (2016).
- [13] Guidance-AI: Guidance: A guidance language for controlling large language models, GitHub repository, <https://github.com/guidance-ai/guidance> (accessed 2026-02-20).