

事業評価票の評価システム作成と支援システム の必要機能についての検討

2026年9月

公益財団法人中部圏社会経済研究所

本レポートは、当財団『研究の場』事業の2025年度の成果報告となります。『研究の場』は、当財団の研究力強化の一環として2024年度に創設されました。研究者の卵である大学院生などをインターンとして迎え入れて共同研究を行い、大学の研究と社会課題を結びつけることを企図しています。

『研究の場』では、4月のインターン受け入れ以降、研究テーマのすり合わせ・調査方針の決定から始まり、調査・データ分析の実施、財団内での中間報告・指導内容の反映など、研究の深掘りを経て、翌年3月に財団内での最終報告を実施します。

この間、本レポートの著者であるインターン（橋本 慧海氏）の担当教授である名古屋工業大学 白松 俊教授をはじめ、当財団のEBPM事業においてご指導いただいている愛知学院大学 石橋 健一教授など、多くの方のご指導・ご助言をいただきました。ご支援いただいた皆様にはこの場をお借りして改めてお礼申し上げます。

事業評価票の評価システム作成と支援システムの必要機能についての検討

名古屋工業大学大学院工学研究科工学専攻 橋本 慧海

【ポイント】

本レポートでは、地方自治体における事業評価票作成支援システムの実現に向けて、大規模言語モデル（Large Language Model：以下 LLM）を用いた事業評価票評価システムを試作し、その有用性と支援システムに必要な機能を検討した。実際の事業評価票に適用した結果、評価票の改善支援に一定の効果が確認できた。

① 事業評価票に共通する課題（図 2、3、4）

中部圏内の 11 自治体、全 44 件の事業評価票を対象に評価を行い、スコア傾向を分析した。その結果、論理的な整合性を示すスコアは自治体・事業間で大きな差がない一方、記述の具体性を示すスコアには差が生じることが分かった。

② 実際のシステム利用例（図 5）

本システムを実際の評価票に適用し、指摘された問題点と改善案を反映することで、評価票が改善されてスコアが向上していく過程を示した。

③ 支援システムに求められる機能（第 5 章）

スコア傾向と自治体職員へのインタビューを踏まえ、事業のライフサイクルに応じた支援機能の切り替えや、作成支援を通じた職員の学習・育成を促す機能などが必要であることを示した。

1. 研究背景

行政においてエビデンスに基づく政策形成（Evidence Based Policy Making：以下 EBPM）が進められている。有効性や効率性といった観点から自治体は施策を評価している。この取り組みは「事業評価」や「行政評価」と呼ばれ、その結果をまとめたものは「事業評価票」として公表されている。EBPM では、例えば「イベントを実施したかどうか」だけでなく、「参

加者数」「満足度」「費用対効果」といった客観的な指標に基づいて事業を評価することが求められている。しかし、杉谷（2024）や杉谷（2022）では、自治体の人手不足や担当職員ごとの能力の違いといったさまざまな理由で、事業評価票の質にばらつきが生じることが指摘されている。

今後、一定水準以上の事業評価票の作成を実現するためには、それを支援する

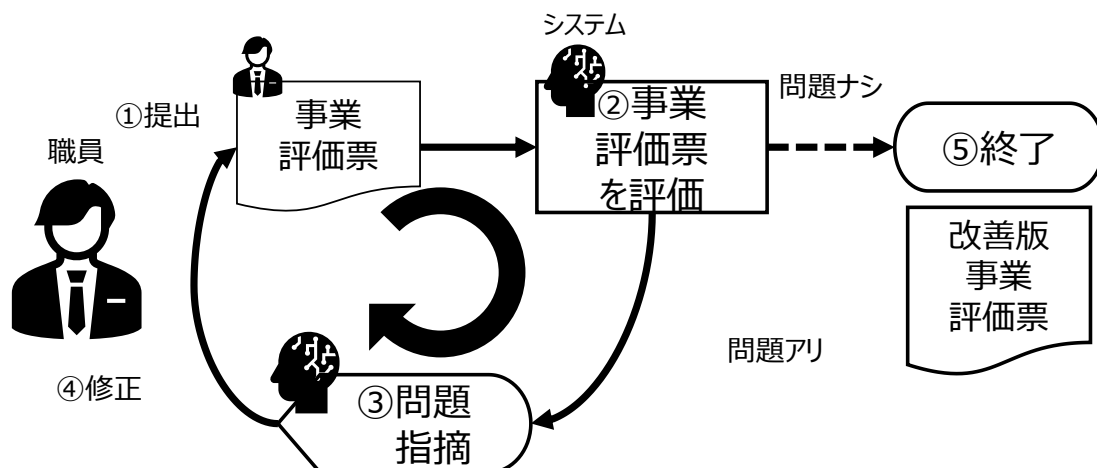


図1 事業評価票作成支援システムの概要図。①システムはユーザから作業中の事業評価票を受け取る、②ガイドラインに沿って評価、③問題がある場合はそれを指摘、④職員に改善を促す。このサイクルによって、より質の高い事業評価票が作成されることが期待される。

システムの導入が望まれる。そこで、本レポートでは図1に示すような事業評価票作成支援システムの開発を最終目標とし、その一環として、LLMを用いた事業評価票の評価機能の開発を行う（図1②）。評価機能の有効性を検証するために、自治体が公表している事業評価票を対象として、評価機能のテストを行い、事業の特性に応じた評価結果の傾向を考察する。

加えて、最終目標達成への足がかりを得るために、本機能の開発プロセスから得られた知見と実験結果を踏まえ、事業評価票作成支援システムが持つべき機能を指摘する（図1③④の必要機能についての考察）。

2. 評価機能の開発

適切な支援には、作成された事業評価票を適切に評価し、その問題点を指摘することが必要である。そのため、本レポートでは評価機能の開発を目指す。

本レポートでの評価方法は「事業評価票が適切に記述されているか」を評価する。具体的には、内閣府が示す「行政評価のためのガイドライン¹」を基に筆者が作成した評価項目に対して、個々の事業評価票の記述がどの程度適合しているかをシステムが評価する。

「適切な記述」とは、例えば次のような観点を満たしているかどうかを指す。

- 事業の目的が明確に記述されているか

¹レビューシートを政策立案や予算要求などの将来に向けての「意思決定」の一環として積極的に活用し、基礎的なEBPMを実践するための考え

方・具体的方法をまとめるための要綱(内閣官房行政改革推進本部事務局 2025)

- 成果指標が妥当かつ可能な限り定量的に示されているか
- 次年度以降の改善方針が具体的に記載されているか

自然言語処理分野では、文書評価のためのさまざまな手法が提案されてきた。特に、近年登場した LLM は、既存の自然言語処理技術に比べて高い文脈理解能力と意味理解能力を持ち、長い文章であっても適切な指示（プロンプト）によって、精度の高い評価が可能となっている。また、いわゆる LLM-as-a-judge と呼ばれる、LLM に文書評価者としての役割を担わせる研究では、人間にも匹敵する文書・言語評価性能を有することが示されている。

事業評価票には「必要事項がきちんと書かれているか」などの具体性の問題と、「書かれている内容が筋として整合しているか」などの論理性の問題が混在する。そのため、事業評価票の評価を単一の尺

度で一元的に行うことは難しい。

そこで、本レポートでは、**記載の有無・具体性を測る記述スコア**と、**事業論理の整合性を測る論理スコア**の 2 つの観点から 5 段階評価（0～4）を行った。システムに与えた評価基準を表 1 に示す。

評価システムのプロンプトには、タスクプロンプト（タスク定義・出力規則・スコア基準（表 1））、評価項目、事業評価票（公表されている PDF を手動で整形したもの）を用いた。

事業評価票は各自治体の総合計画に基づいて作成されるため、総合計画書の理解が評価に有用であると思われる。一方で、総合計画書は自治体により記述粒度や分量が大きく異なり、取り扱いが難しい。本レポートでは、総合計画書は入力として用いず、評価項目に対する事業評価票の記述のみを評価の対象とした。

表 1 記述スコアと論理スコアの基準。0 が最も悪い評価、4 が最も良い評価である。

スコア	記述スコア	論理スコア
0	評価項目に対する直接記述がまったくない。	全体の論理が破綻しており、矛盾や不整合が至るところにみられる。
1	評価項目は書かれてはいるが抽象的で、意図や理念にとどまり、測定や対象が分からない。	事業論理の主要部分に明らかな不整合が残り、事業の流れがつかない。
2	評価項目の一部は具体だが、定義や対象の切り分けがあいまいである。	事業論理におおむね筋は通るが、部分的にあいまいさや軽度の不整合が混在している。
3	評価項目は具体的で、対象・基準や目標の情報がそろっている。	事業論理の流れは一貫しており、わずかな表記ゆれ程度の揺らぎしかない。
4	評価項目は完全に具体化され、測定方法や母集団・集計ルールまで明示されている。	事業論理は完全に首尾一貫しており、数値・時系列・役割・スコープがすべて整合している。

3. 分析方法

本レポートでは、中部圏内の自治体 11 市から各市 4 事業を選定し、全 44 件の事業評価票を対象に評価機能をテストした。対象市の選定基準は、事業評価票を公表しており、かつ事業評価票全体の文字数が可能な限り近い市とした。

また、事業の特性ごとの特徴を測るために、事業内容が近い事業（災害対策、福祉、人づくり・街づくり、その他の 4 類型）を各市から収集し、事業類型別にも集計した。

本稿では、二つの観点から分析した。一つ目は、評価スコアに基づく事業評価票の傾向分析である。評価機能による評価結果をもとに、事業評価票全体、事業類型別、自治体別の 3 点からスコアを分析する。二つ目は、評価スコアの傾向に基づく支援機能の検討である。評価スコアの傾向から支援システムに求められる必要機能を明らかにする。なお、評価機能の妥当性についての検討は付録 A にまとめた。

4. 評価結果の傾向と考察

（1）全体スコアの傾向

本レポートでは、EBPM で一般的なロジックモデル（目的→成果→測定 of 連鎖）を事業評価票の評価に適用するため、記述対象を「①事業目標（目的）②アウトカム（結果）③指標設定（結果の測定可能性）④ロジックモデル（因果経路の整合性）」の 4 領域に整理した。これは、政策評価

において目標・アウトカムの設定と、それに対応する測定指標の設定、因果経路の明確化が重要な要素として扱われているからである。

全 44 事業の平均値（標準偏差）は、記述スコア 1.71(0.33)、論理スコア 1.92(0.09)であり、論理スコアの方が高い傾向にあった（図 2）。これは、多くの事業評価票で「目標に向けてどのような因果経路を想定しているか」といった基本的なロジックが事業開始時から一定程度整理されているからだと考える。その一方で、その内容を測定可能な形で具体化して記述する段階には課題が残っていることが示唆される。

記述スコアを項目別にみると、「事業目標」のスコアは 2.20(0.68)と最も高い一方、「指標設定」は 1.30(0.94)と最も低く、標準偏差も大きい。

これは「事業目標」は事業開始時点で枠組みが定まりやすく、記述も年度を通じて洗練されるため、高得点となりやすかったと考えられる。反対に「指標設定」は、内閣府のガイドライン上も状況に応じた指標の見直し（変更）が求められる。そのため、変更理由や測定方法の説明が不足しやすく、自治体・担当者の経験や説明能力に左右されやすかったことが原因だと考える。

特に、アンケート調査を行う事業やイベント参加者の満足度といった定性的な指標をアウトカムにする事業において、

「指標設定」のスコアが低かった。ガイドラインが求めるのは、(1)調査設計・質問設計を通じて効果を測定可能な形に落とし込むこと、(2)アウトカムが見えにくい場合に代替指標を工夫すること、(3)指標採用の根拠を説明することである。本データでもこれらに関する記述が不十分な例が多くみられた。また、指標の根拠や変更理由の説明が求められるにもかかわらず、事業評価票の様式上、説明欄が十分でない場合もある。

以上より、「事業目標や基本的ロジックの整理」は全体として一定程度達成されている一方で、「効果測定の設計」および「アウトカムが見えにくい事業の指標工夫・根拠の明示」が共通の弱点であることが明らかになった。

(2) 事業類型別スコア

図3に事業類型別のスコアを示す。

事業類型別に見ると、「災害対策」関連事業の記述スコアは1.82であるのに対して、「人づくり・街づくり」「福祉」「その他」は1.60～1.72とやや低い値を示している。一方、論理スコアはいずれの類型でも1.92～1.93程度でほぼ同水準であった。

事業類型間で記述スコアに差があった原因として、災害対策関連事業は計画やマニュアルが比較的整備されており、想定シナリオや対応プロセスが明確であるため、事業目的や手段を文章としても記述しやすかった可能性がある。一方、災害対策関連以外の事業では、抽象的な理念や多目的な事業を扱うことが多く、記述の具体化に苦慮している可能性がある。

特に「人づくり・街づくり」では、指標としてイベント参加者の満足度などを設定する例が多くみられ、指標設定の難しさと相まって記述スコアが伸び悩んだ可

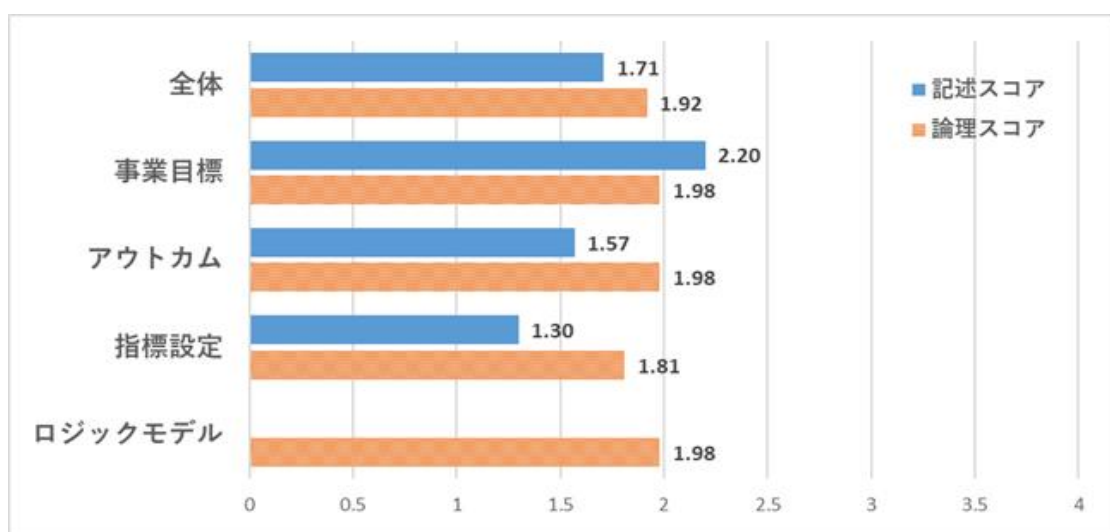


図2 全体のスコアと項目別のスコア。付録Bには全評価指標とその平均値を示している。ロジックモデルに関する指標はすべて論理スコアであるため、記載していない。

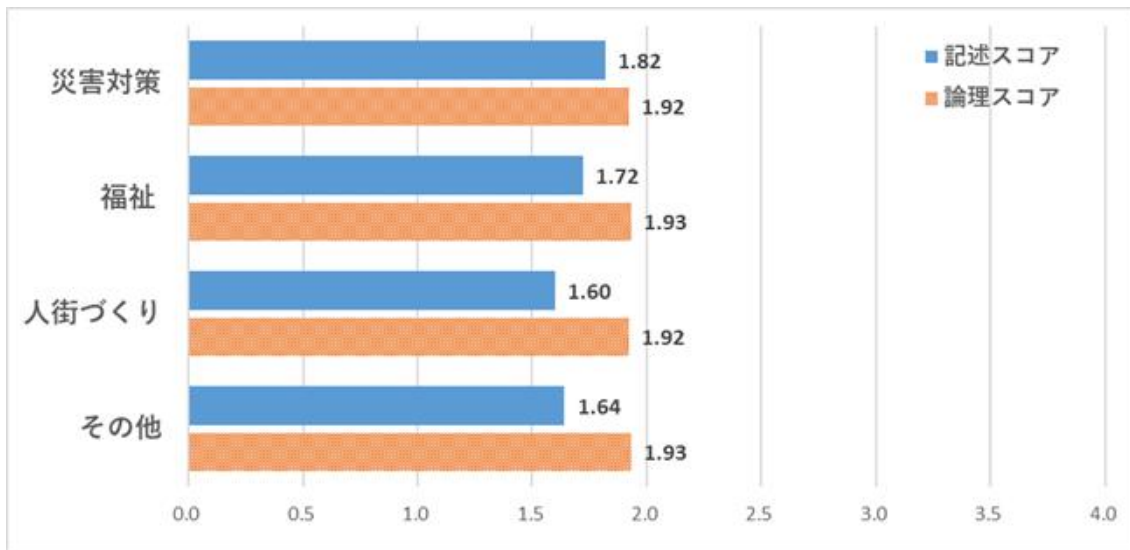


図3 事業類型別スコア

能性がある。今回は、災害対策、人づくり・街づくり、福祉、その他の4分類を行ったが、他にもみられた類型として、体制整備・計画型、イベント・参加促進型、誘致・プロモーション型などがあった。

このような類型に対しても、アウトカムが市民の態度・認知・満足度といった定性的な測定設計に依存しやすい事業（例：イベント参加率・満足度などを指標とする事業）では、記述スコア・論理スコアが相対的に低くなりやすいことが、今回の結果から示唆される。

一方で、アウトプットが明確で活動量を定量指標に落とし込みやすい事業（例：整備件数、訓練実施回数などを指標とする事業）では、記述スコアが高めとなるなど、記述（具体性）の面において類型差が表れやすい傾向が得られる可能性がある。

以上より、アウトカムの指標設計が重要な事業については、指標設計（代替指標

の提案、アンケート設計の具体化、根拠記述の補助など）に関する支援を重点的に行う必要があることが示唆される。

一方、アウトプットが明確な事業では、計画に対して実績が十分に達成されているか（進捗・実施状況）を継続的に確認することが重要である。

加えて、体制整備事業の中に広報活動が含まれる事業（支援金給付体制整備とその認知度向上活動が同一の事業の中で行われる事業）もあった。このような場合、体制整備と広報を分離して記述・評価することで、各活動の目的と成果（アウトプット／アウトカム）の対応関係が明確になり、指標設定の妥当性や改善点を把握しやすい事業評価票となるだろう。

（3）自治体別スコア

図4に自治体別のスコアを示す。自治体間での論理スコアのばらつきは、記述スコアのばらつきに比べて小さいといえ

る。自治体間で特に差が出ているのは「何を書いているか」よりも「どう書いているか」であり、事業目標や指標、ロジックモデルをどこまで具体的・一貫的に記述しているかの部分だといえる。

ロジック面の平均値はどの自治体でも一定程度確保されているため、テンプレートや評価様式の整備により形式的なロジックの枠組みは普及しつつあるものの、その中身の充実度合いは、自治体ごとに差があると解釈できる。

したがって、自治体ごとのスコアの差は、ロジックモデルを定量的な指標や測定方法へ落とし込む記述力（具体性）に起

因している可能性がある。

特に、指標の定義（母集団・算出式・集計周期）や、データ取得方法（既存統計の参照、アンケート設計、業務記録の整備）といった要素は、事業評価票の様式だけでは補いきれず、自治体内の知見蓄積や分掌体制（誰が設計し、誰がデータを持ち、誰が文章化するか）の影響を受けやすいと考えられる。

この示唆に基づけば、自治体間のばらつきを縮小し評価票の品質を底上げするためには、指標候補・測定方法・根拠記述の例示まで含む作成支援などが有効であると考ええる。

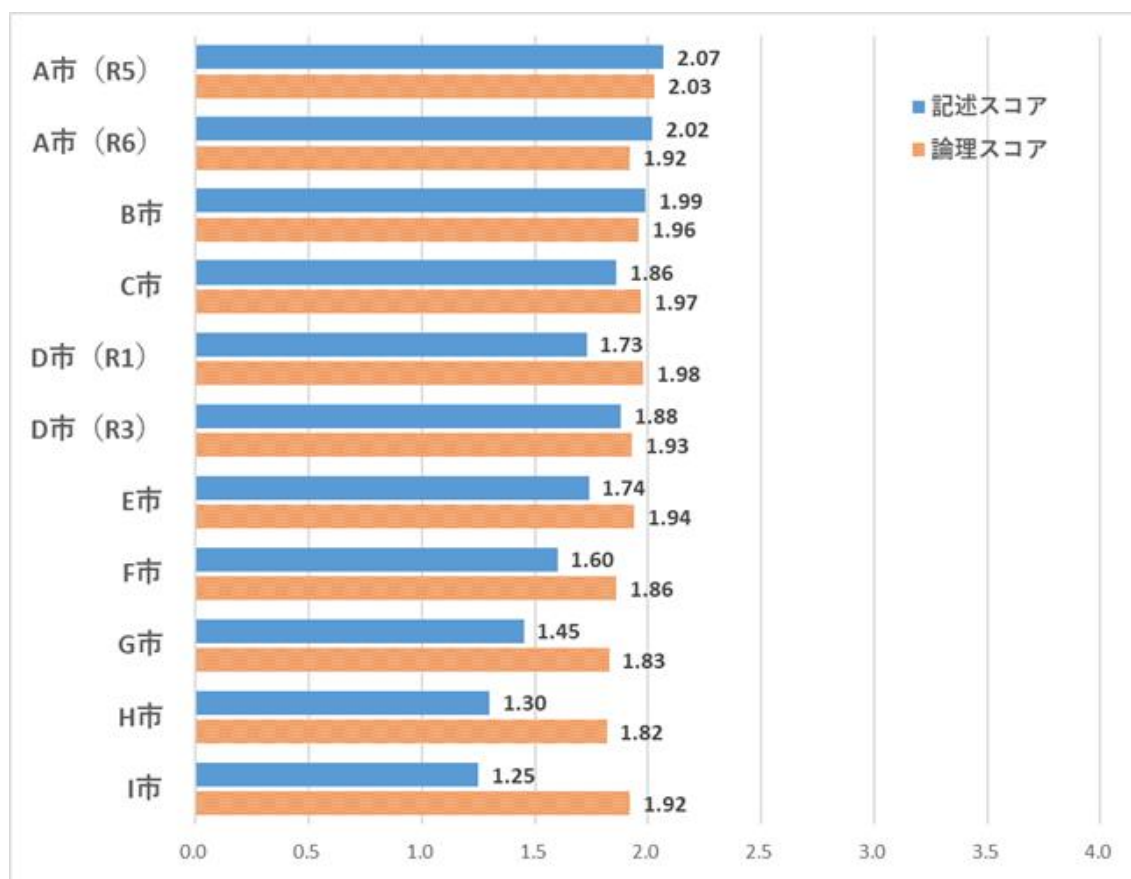


図4 自治体別スコア

(4) 実際の出力例と使用例

ここで、ある自治体の実際の事業評価票を用いて、1つの評価項目のスコアを1点分上げるために、筆者自身が作業を行った例を図5に示した。図では、評価項目「票内のアウトカム指標を採用した理由・根拠が示されているか」の記述スコアを対象としている。

5. 事業評価票作成支援システムについて

(1) 結果の傾向から考える支援機能

前章の分析から、多くの自治体で論理スコアが一定の水準だった一方、記述スコアは伸び悩みがみられた。これは、「アンケート調査の設計」や「アウトカムがみえにくい事業に対する指標設計」、「指標選定の根拠の明示」といった、高度なEBPM実践に関わる項目でスコアが低くなった自治体が多かったためである。

今回の結果を踏まえると、支援システムにおいて重点的に支援すべき領域は、単なる「様式の埋め方」ではなく、次の三点に整理できる。

1. **指標・事業評価設計の質を高める支援**：アンケート調査の質問設計や対象設定、アウトカムが見えにくい事業における代替指標の設定、指標採用の理由の記述など、評価設計そのものを支援する機能。
2. **ロジックモデルの一貫性・具体性を高める支援**：目標、対象、中間成果、

アウトカムの関係を整理し、記述抜け・矛盾を指摘した上で、より具体的な表現や指標案を提示する機能。

3. **自治体や事業類型ごとの「書きぶり」のばらつきをならす支援**：自治体間・事業類型間で記述の具体性に差があることから、他市の優良事例や同類型事業の典型的な書き方を提示し、自庁の記述の補正を支援する機能。

これらに対し、支援機能を以下のように整理できる。

- **事業類型別の指標テンプレート・チェックリスト機能**
- **アンケート設計支援機能**
- **整合性・根拠記述のチェック機能**
- **他市類似事業との比較・リファレンス提示機能**

一方で、こうした支援機能によっても、完全には解決できない課題も存在する。例えば、長期アウトカムの設定や、政策全体の中での優先順位づけ、予算配分の大枠といった事項は、政治的判断や組織内調整を伴うため、システム側で決定することは適切ではない。また、指標の選択に必要な統計データや調査結果そのものが存在しない場合には、LLM はあくまで「こうしたデータがあれば望ましい」といった助言を与えるにとどまり、実際のデータ収集や調査実施は現場のリソースに依存せざるを得ない。

評価項目：票内のアウトカム指標を採用した理由・根拠が明示されているか

	事業の成果や効果を示す指標名(説明)		令和4年度	令和5年度	令和6年度	令和7年度	最終目標値	達成率(%)
	単位	目標						
原票	①	スポーツ推進委員 研修会出席率	% 実績	90 82.1	90 83.5	90 81.7	90 -	90.8 最終目標年度 令和10年度
	式・説明	定例研修会における出席率。研修会出席率＝延べ出席者数／(委員数×年10回開催)						
	②	各種大会・つどいの 参加者数	人 実績	6,000.00 4,146	6,000.00 4,601	6,000.00 4,747	6,000.00 -	79.1 最終目標年度 令和10年度
	式・説明	年度目標は前年度実績をもとに設定						
システム出力	③	アスリート先生派遣 事業実施回数	回 実績	30 27	30 30	30 40	30 -	114.3 最終目標年度 令和10年度
	式・説明	年度目標は前年度実績をもとに設定						
	記述スコア：2	指摘：＜略＞「なぜその指標を選んだか」「目標値をどのように決めたか」の説明は限定的で、採用理由の体系的な根拠や＜略＞が示されていない。したがって、指標の採用根拠の明示は不十分であり、改善により説明の補強が可能である。 改善提案：指標設定の理由を文書で明記する（根拠＝前年度実績）、目標設定方法を数式で明示する（例：前年×1.05）、指標の測定頻度とデータ源を明記する（例：年次集計）						
	システム出力							
改善1回目	①	スポーツ推進委員 研修会出席率	% 実績	90 82.1	90 83.5	90 81.7	90 -	90.8 最終目標年度 令和10年度
	式・説明	定例研修会における出席率。研修会出席率＝延べ出席者数／(委員数×年10回開催)						
	②	各種大会・つどいの 参加者数	人 実績	6,000.00 4,146	6,000.00 4,601	6,000.00 4,747	6,000.00 -	79.1 最終目標年度 令和10年度
	式・説明	参加者数が増えることで、市民の豊かなスポーツライフに繋がる。年度目標はこれまでの実績をもとに設定。						
システム出力	③	アスリート先生派遣 事業実施回数	回 実績	30 27	30 30	30 40	30 -	114.3 最終目標年度 令和10年度
	式・説明	アスリートとの接点を増やすことは、市民のスポーツへの関心向上に繋がる。年度目標はこれまでの実績をもとに設定。						
	記述スコア：2	指摘：＜略＞。「参加者数が増えることで豊かなスポーツライフに繋がる」「興味を持つようになるため」といった文言は因果関係の想定を述べるにとどまり、期待する具体的な成果＜略＞は不十分である。＜略＞ 改善提案：各指標と期待する成果の因果を定量化する（例：参加者増→健康指標5%改善）、指標の測定頻度とデータ源を記載する（例：年1回の集計を施設集計で実施）、目標設定の根拠を示す（例：過去3年の実績平均を基準に年次目標を設定）						
	システム出力							
改善2回目	①	スポーツ推進委員 研修会出席率	% 実績	90 82.1	90 83.5	90 81.7	90 -	90.8 最終目標年度 令和10年度
	式・説明	【採用理由】スポーツ推進委員の研修受講は、指導・運営能力の向上を通じて事業の質を高める。 【算定】研修会出席率＝延べ出席者数／(委員数×年10回開催)。【目標】直近実績から90%を目標とする						
	②	各種大会・つどいの 参加者数	人 実績	6,000.00 4,146	6,000.00 4,601	6,000.00 4,747	6,000.00 -	79.1 最終目標年度 令和10年度
	式・説明	【採用理由】大会・つどい等への参加が増えることは、市民のスポーツ実施の裾野拡大を示すため。【算定】各事業の参加者名簿に基づき、年度の延べ参加者数を集計。【目標】直近3年実績を基準に、改善目標として6,000人を設定する						
システム出力	③	アスリート先生派遣 事業実施回数	回 実績	30 27	30 30	30 40	30 -	114.3 最終目標年度 令和10年度
	式・説明	【採用理由】アスリートとの接点を増やすことは、市民(特に子ども等)のスポーツへの関心・参加意欲の向上に繋がるため。【算定】今年度の実施回数(回)。【目標】前年実績と前年目標の平均を用いて設定する。						
	記述スコア：3							
	システム出力							

図5 実際の評価票を用いて評価を行い、改善する様子（筆者が作業） 上段：原票、指標②、③についての指摘が中心。中央：原票への改善提案を基に手を加えた票（1度目）、依然スコアは2である。下段：何度かの改善を通して、最終的なスコアが3となった票。下段票ではアウトカム設計への詳細な記述を増やした。しかし、このような記述を行うには、共通の票構造から変更する必要があると、記述を数語追加する程度では、十分にスコアアップが望めない可能性も示唆されている。

(2) 開発プロセスを通して得た要望機能

本レポートの執筆に際し、さまざまな専門家と評価票支援システムについて議論を行ってきた。その過程で考察した機能について報告する。

- 指標モニタリングとログ蓄積
- ロジックモデルからの評価票生成
- ロジックモデルを強化するためのデータ要求機能
- ロジックモデル構築支援
- 途中評価・改善提案の生成

また、筆者らは 2025 年 8 月頃に愛知県のある自治体に対してインタビューを行った。本節では実際に評価票作成経験のある職員に行ったインタビューから得られた要望機能についてもここで報告する。

- 他市類似事業との比較・参照
- 事業単位の切り分け自動化の補助
- 予算紐づけの自動化
- 評価機能と連動したフィードバック提示
- 職員教育機能

(3) まとめ

本レポートで挙げた要望機能は、事業の進行段階によって必要度や有効性が異なることが予想される。

そこで、事業評価票の作成を単発の書類作業としてではなく、事業設計・実施・評価を一体的に高めるプロセスとして運用するために、支援機能を事業のライフサイクルに沿って整理する。

本レポートでいう事業のライフサイクルとは、年度単位の事業運営における(1)構想・設計段階、(2)実施・運用段階、(3)評価・意思決定段階の三段階に、加えて、(4)評価で得られた知見を次年度の企画・設計へ確実に還流させるための学習・組織化段階(横断機能)を含む四段階である。

本レポートでは、行政の総合計画のような複数年の政策体系を管理する枠組みを大きな PDCA サイクルとして、年度単位で事業の運用状況を点検し次年度方針に接続する事業評価票(事業報告)を小さな PDCA サイクルとして考察する²。

本節では、両者のうち事業評価票に対応する小さな PDCA サイクルを対象に、支援機能を整理する。

この観点から、構想・設計段階は小さな PDCA の Plan、実施・運用段階は Do を中心にモニタリングなどの小さな Check / Action を伴う段階、評価・意思決定段階は Check/Action として次年度の Plan へ接続する段階とみなせる。さらに、学習・

²政策は政策立案 (Plan)、政策実施 (Do)、政策評価 (Check)、見直し (Action) の大きな PDCA サイクルによって構成され、政策実施は具体的部分の決定 (Plan)、実施 (Do)、モニタリング・評

価 (Check)、見直し (Action) の小さな PDCA サイクルによって構成されるとされている [厚生労働省]。

組織化段階は、評価で得られた知見をテンプレートや教材などの形で定着させ、次年度以降の Plan を強化する横断的な段階として位置づけられる。

事業のライフサイクル全体を通じて支援機能を配置することで、事業評価票の作成を単発の書類作業としてではなく、「事業設計・実施・評価を一体的に高めていくプロセス」の一部として位置づけることが可能となる。

本レポートで試作した評価機能は、その中でも特に評価時点におけるフィードバックの役割を担うものであり、将来的には上記の各機能と連携させることで、より包括的な事業評価票作成支援システムへと発展させうると考えられる。

構想・設計段階（Plan）における支援

- ロジックモデル構築支援
- ロジックモデルを強化するためのデータ要求機能
- 他市類似事業との比較・参照
- 事業単位の切り分け自動化の補助

実施・運用段階（Do）における支援

- 指標モニタリングとログ蓄積
- 予算紐づけの自動化
- 途中評価・改善提案の生成

評価・意思決定段階（Check、Action）における支援

- ロジックモデルからの評価票生成
- 評価機能と連動したフィードバック

提示

学習・組織化段階（横断的：次の Plan を強化）における支援

- 職員教育機能

6. おわりに

本レポートでは、EBPM の潮流を踏まえ、中部圏内の 11 自治体・全 44 件の事業評価票を対象として、LLM を用いた評価機能の試作と検証を行い、スコアの結果を分析した。

事業類型別・自治体別・評価項目別の分析を通じて、多くの自治体で事業目標や指標の「基本的な整理・記述」は一定程度達成されている一方で、アンケートを用いた効果測定的设计や、アウトカムが見えにくい事業における指標の工夫・根拠の明示といった、より高度な EBPM 実践に関わる部分が共通の弱点となっていることが明らかになった。また、論理スコアは項目別で大きな差はなかったものの、記述スコアには項目別で差がみられることが分かった。これらの結果を踏まえた、事業評価票作成支援システムが備えるべき機能の方向性も整理した。

本レポートで扱ったのは、あくまで「事業評価票の記述内容がガイドラインにどの程度沿っているか」の観点であり、プロセス評価（会計不正がないか、事務処理が適正に行われているかなど）や、政策評価・事業成果評価（十分な事業効果があっ

たかどうかなど) そのものは対象としていない。

また、本レポートの評価スコアは、事業そのものの優劣や政策の有効性を直接判定するものではなく、「事業評価票が、EBPM の観点から望ましい情報をどこまで含んでいるか」を測る指標であることに留意が必要である。

最後に、「AI が出力したスコアを過信し、何らかの事故が起きた場合の責任」といった倫理的な問題が挙げられる。このような問題は支援システムの機能のみならず、法制度面からも検討する必要がある。これらの議論の参考として、自動運転などにおける法整備が参考となる可能性がある。また、LLM のハルシネーションやバイアスの問題に留意しつつ、人間の専門的判断を補完する形での活用範囲と限界を明示することも重要である。

今後の展望としては、本レポートで得られた知見を基に、事業の設計段階から評価時点までを通貫する支援機能と、本レポートで試作した評価機能とを統合し、事業設計・実施・評価を一体的に支援するシステムの具体化が挙げられる。

参考文献

厚生労働省「3.政策策定と評価 政策の実施・評価・見直し」

<https://www.mhlw.go.jp/content/10900000/001524235.pdf>.

杉谷和哉 (2022) 『政策にエビデンス

は必要なのか EBPM と政治のあいだ』ミネルヴァ書房.

杉谷和哉 (2024) 『日本の政策はなぜ機能しないのか? EBPM の導入と課題』光文社新書.

内閣官房行政改革推進本部事務局

(2025)「行政事業レビューシート作成ガイドブック」2025年3月31日.

https://www.gyokaku.go.jp/review/img/R07sakusei-guidebook_ver1-2.pdf

付録

A. 評価の妥当性の検討

LLM は高い言語評価性能を持つ反面、ハルシネーションに代表されるように、事実と異なる内容をもっともらしく生成してしまう問題も指摘されている。

そのため、本節では、LLM を本レポートに用いるにあたって、その評価結果がどの程度妥当といえるのかについても検証を行う。

本来であれば、専門家などによる評価スコアを基準として、本システムが予測したスコアの妥当性を定量的に検証すべきである。

しかし、時間的・人的リソースの制約や評価依頼に伴う調整コストの観点から、本レポートで十分な規模の専門家評価データを収集することは困難であった。

そのため、定性的な評価にとどめ、文字数とスコアの相関関係と外部指摘との比較による妥当性の検証を行う。

（１）文字数とスコアの相関関係

より長く書けば評価（特に記述スコア）が上がる可能性があった。そこで、事業評価票の文字数とスコアの相関係数を調べた。その結果、記述スコアで 0.24、論理スコアで 0.36 と弱い正の相関にとどまっていることが分かった。

つまり、「長く書けば評価が上がる」といった単純な関係ではなく、ある程度の分量は必要であるものの、その中でどれだけ論理構造や根拠が明確に説明されているかが、LLM による評価にとって重要であると解釈できる。

（２）第三者による指摘との比較

第三者による評価票への指摘と本システムの評価結果を比較することで、評価機能の妥当性を補足的に検証する。

一部の事業評価票には、外部評価（評価委員会や専門家によるコメント）が記載されている。外部評価では、「事業の継続、縮小、廃止」などを検討し、結論付けることが多い。しかし、1 章でも示した通り、本システムはあくまでも評価票がガイドラインに沿っているかを判定することが目的であり、これらと比較することは本来の目的から外れている。

そこで、本レポートでは、このような事業の今後を決める外部評価ではなく、評価票の記載に対するコメントとしての外部評価との比較を行い、あくまでも事業評価票の内容および事業内容についてのコメントとの比較を行う。具体的には、指

摘コメントが付されている事業を抽出し、その指摘内容とシステムが出力した指摘を比較した。

対象とした評価票は、ある自治体の市民の自主的な防災活動を自治体が支援する事業であり、この事業の成果指標は「自主防災倉庫設置数」のみであった。

この事業に対して専門家は、「自主防災倉庫の設置数を活動指標にするのはピンと来ない」と述べ、あわせて「中間成果は貸与された資機材で訓練活動が行われているかを測れるものであるべき」と指摘した。それに対して本システムは、「倉庫設置数がアウトプット寄りであり、最終成果に資することを示す因果検証手法が欠落している」点を指摘した。

このように、専門家の指摘とシステムの指摘が近い問題意識を共有しているといえる。これを、人間の評価観点に近い精度で本システムが評価票の問題点を検出できている可能性を示す補助的根拠とする。

ただし、本比較は単一事例に基づくものであり、専門家の指摘内容を網羅的に対応付けたわけではない点に留意が必要である。また、専門家の指摘は具体的かつ実務的であり、現場の状況を踏まえて課題を特定したうえで改善の方向性まで踏み込んだ提案（新たな指標の提案）になっているが、システムはあくまでも問題点が存在することを示すにとどまっている。つまり、専門家の指摘の方が、実際に職員

が評価票へ反映しやすいと考えられるため、システムもこのような指摘ができるように改善する必要があることが示唆された。

B. 項目ごとのスコア

評価項目	記述	論理
事業目標と長期アウトカムには整合性がある課題・目的が設定されているか	2.04	2.00
抽象的な事業目標（理念や将来像）が、定量・定性の指標や行動計画として具体化されているか	2.04	1.98
定性的な評価しかないが、定量的な評価をするための指針が書かれているか	1.98	2.02
事業目標が指標や数値を含め、担当課の年度方針として明確に示されているか	1.96	2.00
現在の指標が過去（今年度よりも前）に設定した指標から変更されているが、アウトプットからのつながりや成果実績及び目標値の根拠として、その理由についても補足されているか	1.86	2.00
短期・中期の目標と指標が母集団や基準値と整合し、現実的かつ測定可能な形で設定されているか	1.83	1.96
広報・周知型事業（市民を対象としたイベント・教室実施等の事業）だが、ターゲットとなるステークホルダーが具体的に（なぜそのターゲットなのかも含めて）定義されているか	1.77	2.00
最終成果に到達するために、内容および因果ロジックを伴った中間成果指標が適切に設定されているか	1.77	1.96
票内のアウトカム指標を採用した理由・根拠が明示されているか	1.50	2.04
施設整備型事業だが、施設着工から運用開始までのマイルストーンが明確か	1.50	2.00
抽象的な長期アウトカムが設定されているが、他事業の影響を排除しづらい状況と評価方法が説明されているか	1.51	1.97
アウトカム指標の目標値について、過去実績や計画に基づく数値根拠・算定ロジックが明示されているか	1.44	1.92
広報・周知型事業（市民を対象としたイベント・教室実施等の事業）だが、周知活動のインパクトを測定する指標（リーチ数・行動変容など）が設定されているか	1.35	1.96
事業目標が閣議決定等の文言引用だけに留まらず、実際の事業目的や課題が具体的に記載されているか	2.25	
継続型事業（市職員等の行政職員を対象とした研修等の事業）だが、研修効果の測定(受講後アンケート・スキルテスト等)の仕組みが用意されているか	2.00	
定性的なアウトカムが設定されているが、定量的な指標を設定できるのにしていないか	1.92	
直接解決を目指す具体的な社会課題が明示されているか	1.92	
事業票内に事業目的が複数あるが、各目的に応じた効果検証を行うための指標を設定できているか	1.88	
事業単独で達成を目指す事業目的 と 政策全体としての大きな事業目的 の区別有無	1.79	
ポジティブな変化（アウトカム）が見えにくい迅速性（質の改善）を測る事業だが、「●	1.75	

人を配置」、「●件の設備投資」などの活動を測る指標がきちんと設定されているか		
事業票内で短期・長期アウトカムが設定されていないが、本当に不要（発現経路がシンプルな事業）なのかにに対して検討されているか	1.63	
事業票内で短期・長期アウトカムが設定されていないが、効果測定タイミングを設定すべきではないのかに対して検討されているか	1.55	
実績値欄や目標値欄に空欄があるが、その理由が説明されているか	1.31	
ポジティブな変化（アウトカム）が見えにくい実効性／効率性（質の改善）を測る事業だが、活動が的確・有効に行われているかを見る指標が設定されているか	1.09	
専門評価・外部評価の活用をしているか	1.00	
アンケートを用いた定性的な評価をしているが、単なる満足度調査ではなく、対象者の意識や行動の変化につながったかを測定できているか	0.98	
アンケートを用いた定性的な評価をしているが、事業効果を把握し改善に活かすため、質問内容・方法・時期・対象者を明示した調査設計を行っているか	0.88	
ポジティブな変化（アウトカム）が見えにくい実施規模(量)を測る事業だが、「●日以内に情報が回付される」などのスピードを測る指標が設定されているか	0.54	
アクティビティが中間成果を経て最終成果に至る因果の道筋と内容が記載されているか		2.04
エビデンスがエピソードベースだが、工夫・改善 → 成果につながる論理がちゃんと示されているか		2.00
ポジティブな変化（アウトカム）が見えにくい迅速性（質の改善）を測る事業だが、「●人を配置」、「●件の設備投資」などの活動を測る指標にきちんと整合性はあるか		2.00
中間成果目標に対応する定量・定性指標が、母集団や成果内容と論理的に整合しているか		2.00
抽象的な長期アウトカムが設定されているが、事情を踏まえて妥当な設定になっているか		2.00
ポジティブな変化（アウトカム）が見えにくい実施規模(量)を測る事業だが、「●日以内に情報が回付される」などのスピードを測る指標にきちんと整合性はあるか		1.88
事業票内に事業目的が複数あるが、目的ごとに対象集団・指標が整理され、効果発現経路が一貫しているか		1.88
ポジティブな変化（アウトカム）が見えにくい実効性／効率性（質の改善）を測る事業だが、活動が的確・有効に行われているかを見る指標にきちんと整合性はあるか		1.82
抽象的な長期アウトカムが設定されているが、達成時期も現実的な範囲（5年以内）で設定されているか		1.69