

生成AIは利益を生むか？ — 事業インパクト測定に向けたLLM活用型評価基盤の初期検討 —

近藤 真暉[†]株式会社 ARISE analytics[†]

1 はじめに

生成AIの業務活用が急速に進む中、その導入効果をどのように定量的に評価するかは重要な経営課題となっている。約80%の企業が生成AIを導入しているものの、売上高や利益率の向上を実感できていない[1]。効果測定が工数指標に留まり、事業成果への寄与を示せていないためである。生産性効果はRCT等により実証されているが[2]、工数削減のみの測定には本質的な限界が存在する。

提案書作成を例に挙げると、作成工数が削減されても品質が低下していれば受注率は向上しない。逆に、時間を要しても高品質な提案書を作成できれば、受注率向上という事業成果に直結する。すなわち、「工数削減」と「価値創出」は必ずしも一致しない。

関連研究として、ロジックモデル[3]は因果構造を示すが、効果測定の具体手法は提供していない。AI導入効果の測定指標としては、Speed（処理速度）・Quality（品質）・Effort（作業負荷）の3軸が提案されている[4]。また、品質と事業成果の関係性は実証されており[5]、品質は事業インパクトを測定する上で重要な媒介変数と位置付けられる。LLM-as-a-Judge[6]は品質評価手法として提案されている。しかし、成果物の品質を定量的に測定し、事業成果との関係を体系的に分析する枠組みは十分に整理されていない。また、品質評価は人手による主観の評価に依存しており、スケーラビリティや再現性に課題がある。

そこで本稿では、因果層・測定層・分析層の3層構造から成る評価フレームワークを提案する。工数に加えて成果物の品質および事業成果を測定・統合する点に特徴がある。本稿ではフレームワークの設計方針と、LLM-as-a-Judgeを用いた品質評価手法の妥当性検証について報告する。

2 提案手法

1章で述べたロジックモデル[3]を業務プロセスに適用すると、工数（Input）→タスク（Activity）→成果物（Output）→事業成果（Outcome）という流れとなる。

ロジックモデルは構造を示すが、それだけでは効果を検証できない。そこで本研究では、このロジックモデルを因果層として採用し、測定層・分析層を加えた3層構

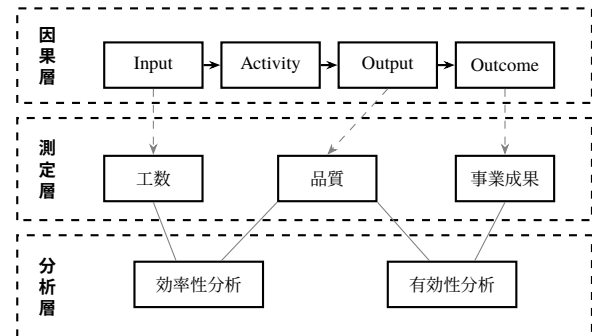


図1 提案フレームワークの構成

造のフレームワークを提案する（図1）。

測定層では、工数（Input）、品質（Output）、事業成果（Outcome）を定量化する。工数と品質は、AI導入効果の測定指標として提案されているSpeed・Qualityに対応する[4]。事業成果は、受注率や売上高等の既存の指標で測定可能である。なお、Effortは主観評価が主であり客観的測定が困難なため、本研究では採用しない。

分析層では、効率性 E と有効性 V を以下のように定義する。効率性は単位工数あたりの品質として $E = Q/T$ と定義する。ここで T は工数、 Q は品質である。導入効果は $E_{\text{after}}/E_{\text{before}}$ で評価する。有効性は、品質水準による事業成果達成率の差として $V = P(O|Q \geq \theta) - P(O|Q < \theta)$ と定義する。ここで O は事業成果の達成（受注、承認等）、 θ は品質の閾値である。 $V > 0$ であれば、品質向上が事業成果に正の寄与を持つ。

既存の効果測定では工数のみを測定していたため、工数削減が事業成果に結びついていないか検証できなかった。提案フレームワークでは、生成AI導入前後の効率性・有効性を比較することで、導入効果を評価する。生成AI導入により、同一工数での品質向上、または同一品質での工数削減が実現すれば、効率性が向上したと評価できる。さらに、品質向上が事業成果の改善に繋がれば、有効性も検証できる。

3 品質評価の初期検討

測定層では工数・品質・事業成果を測定する。工数は工数管理ツール等で、事業成果は受注率や売上高等の既存の指標で測定可能である。一方、成果物の品質は従来人手で評価されてきたが、スケーラビリティに課題がある。そこで本稿では、品質の測定手法に焦点を当てる。

品質の測定手法として、人手評価はスケーラビリティに課題があり、ルールベースの自動評価は多様な成果物

Measuring the Business Value of Generative AI: A Preliminary Study on LLM-based Evaluation Framework

[†] Masaki Kondo, ARISE analytics Inc.

件名:【ご確認依頼】Q4 営業戦略資料のレビュー (12/20 締切)
田中部長
お疲れ様です。営業企画課の山田です。営業戦略資料のドラフトが完成しましたので、レビューをお願いいたします。
■ ご確認いただきたい点
1. 目標設定の妥当性 (スライド 5-8)
2. リソース配分の方針 (スライド 12-15)
■ 期限: 12 月 20 日 (金) 17:00 まで

図 2 高品質メールの例

表 1 評価対象メールの概要

品質	件名	期限	確認ポイント
高	目的・期限を明示	日時指定	箇条書きで明示
中	簡潔	「今週中」	概要のみ
低	なし	なし	不明確

への対応が困難である。本稿では、品質評価の自動化・標準化に向けた初期検討として、スケーラブルかつ柔軟な評価が可能な LLM-as-a-Judge[6] に着目する。

評価対象として、業務で頻繁に作成され、かつ品質差が明確なビジネスメールを用いた。これは、品質評価の妥当性を初期的に検証するための代理タスクであり、将来的には提案書や報告書等への展開を想定している。評価観点は、実務上重要とされる Clarity (明確さ)、Conciseness (簡潔さ)、Tone (トーン) に加え、事業成果への影響を意識した Actionability (アクション明確さ) の 4 観点とした。検証では、GPT-5.2 を用いて品質水準が異なる 3 種類のメール (高品質・中品質・低品質) を各 5 通、計 15 通作成し、評価スコアが期待される品質水準と整合するかを確認した。評価対象メールの概要を表 1 に、高品質メールの例を図 2 に示す。

LLM-as-a-Judge では、評価基準を明示したルーブリックベースの評価が人間評価と約 80% の一致率を達成することが示されている [6]。本研究でもルーブリックベースを採用し、評価対象メールと 4 観点の評価基準 (各 1-5 点) を入力とし、各観点の評価理由とスコアを出力する構成とした。評価観点と評価基準を表 2 に示す。評価モデルには Claude Opus 4.5 を用い、評価理由を先に記述させる Chain-of-Thought 形式を採用することで、評価の一貫性向上を図った。なお、本検討では単一モデル・単一設定での評価に限定しており、再現性やモデル依存性の検証は今後の課題とする。

結果を表 3 に示す。評価スコアは期待される品質水準と整合しており、設計した評価プロンプトが品質の違いをスコアに反映できることを確認した。

本検討はサンプル数 15 通、単一モデルでの評価であり、汎化性能の検証には至っていないものの、LLM-as-a-Judge を用いた品質評価が測定層の基盤として機能する可能性を示唆している。本格的な運用に向けては、人

表 2 評価観点と評価基準

観点	定義	基準 (1 点→5 点)
Clarity	目的と内容が明確に伝わるか	目的が不明確→目的が冒頭で明示され論理的に構成
Conciseness	必要な情報が過不足なく簡潔か	冗長または情報不足→必要十分な情報が無駄なく記載
Tone	ビジネスに適切な丁寧さか	カジュアルすぎる→適切な敬語と相手への配慮
Actionability	次のアクションが明確か	何をすべきか不明→期限・担当・アクションが明示

表 3 評価スコアの平均値

メール	Clarity	Conciseness	Tone	Actionability	合計
高品質	5.0	5.0	5.0	5.0	20.0
中品質	2.8	3.2	4.0	3.0	13.0
低品質	1.6	1.8	3.0	1.4	7.8

間評価との一致度検証、業種・職種に応じた評価観点のカスタマイズ、および提案書や報告書等の他の成果物への展開が必要となる。

4 おわりに

本稿では、生成 AI 導入効果を事業インパクトとして測定する 3 層構造の評価フレームワークを提案した。本フレームワークは、品質を媒介変数として効率性と有効性を検証する点に特徴がある。

初期検討として、測定層における品質評価基盤の構築に焦点を当て、LLM-as-a-Judge を用いた評価手法の妥当性をビジネスメールを対象に検証した。その結果、品質水準の違いを定量的に捉えられる可能性を確認した。

今後は、人間評価との一致度検証、複数モデルによる再現性検証、および他の成果物への展開により測定の信頼性を高めた上で、分析層の効率性分析・有効性分析を実データに基づき実施し、生成 AI 投資の事業インパクトを定量的に示すことを目指す。

参考文献

- [1] McKinsey & Company: Beyond the hype: Unlocking value from the AI revolution (オンライン), 入手先 <<https://www.mckinsey.com/>>, (参照 2026-01-07).
- [2] Noy, S. and Zhang, W.: Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence, Science, Vol.381, pp.187-192, 2023.
- [3] W.K. Kellogg Foundation: Logic Model Development Guide, 2004.
- [4] Cambon, A. et al.: Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity, Microsoft Research, 2023.
- [5] Prakash, A. et al.: Productivity, quality and business performance: an empirical study, Int. J. Prod. Perform. Manag., Vol.66, No.1, pp.78-91, 2017.
- [6] Zheng, L. et al.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, Proc. NeurIPS, 2023.