Uncovering the Risk of Evaluation Formalism: A Debiased LLM-as-a-Judge Study in Japan's Government Project Reviews

Masahiro Asami[1*] and Kai Fukunaga[2]

[1*]Independent Researcher, Tokyo, Japan.
[2]Research Institute of Economy, Trade and Industry, Tokyo, Japan.

*Corresponding author(s). E-mail(s): m.asami.moj@gmail.com;
Contributing authors: fukunaga.kaii@gmail.com;

Abstract

This study examines the challenges of applying the *logic model* framework to *Plan* projects, which include Proof of Concept (PoC) and research initiatives. Despite efforts to standardize the framework, we argue that *Plan* projects often produce low-quality *logic models* due to inherent limitations in the framework. Using a newly released administrative project review system (RS system), we assess thousands of *logic models* for *Plan* projects through a hybrid approach combining Large Language Models (LLMs) and regression adjustments, which correct for known biases in LLM scoring, particularly the tendency to favor longer inputs. Our findings suggest that *Plan* projects exhibit lower-quality models on average compared to other project types, highlighting the risk of evaluation formalism. We propose an alternative management and evaluation framework tailored to *Plan* projects, emphasizing hypothesis testing, evidence gap analysis, and feasibility assessments. Additionally, this study demonstrates the potential for scalable monitoring of government evaluations using LLMs.

Keywords: logic model, EBPM, LLM-as-a-Judge, NLP, causal inference

1 Introduction

In recent years, the Japanese government has increasingly emphasized the promotion of Evidence-Based Policy Making (EBPM) as a guiding principle for public administration. Among various initiatives undertaken to promote EBPM, the Administrative

Project Review (*Gyosei Jigyo Review*) has recently attracted particular attention due to its experimental reforms and intensified efforts to embed evidence-based practices. In particular, the introduction of standardized *logic model* descriptions—initiated in 2022 and fully implemented in 2023—has required ministries and agencies to clearly articulate the causal pathways linking their projects' inputs, activities, outputs, and intended societal outcomes [21]. Furthermore, the launch of the *RS System* (Review Sheet System) has centralized the management and public disclosure of review data, significantly enhancing accessibility to administrative project information.

Despite these advances, concerns persist regarding the operational challenges of the Administrative Project Review. Previous studies have highlighted issues such as *evaluation fatigue* and *goal displacement*, where evaluation practices risk becoming ritualistic rather than substantively improving policy outcomes [21, 23, 24].

In this study, we focus particularly on the structural issue that the Administrative Project Review applies uniformly to all publicly funded projects, regardless of their nature or developmental stage. We raise concerns about the current framework's expectation that even *Plan* projects—such as Proof of Concept (PoC) and research initiatives, which correspond to the "Plan" phase of the PDCA cycle—be managed within the same evaluative structure, albeit with somewhat relaxed requirements [7].

1.1 Policy Contribution

Leveraging the newly released *RS System*, we compiled the first comprehensive dataset registered in fiscal year 2024. Our empirical analysis reveals that *Plan* projects tend to produce *logic models* that are substantially less coherent and internally consistent than those of other projects. We interpret this systematic weakness as an empirical marker of *evaluation formalism*—a situation in which compliance with reporting rules becomes ritualistic and ceases to foster substantive learning. On the basis of these findings, we offer concrete policy recommendations for revising the current Administrative Project Review guidelines.

1.2 Methodological Contribution

Our second contribution is methodological. Because the analysis relies on thousands of unstructured Administrative Project Review sheets, we adopt the emerging *LLM-as-a-Judge* paradigm to automate scoring at scale [4]. Yet recent evidence shows that LLM evaluations can be systematically biased, with one commonly reported issue being a *verbosity bias*, whereby longer inputs are rewarded with higher scores [4, 16, 19, 28, 29]. This matters especially here, because review sheets for *Plan* projects are, on average, substantially longer than those for other projects. To obtain the most unbiased assessment possible, we introduce the ***D**ebiased **L**LM-as-a-**J**udge **2**-stage* (**DLJ-2**) framework, which couples an initial LLM scoring pass with a streamlined statistical adjustment that explicitly controls for observed covariates—such as input text length in the present case.

Additionally, the present analysis showcases how LLMs can underpin *scalable, continuous monitoring of government evaluations*, opening the door to even broader applications in future oversight efforts.

2 The Administrative Project Review System in Japan

2.1 Institutional Overview

The **Administrative Project Review** (*Gyosei Jigyo Review*) was institutionalized in 2013 through a Cabinet Decision[1]. It requires ministries and agencies to conduct annual self-assessments of their budgetary projects, evaluating them in terms of necessity, efficiency, and effectiveness. Following the "one project, one sheet" principle, each project is documented in a standardized *Administrative Project Review Sheet*, and selected projects undergo external third-party review. Public evaluations are conducted through the *Public Process*, while comprehensive cross-ministerial verifications are performed during the *Autumn Review* [24].

A key feature of the Administrative Project Review is the emphasis on explicitly constructing a *logic model* that links project inputs, outputs, and outcomes. This emphasis was further strengthened through reforms initiated in 2022 and fully implemented in 2023, reflecting efforts to promote more systematic EBPM practices [21].

2.2 Logic Model

A *logic model* is a structured representation of the causal pathways linking program resources (inputs), activities, immediate outputs, and longer-term outcomes. It provides a theoretical framework to clarify how administrative actions are expected to generate social value, thereby supporting coherent project design, implementation, and evaluation [27].

In Japan, the use of *logic models* has been increasingly emphasized within government evaluation guidelines. Recent guidelines [6, 7] require ministries to explicitly map the causal chain of "Inputs → Activities → Outputs → Outcomes" in a clear and traceable manner.

Specifically, ministries are instructed to:

- Define project objectives and target outcomes;
- Identify necessary inputs and planned activities;
- Specify measurable outputs and intermediate indicators;
- Articulate the assumed causal linkages to final societal impacts;
- Employ the *logic model* framework for ongoing monitoring and adaptive management.

Through these requirements, the Administrative Project Review aims not merely to assess financial or operational performance, but to embed causal reasoning and empirical verification into the policy-making process [7].

Fig. 1 Overview of the RS System interface https://rssystem.go.jp/(Accessed: April 26, 2025). The left box ("Search Sheets") allows users to search for specific budgetary projects by setting various criteria. The right box ("Aggregate and Analyze Sheets") provides functionalities for aggregating and analyzing review sheets by ministries and agencies. In addition, the "CSV Data" button located at the top menu enables users to download the entire dataset in bulk for external analysis.

2.3 Recent Initiatives: RS System

The RS System[2] (Review Sheet System) is an integrated online platform for managing the Administrative Project Review. Developed by the Administrative Reform Promotion Headquarters, Cabinet Secretariat of Japan, it aims to promote EBPM and enhance administrative digital transformation efforts (Figure 1).

The system was officially launched in April 2024, enabling ministries and agencies to prepare and manage their *Administrative Project Review Sheets* directly through a centralized interface. Subsequently, in September 2024, the completed review sheets were made publicly accessible through the *Administrative Project Review Visualization Website*, allowing cross-ministerial search, aggregation, and analysis according to user interests.

The RS System database includes *review sheets* (for individual projects), *segmented sheets* (for grouped projects or policy segments), and *fund management sheets* (for subsidy and fund-related projects), covering data from fiscal year 2021 onwards. By centralizing previously decentralized and manually managed processes, the RS System significantly reduces administrative burdens and improves operational efficiency, while simultaneously supporting governmental transparency and public accessibility to policy information.

[1] Cabinet Decision on the Implementation of the Administrative Project Review (April 5, 2013) https://www.kantei.go.jp/jp/singi/gyoukakusuisin/dai2/siryou01.pdf(Accessed: April 26, 2025)
[2] The official website is available at https://rssystem.go.jp/ (Accessed: April 26, 2025)

3 Problem Framing and Analytical Focus

3.1 Institutional Critiques in Prior Research

While the Administrative Project Review has contributed to enhancing transparency and accountability, recent analyses have identified persistent challenges. First, institutional redundancy between the Administrative Project Review and other evaluation mechanisms, such as the Policy Evaluation System, has been noted as a source of inefficiency.[3] Second, Sugitani [21], Tokuda [23] highlight *evaluation fatigue* and *goal displacement*, where evaluation activities risk becoming procedural formalities, detached from substantive policy improvement.

3.2 Our Analytical Focus: Evaluation Formalism in *Plan* Projects

We argue that this issue is relevant in the context of the current RS system. Nearly all budgetary projects are registered within the system. The official tutorial [7] specifies that projects such as *Proof of Concept (PoC)* and *research* should be included under the *Plan* phase of the PDCA cycle, noting that while quantitative outcomes are not necessarily required, their inclusion is expected.

It may be inappropriate to assess all budgetary projects using a uniform format. The *logic model* is designed to evaluate the *Do* phase of PDCA, and it may not be appropriate for the preparatory *Plan* phase. Ono [15] argues that imposing unreasonable outcome-setting can lead to *evaluation formalism*—a form of institutional *decoupling* in which the review ritual is maintained for legitimacy while substantive learning is neglected [14]. In this context, forcing the use of *logic models* for *Plan* projects may exacerbate the risk of *evaluation formalism*, as it imposes a rigid framework that may not be suitable for the preparatory phase.

In this study, we therefore focus specifically to *Plan* projects: leveraging the RS System's large number of *logic models*, we test whether forced and internally inconsistent outcome setting—an empirical marker of evaluation formalism—appears more frequently in *Plan* than in *Do* projects.

The overall conceptual framing of our study is summarized in Table 1.

4 Data and Definition of Target Projects

We extracted project data from the RS System described in subsection 2.3 and evaluated the quality of their *logic models*. A total of 5,896 projects were registered for the fiscal year 2024. For the analysis, we utilized the `main_expenses` column from the RS system, extracting those main expenses categories that contained over 100 samples. The categories *Other Expenses* (2,472 cases) and *Economic Cooperation Expenses* (207 cases) were excluded from the analysis as they are outside the scope of this study.

[3] Besides the Administrative Project Review, Japan maintains multiple administrative evaluation mechanisms, including evaluations under the Policy Evaluation Act, administrative evaluations based on the Ministry of Internal Affairs and Communications Establishment Act, budget execution surveys based on the Public Finance Act and the Accounting Act conducted by the Ministry of Finance, and audit inspections performed by the Board of Audit [24].

Table 1 Conceptual Overview of the Study

Research Motivation	Potential risk of *evaluation formalism* in current policy review practices
Hypothesis	The uniform application of a single *logic model* format across all project types may be inappropriate
Empirical Marker	*Logic models* of *Plan* projects exhibit systematically lower structural consistency and coherence
Empirical Strategy	Large-scale analysis using FY2024 RS System data; development of the DLJ-2 (Debiased LLM-as-a-Judge 2-stage) framework, combining LLM scoring with bias-adjusted regression modeling

Additionally, we operationalize *Plan* projects as follows. Since the RS system does not include a specific column for this, we defined this variable by checking whether the project's *output* column contains terms such as *Research* or *PoC*. This analysis focuses on 2,835 *logic models*, as summarized in Table 2.

Table 2 Main Expenses Categories and Plan Projects in the RS System for Fiscal Year 2024 (with over 100 projects registered per category). The categories *Other Expenses* (2,472 cases) and *Economic Cooperation Expenses* (207 cases) were excluded from the analysis as they are outside the scope of this study.

Main Expenses	**Number of Projects**	***Plan* Projects**
Science and Technology Promotion Expenditure	567	87
Energy Policy Expenditure	390	93
Public Health and Sanitation Expenditure	358	39
Food Security and Supply Expenditure	355	72
Defense Expenditure	320	22
Social Welfare and Living Assistance Expenditure	285	31
Employment and WorkersCompensation Expenditure	218	18
Educational Promotion Grants	216	34
Small and Medium Enterprise Support Expenditure	126	34

4.1 Plan-Project Classification Validation

To validate our simplified classification, we sampled 50 projects from the *Small and Medium Enterprise Support Expenditure* category and compared the heuristic labels with `GPT-4o` classifications. The resulting confusion matrix is reported in Table 3.

Overall agreement was high (accuracy = 0.90, precision = 0.75, recall = 1.00), indicating that the heuristic introduces only a few false positives and no false negatives in this sample.

Table 3 Confusion matrix comparing the rule-based heuristic with `GPT-4o` classification for identifying *Plan* projects in the *Small and Medium Enterprise Support Expenditure* sample ($n = 50$). Rows correspond to GPT-4o labels (reference standard) and columns to heuristic labels.

GPT-4o (reference)	**Heuristic**	
	Non-Plan (0)	**Plan (1)**
Non-Plan (0)	30	5
Plan (1)	0	15

5 Debiased LLM-as-a-Judge 2-stage framework

In this section, we present ***D**ebiased **LLM**-as-a-**J**udge **2**-stage* (**DLJ-2**) framework, which combines LLMs and traditional statistical methods. This approach was designed to enable scalable assessment of *logic model* quality while addressing potential biases inherent to LLM-generated outputs. By integrating human-like reasoning capabilities of LLMs with post-estimation adjustments through regression modeling, we aim to achieve more analytically robust and unbiased evaluations.

1st stage: LLM Scoring

Let $\mathbf{x}_i$ denote the i-th logic-model text and let $\mathcal{C}$ be the set of evaluation criteria inserted into the prompt. A frozen LLM parameterised by θ induces a conditional distribution $p_\theta(y \mid \mathbf{x}_i, \mathcal{C})$ over the judgement output y (a numerical score with an accompanying rationale). Feeding the concatenated prompt $\mathbf{x}_i \oplus \mathcal{C}$ to the model yields

$$y_i \;=\; g_\theta\big(\mathbf{x}_i \oplus \mathcal{C}\big), \tag{1}$$

where $\oplus$ denotes string concatenation and g_θ is the LLM's sampling function [4].

2nd stage: Debiasing Regression

Define variables $(d_i,\, \ell_i)$ for each sheet: $d_i \in \{0, 1\}$ is a *Plan* dummy, ℓ_i the $\mathbf{x}_i$'s length as a covariate. The score y_i follows

$$y_i = \beta_0 + \beta_d d_i + \beta_\ell \ell_i + \epsilon_i, \tag{2}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ captures *unobserved, project-specific quality factors* that remain after controlling for d and ℓ.

Parameter of Interest: Misfit Cost (Plan Effect)

The *Plan–phase misfit cost* is defined as

$$\hat{\beta}_d \;=\; \mathbb{E}[\beta_d \mid \mathcal{D}] \tag{3}$$

where $\mathcal{D} = \{y_i, d_i, \ell_i\}_{i=1}^{N}$ denotes the observed data.

Equations 1-3 together constitute the **DLJ-2** framework: an LLM-derived score followed by regression adjustment, with $\hat{\beta}_d$ quantifying the degree of mismatch between the evaluation template and the nature of *Plan* projects.

The subsections that follow (i) survey prior work on the LLM-as-a-Judge paradigm, (ii) explain our prompt-construction strategy, and (iii) specify the bias we address in the present analysis.

5.1 Prior Work on LLM-as-a-Judge

The rationale for utilizing LLMs as evaluators—commonly referred to as the *LLM-as-a-Judge* framework—originates from well-documented limitations of human judgment, including susceptibility to subjective bias, inconsistency across raters, and gradual drift in evaluative standards over time. LLM-based evaluation offers a more stable and reproducible process that adheres consistently to predefined criteria [29].

This approach has gained traction across a wide range of domains. In the legal and judicial context, LLMs have been employed to simulate judicial decision-making and assess legal arguments [2, 4, 5, 12]. In the political domain, LLMs have been used to enhance transparency and accountability: for example, Lilley and Townley [9] applied GPT-based models to classify and visualize UK parliamentary voting records, while Asatryan et al. [1] conducted a large-scale sentiment analysis of EU cohesion policy evaluations to detect potential biases and ritualistic tendencies in official assessments.

Our study follows this growing line of research by applying the LLM-as-a-Judge paradigm to the Japan's *Administrative Project Review* system—a newly standardized nationwide framework aimed at promoting evidence-based policy.

5.2 Our Prompt Design

We designed the following evaluation procedure.[4] Figure 2 illustrates the overall evaluation procedure implemented in this study.

[4] Pairwise comparison is a well-established evaluation method that has significantly influenced various academic and applied fields [18]. As noted by Liu et al. [10], the agreement between LLM-based evaluations and human judgments tends to be higher in the context of pairwise comparisons than in score-based evaluations. Moreover, numerous studies have shown that pairwise comparison outperforms alternative evaluation methods in terms of positional consistency and reliability [11, 29].
However, in the present study, the total number of possible pairwise comparisons among the 2,835 projects is ${}_{2,835}C_2 = 4,017,195$, making full pairwise evaluation impractical due to computational and API cost constraints.
Therefore, we designed an alternative evaluation procedure described in the following subsection.

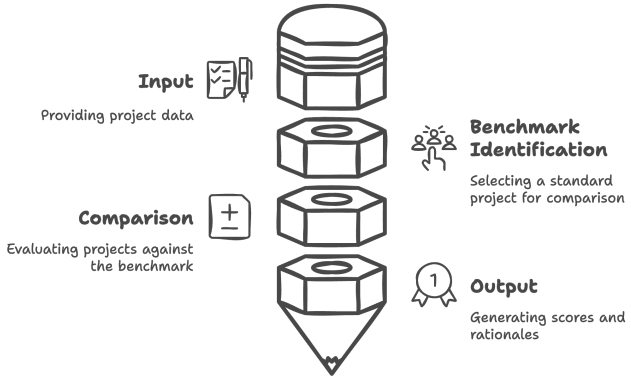

Fig. 2 Flowchart of the LLM-as-a-Judge Evaluation Procedure

The following process was carried out for each Main Expenses Category using the LLM: First, the LLM was provided with an average benchmark for each Main Expenses Category. Next, the LLM evaluated each project on a five-point scale, referring to the aforementioned benchmark. To enhance output stability and reduce the risk of hallucinations, the model was instructed to provide not only a numerical score but also a textual rationale for each evaluation, a method shown to improve LLM judgment reliability [13]. [5]

The evaluation criteria focused on the *consistency of outputs and outcomes*, including the consistency from outputs to outcomes and the alignment of short-term, medium-term, and long-term outcomes (i.e., the presence or absence of leaps in logic). This is a valid criterion for assessing *logic models*, as it aligns with established standards for *logic model* evaluation [6, 27]. It is important to note that the objectivity and measurability of the metrics are explicitly excluded from this criterion; it is self-evident that the *Plan* projects are inferior in these aspects, and the official tutorial explicitly disclaims these aspects for the projects [7].

The prompt and LLM model used for evaluation are detailed in Appendix A.

[5] We also account for *position bias*—a systematic tendency of LLMs to vary outputs based on the order of inputs [20, 26, 28]—by randomly shuffling the order of project entries before submission.

(Partial excerpt from the evaluation prompt)
Evaluation Criteria: Consistency Between Outputs and Outcomes

The evaluation focuses on the logical consistency between the stated outputs and outcomes of each project, as interpreted from the project summary. The criteria are defined as follows:

- **Minor Deduction (-1)**: A point is deducted if there is a logical leap between the output and the stated outcome.
- **Minor Deduction (-1)**: A point is deducted if there are gaps or inconsistencies across short-, medium-, and long-term outcomes.
- **Major Deduction (-3)**: A significant deduction is applied if the outcome appears to be forced or artificially constructed, lacking substantive meaning or relevance.
- **Major Deduction (-3)**: A significant deduction is also applied if the output itself does not logically address or relate to the problem stated in the project summary.
- **Note**: Objectivity, specificity, or clarity of expression are not part of the evaluation. Even if an outcome is described in clear or concrete terms, this alone should not result in a high score.

5.2.1 Rank Invariance Assumption

To enhance robustness, one might consider ensemble evaluations using multiple LLMs. However, due to API budget constraints, we employed a single model (`GPT-4o`) throughout the study. To address potential concerns about model-specific biases, we invoke the assumption of *rank invariance*.

Let θ denote the internal parameters of the LLM and $g_\theta(x)$ the scalar score assigned to input x. For any two inputs x_i and x_j within the same category, we assume that the relative ranking is preserved across plausible LLMs:

$$g_\theta(x_i) > g_\theta(x_j) \iff g_{\theta'}(x_i) > g_{\theta'}(x_j), \quad \forall \theta, \theta' \in \Theta,$$

where Θ represents the set of instruction-following, high-capacity LLMs. This assumption implies that the influence of θ acts as a fixed effect, which cancels out in relative comparisons.

This assumption is supported by recent empirical evidence from Thakur et al. [22], who show that even judge models with relatively low score-level alignment can nonetheless maintain high consistency in rank-ordering. For instance, models such as `Mistral 7B` and the lexical `Contains` metric achieved Spearman's ρ values of 0.98 and 0.99, respectively—on par with `GPT-4 Turbo`—despite lower agreement with human-assigned scores. These findings suggest that rank invariance is a plausible and practically sufficient assumption in comparative evaluation settings.

While our design does not employ full pairwise or tournament-style comparisons, it remains inherently comparative: each project is evaluated relative to a benchmark

example within its Main Expenses Category. This localized evaluation structure makes the rank invariance assumption both reasonable and methodologically appropriate in our setting.

5.3 Identifying Backdoor-Relevant Biases in LLM Scores

Large language models (LLMs) are known to exhibit various evaluation biases[4]. However, not all such biases need to be adjusted for in causal analysis. Following the *back-door criterion* [17], we adjust only for variables that influence both the variable of interest d and y. This ensures an unbiased estimate of the parameter β_d.

Among the potential sources of bias, we focus on the well-documented *redundancy bias*—the tendency of LLMs to assign higher scores to longer or more verbose inputs [4, 16, 19, 28, 29].

Figure 3 shows the distribution of input text length ℓ (covering project descriptions, outcomes, and other relevant text) across project types. It reveals that *Plan* projects tend to be more verbose: the median length is 438.5 for *Plan* projects and 335.0 for other projects.

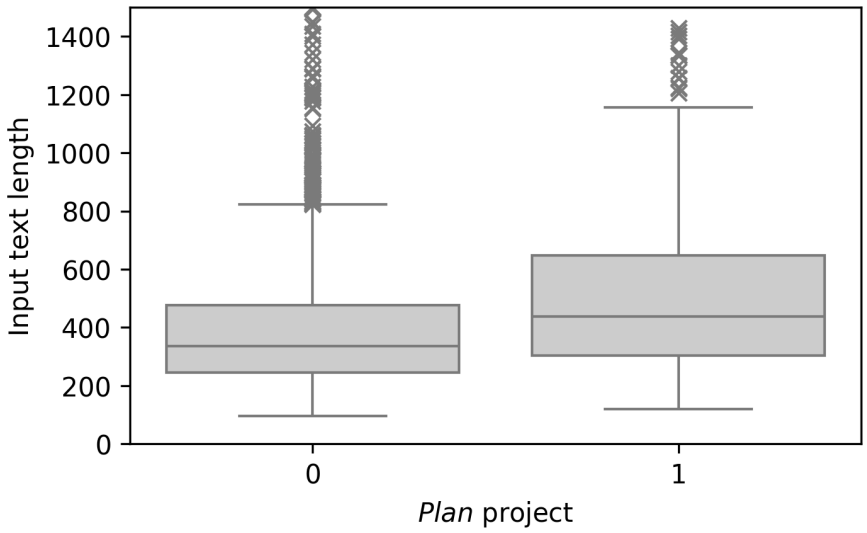

Fig. 3 Distribution of input text length ℓ for project descriptions, outcomes, and other relevant text data fed into the LLM. *Plan* projects tend to have longer text compared to others, with the median length being 438.5 and 335.0, respectively.

Given that verbosity ℓ is associated with both d and y, failing to adjust for ℓ would violate the back-door criterion and bias the estimate of β_d. Figure 4 presents the assumed causal structure, in which ℓ acts as a confounder. Accordingly, we include ℓ as a covariate in the regression model (Equation 2).

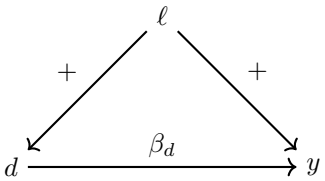

Fig. 4 Directed acyclic graph (DAG) representing the assumed causal structure. The variable of interest d (*Plan* project flag) directly affects the outcome y (LLM-based score) with coefficient β_d. The covariate ℓ (logic-model text length) influences both d and y, and thus must be adjusted for under the back-door criterion[17].

6 Fitting and Results

6.1 1st-stage Results: LLM Score Distributions

Table 4 shows the evaluation results of the LLM. Scores range from 1 (lowest) to 5 (highest), with 5 representing the most favorable evaluation. The average score for *Plan* projects is 2.89, while for other projects it is 2.77, suggesting that *Plan* projects receive marginally higher scores on average. However, as pointed out in Section 5.3, this evaluation includes biases inherent to the LLM, such as those caused by long text inputs. Therefore, this aggregation alone does not allow for a conclusive identification of the quality of *Plan* projects.

Table 4 Score Distributions and *Plan* Project Status

Score	***Plan* Project = 0**		***Plan* Project = 1**	
1	257	(11.43%)	39	(9.49%)
2	664	(29.52%)	114	(27.74%)
3	776	(34.50%)	137	(33.33%)
4	442	(19.65%)	95	(23.11%)
5	110	(4.89%)	26	(6.33%)

6.2 2nd-stage Regression Modeling

Having constructed the data set $\mathcal{D} = \{y_i, d_i, \ell_i\}_{i=1}^{N}$, we fit the 2nd-stage regression model defined in Equation 2. As noted in Section 5.2, since the LLM score y_i is assigned on a *within–Main Expenses Category* basis, we estimate the model in a hierarchical (multilevel) form that allows category-specific effects. Two specifications are explored: *random-intercept* model and *random-intercept-and-slope* model.

6.2.1 Random-Intercept Model

The *Random-Intercept* model can be specified as:

$$y_i = \beta_{0j} + \beta_d d_i + \beta_\ell \ell_i + \epsilon_i, \tag{4}$$

Here, β_{0j} represents the random intercept for the j-th category, determined by the expense category main_expenses$_j$. The random intercept β_{0j} is modeled as:

$$\beta_{0j} = \mu_{\beta_0} + u_{0j}, \quad u_{0j} \sim \mathcal{N}(0, \sigma^2_{\beta_0})$$

In this case, μ_{β_0} represents the overall mean intercept for all categories, while u_{0j} is the random deviation for the j-th category, assumed to follow a normal distribution with mean 0 and variance $\sigma^2_{\beta_0}$.

This model assumes that only the intercept varies across categories, while the coefficients of the *Plan* project indicator d_i and text length ℓ_i remain fixed across all categories.

6.2.2 *Random-Intercept-and-Slope* Model

The *Random-Intercept-and-Slope* model extends the previous model by allowing both the intercept β_0 and the coefficient for the *Plan* project indicator d_i to vary across categories. The model is specified as:

$$y_i = \beta_{0j} + \beta_{dj} d_i + \beta_\ell \ell_i + \epsilon_i, \tag{5}$$

Here, β_{0j} is the random intercept for the j-th category, β_{dj} is the random coefficient for the *Plan* project indicator d_i, which allows the effect of *Plan* projects to vary across categories, and β_ℓ is the fixed effect associated with the text length ℓ_i.

The random coefficient β_{dj} are modeled as:

$$\beta_{0j} = \mu_{\beta_0} + u_{0j}, \quad u_{0j} \sim \mathcal{N}(0, \sigma^2_{\beta_0})$$
$$\beta_{dj} = \mu_{\beta_d} + u_{dj}, \quad u_{dj} \sim \mathcal{N}(0, \sigma^2_{\beta_d})$$

In this model, μ_{β_0} and μ_{β_d} represent the overall means for the random intercept and random coefficient, respectively. The terms u_{0j} and u_{dj} represent the random deviations for the j-th category in the intercept and coefficient, respectively, assumed to follow normal distributions with mean 0 and variances $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_d}$.

This model allows both the baseline evaluation score and the effect of the *Plan* project indicator d_i to vary across categories, offering a more flexible approach compared to the *Random-Intercept* Model (Equation 4).

We impose weakly informative priors:

$$\mu_{\beta_0},\ \mu_{\beta_d} \sim \mathcal{N}(0, 5^2), \qquad \sigma_{\beta_0},\ \sigma_{\beta_d} \sim \text{HalfNormal}(5).$$

6.3 2nd-stage Regression Results

6.3.1 Model Comparison

We trained and compared the models specified in Model 4 (*Random-Intercept model*) and Model 5 (*Random-Intercept-and-Slope model*), in addition to a Naive model based on Model 4 that does not include any covariate. The results of the learned models are shown in Table 5.

First, the Naive model without a covariate has a positive estimate for the parameter of interest, consistent with the aggregation results in section 6.1. In Model 4, by including the covariate, the bias is adjusted, and the parameter of interest becomes negative. This suggests that the quality of the *logic models* for *Plan* projects is comparatively low.

Table 5 Bayesian Regression Model Comparison

Model	**Naive Model**	**Model 4**	**Model 5**
Parameter of Interest $\hat{\beta}_d$	0.040	-0.114	*Category dependent*
	(0.054)	(0.055)	
Parameters Included			
Covariate (ℓ)		✓	✓
Random Intercept (β_{0j})	✓	✓	✓
Random Slope (β_{dj})			✓
elpd_loo	-3831.43	-3711.43	-3696.69

Furthermore, the results from Leave-One-Out Cross-Validation (elpd_loo in the table) indicate that Model 5 best explains the observed phenomenon [25]. This finding implies that the quality of *logic models* for *Plan* projects exhibits heterogeneity depending on the main expenses category. In the following, we present the estimated values for the parameter of interest based on Model 5. Note that the detailed estimation for Model 5 can be found in Appendix B.

6.3.2 Estimates of the Parameter of Interest

The posterior distribution of the parameter of interest estimated by Model 5, along with the 94% HDI, is reported in Figure 5. Generally, the values are negative, but for certain main expense categories, they are positive, indicating heterogeneity across categories.

There can be multiple factors contributing to heterogeneity. First, even for *Plan* projects, there may be cases where the project itself has a spillover effect, resulting in a valid *logic model*. For example, the project "*Regional Co-Creation and Cross-Sector Carbon Neutral Technology Development and Demonstration Project*"(Budgetary Project ID: `5019`) under Main Expenses: Energy Policy Expenditure, is classified as

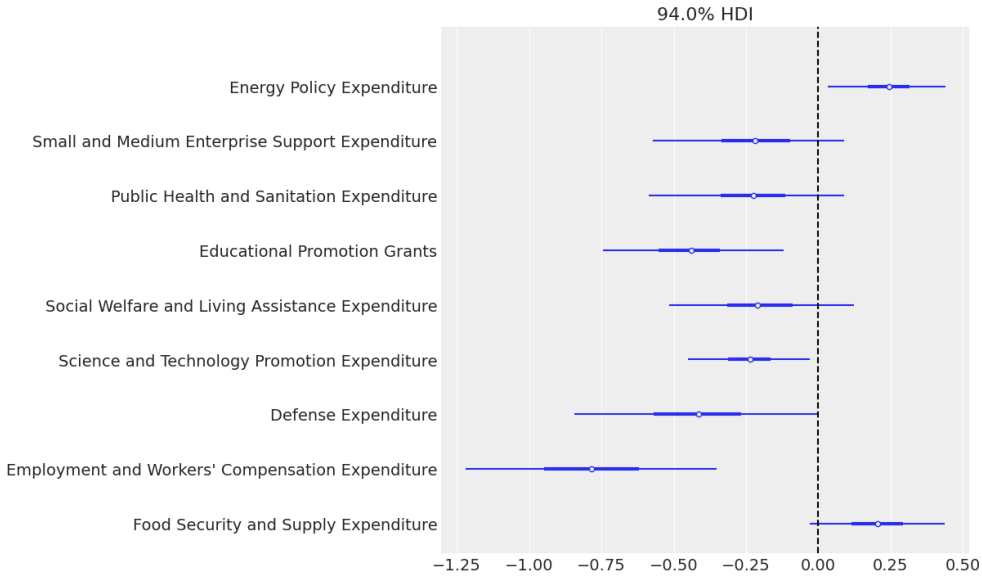

Fig. 5 Posterior distribution of the parameter of interest estimated by Model 5 with the 94% Highest Density Interval (HDI). The plot illustrates the variation in the parameter across different main expense categories. The HDI represents the region within which the true parameter value lies with 94% probability, providing an indication of the uncertainty in the estimate.

a *Plan* project according to our definition. However, the LLM evaluates it by stating that "the development of CO2 reduction technologies is expected to have direct effects," assigning the highest score of 5. As a demonstration project and research and development initiative, it leads to an outcome that contributes to future CO2 reduction, forming a valid *logic model*.

Another factor to consider is that our assignment of "*Plan* projects" is based on a simplified method. As defined in section 4, we classified the projects mechanically. For instance, the project "*Ministry of Economy, Trade, and Industry Green Innovation Fund Project*" (Budgetary Project ID: 7204) under Main Expenses: Energy Policy Expenditure, which received a score of 5 from the LLM, is a mix of multiple initiatives. While it contains the terms "Research / PoC," it also includes substantial policies, which results in a high-quality *logic model*. For such projects, while prominent cases could be organized manually, we have opted to report the study using this simplified classification due to the high risk of cherry-picking in research.

7 Case Study: Are *Logic Models* Appropriate for *Plan* projects?

In the previous sections, it has been demonstrated that, despite the presence of heterogeneity, the overall quality of the *logic models* for *Plan* projects is low. Before drawing conclusions and providing recommendations, it is important to examine a typical case.

7.1 Case: Smart Safety Technology PoC Project

We examine the case of the "Smart Safety Technology PoC Project" (Budgetary Project ID: 006080), administered by the Ministry of Economy, Trade and Industry (METI) under the Small and Medium Enterprise Support Expenditure.[6] This project responds to a critical policy concern: in the fields of industrial safety, including high-pressure gas, electricity, city gas, and LPG, a significant wave of retirements among experienced workers is expected, exacerbated by difficulties in recruiting younger personnel. To address the potential degradation of Japan's industrial safety standards, the project investigates the feasibility of introducing new technologies such as IoT, AI, and drones. It is fundamentally a PoC initiative aimed at testing the applicability of emerging technologies to industrial safety.

The *logic model* of this project, as registered in the RS System, is presented in Table 6.

Table 6 *Logic model* structure of the "Smart Safety Technology PoC Project." Budgetary Project ID: 006080. RS System. Available at: https://rssystem.go.jp/project/df9ca39e-bed4-4c4b-80f6-6bfc5ce860df?activeKey=basic-information (Accessed: April 26, 2025).

Goal Type	Goal Description	Indicator
Output	Conduct a sufficient number of technology PoC initiatives to clarify the merits of introducing new technologies such as IoT, AI, and drones into the industrial safety field.	Number of technology PoC initiatives conducted
Outcome (Short-term)	Expansion of best practices for smart safety implementation	Number of best practices for smart safety implementation
Outcome (Medium-term)	Expansion of best practices for smart safety implementation	Number of best practices for smart safety implementation
Outcome (Long-term)	Maintenance and improvement of industrial safety levels through smart safety initiatives; improvement of operational efficiency	Increase in certifications under the accredited business operator system through smart safety

[6] Budgetary Project ID: 006080. RS System. Available at: https://rssystem.go.jp/project/df9ca39e-bed4-4c4b-80f6-6bfc5ce860df?activeKey=basic-information (Accessed: April 26, 2025).

7.2 Our Critical Evaluation of the Case

Under the LLM-as-a-Judge framework, this project received a score of 2. The evaluation indicates a significant disconnect between the output and the described outcomes. While quantitative indicators such as the number of best practices and certifications are presented, these metrics do not fully capture the project's fundamental purpose.

Rather, the essential aim—as stated in the output description—is to assess the conceptual feasibility of "smart safety" technologies and to investigate the bottlenecks hindering their broader industrial adoption. However, the *logic model*, as registered, seems to confine the project within a conventional output-outcome framework, emphasizing easily quantifiable metrics, such as the number of best practices. These metrics are not appropriate measures of success for a PoC initiative.

For PoC projects, success should be defined not by the number of tangible deliverables, but by the degree to which the project contributes to clarifying feasibility conditions, identifying barriers to adoption, and informing the design and refinement of subsequent policy interventions.

Thus, the evaluation framework, which treats PoC activities as if they were fully operational programs, risks incentivizing superficial compliance with quantitative targets at the cost of meaningful evidence production.

8 Conclusion and Future Work

8.1 Empirical Findings and Recommendations

We argue that applying the *logic model* framework to *Plan* projects is inherently problematic. As demonstrated in the preceding sections, the current implementation tends to result in relatively low-quality *logic models* for these projects. As defined in Section 5.2, the notion of "low quality" in this context refers **to the logical consistency and coherence of outcome setting—not to its objectivity or measurability**. This study highlights the operational issue whereby administrators are effectively compelled to construct logically tenuous *logic models* for *Plan* projects.

Moreover, it is important to note that *Plan* projects are not marginal in number. Table 7 provides an approximate estimate of the proportion of *Plan* projects across ministries for fiscal year 2024. Given this prevalence, we are concerned that the current situation may contribute to what Ono [15] describes as the formalization or ritualization of evaluation practices. We argue that, as soon as practicable, *Plan* projects should be excluded from the current evaluation scheme.

At present, the official operational manual does not explicitly exempt *Plan* projects from the *logic model* framework. While it does allow for the omission of objectively measurable outcomes, it still requires *Plan* projects to conform to the same basic *logic model* structure as *Do* projects [7].

While we critically assess the direct application of traditional *logic models* to *Plan* projects, we recognize that some form of management and evaluation framework remains necessary, given that these projects are still components of publicly funded budgets. Designing a tailored management framework suitable for the characteristics

of *Plan* projects—emphasizing preliminary investigations, feasibility assessment, and evidence-gap exploration—represents an important area for future work.

We propose the following revisions to project management and evaluation frameworks:

Recommendations

- Introduce a dedicated management category within the RS system specifically for *Plan* projects.
- Develop a separate management and evaluation format tailored to the unique objectives of *Plan* projects, emphasizing hypothesis testing, evidence gap analysis, and feasibility assessment rather than linear output-outcome chains.

8.2 Further Contributions

This study introduced the **DLJ-2** framework, a two-stage framework designed to address the challenges of evaluating *Plan* projects. In the present case, we focus on controlling for input verbosity—identified in Section 5.3 as a source of LLM bias that is both correlated with the outcome score and with the variable of interest d (Plan-project indicator). Importantly, our framework is extendable: as long as a measurable source of LLM bias satisfies the *back-door criterion* [17], the DLJ-2 can be smoothly adapted to other contexts.[7]

Beyond this technical contribution, our study also demonstrates the potential of LLMs to support scalable evaluation of government programs, particularly those involving unstructured materials such as text or slides. With the advent of centralized systems like Japan's *RS System*, we anticipate broader applications of LLMs for continuous monitoring—enabling early detection of evaluation formalism and more adaptive, data-driven governance.

[7] In cases where the set of confounding covariates becomes high-dimensional, the adjustment stage may benefit from the use of modern machine learning techniques such as Double/Debiased Machine Learning (DML) [3], which enable the use of flexible, potentially nonparametric models for nuisance components while ensuring valid inference for the parameter of interest through orthogonalization and sample-splitting.

Table 7 Preliminary and Approximate Aggregation of Plan Project Ratios by Ministry (FY2024)

Ministry	**Total Projects**	**Proportion of *Plan* Projects (%)**
Ministry of the Environment	217	29%
Ministry of Economy, Trade and Industry	514	23%
Ministry of Internal Affairs and Communications	247	23%
Ministry of Agriculture, Forestry and Fisheries	414	20%
Cabinet Office	235	19%
Ministry of Land, Infrastructure, Transport and Tourism	692	19%
Nuclear Regulation Authority	54	19%
Children and Families Agency	99	18%
Ministry of Education, Culture, Sports, Science and Technology	534	15%
Digital Agency	235	14%
National Police Agency	62	11%
Ministry of Health, Labour and Welfare	1139	10%
Ministry of Finance	61	10%
Ministry of Justice	50	8%
Ministry of Defense	320	8%
Reconstruction Agency	160	6%
Ministry of Foreign Affairs	437	5%
Others	134	30%

Note: Others includes ministries with fewer than 50 projects.

Statements and Declarations

Conflict of Interest:

The authors have no conflicts of interest to this article.

Data Availability:

Japan's Cabinet Secretariat [8]. RS System: Administrative Project Review Dataset, FY2024. Available at: https://rssystem.go.jp/download-csv/2024 (Accessed: April 26, 2025).

Code Availability:

All code used in the analysis is available in the following GitHub repository: https://github.com/MasaAsami/rs-system-llm-eval

Appendix A Prompt Texts and Parameter Settings

This appendix presents the complete prompt texts provided to the LLM, as well as the relevant parameter settings used for evaluation.

A.1 System Prompt

The following system prompt was used: [8]

```
# Role
You are assigned the role of a highly competent Japanese government official
responsible for evaluating budgetary projects proposed by various ministries.
You must perform the evaluations with strict rigor.
Carefully review the provided data and evaluate each project based on
the specified perspectives and criteria.

# Task
You are instructed to evaluate each project on
a five-point scale according to the specified criteria:
5 | Top 5 projects (excellent)
4 | Relatively good
3 | Average
2 | Relatively poor
1 | Bottom 5 projects (poor)

# Evaluation Criteria: Consistency Between Outputs and Outcomes
The evaluation focuses on the logical consistency between the stated outputs
and outcomes of each project, as interpreted from the project summary.
The criteria are defined as follows:

- Minor Deduction (-1):
A point is deducted if there is a logical leap between the output and
the stated outcome.
- Minor Deduction (-1):
A point is deducted if there are gaps or inconsistencies across short-,
medium-, and long-term outcomes.
- Major Deduction (-3):
A significant deduction is applied if the outcome appears to be forced or
artificially constructed, lacking substantive meaning or relevance.
- Major Deduction (-3):
A significant deduction is also applied if the output itself does not
logically address or relate to the problem stated in the project summary.
- Note:
Objectivity, specificity, or clarity of expression are not part of the evaluation.
Even if an outcome is described in clear or concrete terms,
this alone should not result in a high score.

# Procedure
```

[8] Note: The actual prompts used in the evaluation were written in Japanese. The English version presented below is a faithful translation for documentation purposes.

```
## Step 1: Establish Benchmarks
- First, establish a benchmark by identifying the overall average level of
project quality across the dataset.
- Additionally, select a few representative examples from the top 5 (score = 5)
and bottom 5 (score = 1) projects to help anchor your scoring decisions.

## Step 2: Evaluation
- Assign a 5-point score to each project by comparing it to the benchmark level
identified in Step 1.
- Always evaluate relatively, based on the benchmark examples.

# Important Notes
- Evaluate all given budgetary projects.
- Be sure to assign at least 5 projects a score of 5, and 5 projects a score of 1.
- Grade strictly and critically.
- First consider the rationale for the score, and then assign the score accordingly.
- Avoid using score 3 (average) unless absolutely necessary.

# Output Format
Output must be in dictionary format only.
(Do not include any additional explanation or formatting.)

Each record should contain:
- `id`: Corresponding project ID
- `point`: Assigned score
- `reason`: Rationale for the assigned score

# Example Output Format
[
    {"id": 3503, "point": 4, "reason": "Because ..."},
    {"id": 3533, "point": 1, "reason": "Because ..."}
]
```

A.2 User Prompt

Figures A1 presents the user prompt used for scoring. {`data_text`} corresponds to the content of the Administrative Project Review Sheets.

> Evaluate the following.
>
> Data: {`data_text`}

Fig. A1 The user prompt used for scoring *logic models*. {`data_text`} corresponds to the content of the Administrative Project Review Sheets.

A.3 Parameter Settings

The LLM was invoked with the following parameters:

- **Model:** `GPT-4o`
- **Temperature:** 0

A temperature setting of zero was used to minimize randomness and promote more consistent outputs across repeated queries.

Appendix B Bayesian Inference and MCMC Convergence Analysis

The estimation was performed using PyMC[9], a probabilistic programming library in Python.

To assess the convergence of the Markov Chain Monte Carlo (MCMC) chains, we examined the trace plots and summary statistics of the model parameters. The trace plots, shown in Figure B2, indicate good mixing, with no evident signs of autocorrelation or divergence in the chains. Additionally, the posterior distributions for all parameters, including the coefficients associated with the `coef_plan_project` categories and the intercept terms, display reasonable convergence to stable distributions.

Furthermore, the summary statistics presented in Table B1 confirm that the potential scale reduction factor $\hat{R}$ for all parameters is equal to 1, indicating that the chains have converged to the target posterior distribution. The effective sample sizes (ESS) for both bulk and tail of the posterior distributions are sufficiently large, suggesting that the MCMC has adequately explored the parameter space.

Taken together, these diagnostics provide strong evidence that the MCMC chains have converged and that the posterior estimates are reliable.

[9] PyMC, https://www.pymc.io/welcome.html (Accessed: May 1, 2025)

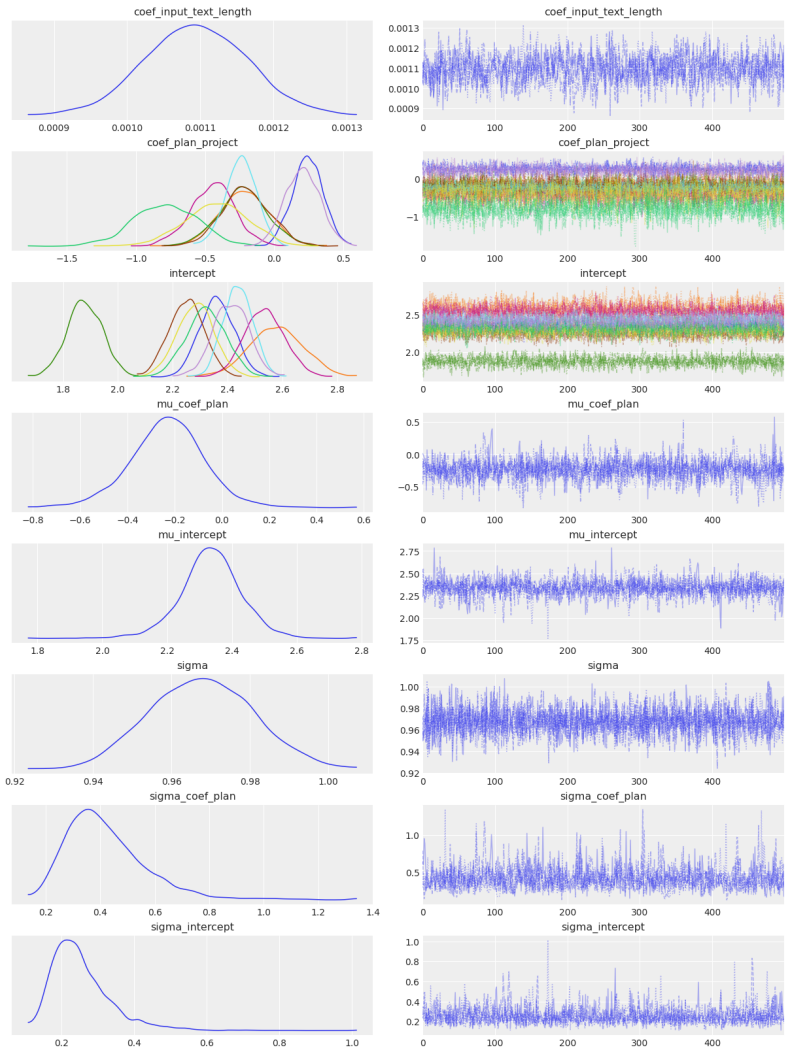

Fig. B2 Trace plots for the MCMC chains of the model parameters. The plots show good mixing of the chains, with no visible signs of autocorrelation or divergence, indicating that the Markov chains have adequately explored the parameter space and have converged to the target posterior distribution.

Table B1 Summary Statistics for Bayesian Model Estimation

Parameter	Mean	SD	ESS (Bulk)	ESS (Tail)	R-hat
coef_input_text_length	0.001	0.000	1045	1388	1.000
coef_plan_project[0]	0.243	0.109	2835	1654	1.000
coef_plan_project[1]	-0.217	0.179	2478	1120	1.000
coef_plan_project[2]	-0.225	0.177	3029	1651	1.000
coef_plan_project[3]	-0.445	0.162	2606	1436	1.000
coef_plan_project[4]	-0.204	0.174	3610	1301	1.000
coef_plan_project[5]	-0.240	0.112	4157	1797	1.000
coef_plan_project[6]	-0.421	0.233	2776	1579	1.000
coef_plan_project[7]	-0.788	0.233	2246	1420	1.000
coef_plan_project[8]	0.205	0.125	2713	1585	1.000
intercept[0]	2.361	0.063	1993	1789	1.000
intercept[1]	2.577	0.099	1870	1496	1.000
intercept[2]	1.878	0.064	2032	1649	1.000
intercept[3]	2.535	0.073	1755	1739	1.000
intercept[4]	2.250	0.063	2149	1421	1.000
intercept[5]	2.438	0.053	1784	1675	1.000
intercept[6]	2.283	0.067	2271	1507	1.000
intercept[7]	2.320	0.073	2177	1383	1.000
intercept[8]	2.412	0.065	1823	1442	1.000
mu_coef_plan	-0.232	0.160	2075	1284	1.000
mu_intercept	2.335	0.096	1907	1162	1.000
sigma	0.968	0.013	4465	1495	1.000
sigma_coef_plan	0.421	0.153	1166	1083	1.000
sigma_intercept	0.255	0.087	1895	1389	1.000

Note: The coefficients associated with the 'coef_plan_project' parameters correspond to the following expenditure categories: [0] Energy Policy Expenditure, [1] Small and Medium Enterprise Support Expenditure, [2] Public Health and Sanitation Expenditure, [3] Educational Promotion Grants, [4] Social Welfare and Living Assistance Expenditure, [5] Science and Technology Promotion Expenditure, [6] Defense Expenditure, [7] Employment and Workers' Compensation Expenditure, [8] Food Security and Supply Expenditure.

References

[1] Asatryan Z, Birkholz C, Heinemann F (2025) Evidence-based policy or beauty contest? An LLM-based meta-analysis of EU cohesion policy evaluations. International Tax and Public Finance 32:625–655. https://doi.org/10.1007/s10797-024-09875-4, URL https://doi.org/10.1007/s10797-024-09875-4

[2] Cheong I, Xia K, Feng KJK, et al (2024) (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '24, p 2454–2469, https://doi.org/10.1145/3630106.3659048, URL https://doi.org/10.1145/3630106.3659048

[3] Chernozhukov V, Chetverikov D, Demirer M, et al (2018) Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21(1):C1–C68. https://doi.org/10.1111/ectj.12097, URL https://doi.org/10.1111/ectj.12097, https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf

[4] Gu J, Jiang X, Shi Z, et al (2025) A Survey on LLM-as-a-Judge. URL https://arxiv.org/abs/2411.15594, arXiv:2411.15594

[5] Guha N, Nyarko J, Ho DE, et al (2023) LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NeurIPS '23

[6] Japan's Cabinet Secretariat (2023) EBPM Guidebook ver 1.2. URL https://www.gyoukaku.go.jp/ebpm/img/guidebook1.2_230403.pdf

[7] Japan's Cabinet Secretariat (2024) Gyosei Jigyo Review Sheet Sakusei Guidebook Ver.1.0. URL https://www.gyoukaku.go.jp/review/img/R06sakusei-guidebook.pdf

[8] Japan's Cabinet Secretariat (2024) RS System: Administrative Project Review Dataset, FY2024. URL https://rssystem.go.jp/download-csv/2024, Accessed: April 26, 2025

[9] Lilley J, Townley S (2024) Tackling transparency in UK politics: application of large language models to clustering and classification of UK parliamentary divisions. Journal of Computational Social Science 7:2563–2589. https://doi.org/10.1007/s42001-024-00317-z, URL https://doi.org/10.1007/s42001-024-00317-z

[10] Liu Y, Zhou H, Guo Z, et al (2024) Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. In: First Conference on Language Modeling, URL https://openreview.net/forum?id=9gdZI7c6yr

[11] Liusie A, Manakul P, Gales M (2024) LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. In: Graham Y, Purver M (eds) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, St. Julian's, Malta, pp 139–151, URL https://aclanthology.org/2024.eacl-long.8/

[12] Ma S, Chen C, Chu Q, et al (2024) Leveraging Large Language Models for Relevance Judgments in Legal Case Retrieval. URL https://arxiv.org/abs/2403.18405, arXiv:2403.18405

[13] Madaan A, Tandon N, Gupta P, et al (2023) SELF-REFINE: iterative refinement with self-feedback. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NeurIPS '23

[14] Meyer JW, Brian R (1977) Institutionalized Organizations: Formal Structure as Myth and Ceremony. American Journal of Sociology 83(2):340–363. https://doi.org/10.1086/226550, URL https://cir.nii.ac.jp/crid/1363388845911085952

[15] Ono T (2011) Validity of target setting and achievement assessment in performance measurement. Regional studies (Tottori University journal of the Faculty of Regional Sciences) 8(2):1–20. URL https://cir.nii.ac.jp/crid/1050578304505257088

[16] Park J, Jwa S, Meiying R, et al (2024) OffsetBias: Leveraging debiased data for tuning evaluators. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp 1043–1067, https://doi.org/10.18653/v1/2024.findings-emnlp.57, URL https://aclanthology.org/2024.findings-emnlp.57/

[17] Pearl J (1993) [Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention. Statistical Science 8(3):266–269. URL http://www.jstor.org/stable/2245965

[18] Qin Z, Jagerman R, Hui K, et al (2024) Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. In: Duh K, Gomez H, Bethard S (eds) Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics, Mexico City, Mexico, pp 1504–1518, https://doi.org/10.18653/v1/2024.findings-naacl.97, URL https://aclanthology.org/2024.findings-naacl.97/

[19] Saito K, Wachi A, Wataoka K, et al (2023) Verbosity Bias in Preference Labeling by Large Language Models. URL https://arxiv.org/abs/2310.10076, arXiv:2310.10076

[20] Shi L, Ma C, Liang W, et al (2025) Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. URL https://arxiv.org/abs/2406.07791, arXiv:2406.07791

[21] Sugitani K (2023) The Reality of Administrative Project Review and its Challenges at a Turning Point: EBPM and "Agile Policy Formulation and Evaluation". Nihon Hyoka Kenkyu (Japanese Journal of Evaluation Studies) 23(2):17–30. https://doi.org/10.11278/jjoes.23.2_17, URL https://doi.org/10.11278/jjoes.23.2_17

[22] Thakur AS, Choudhary K, Ramayapally VS, et al (2025) Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. URL https://arxiv.org/abs/2406.12624, arXiv:2406.12624

[23] Tokuda T (2022) Debates on the Policy Evaluation System: 20 Years After Its Introduction. Rippou to Chousa 443:189–207. URL https://www.sangiin.go.jp/japanese/annai/chousa/rippou_chousa/backnumber/20220218.html

[24] Tokuda T (2023) Evaluation Systems within the Japanese Government: Project Reviews, Policy Evaluation, Administrative Evaluation and Oversight, Budget Execution Surveys, and Audit Inspections. Rippou to Chousa 459:218–227. URL https://www.sangiin.go.jp/japanese/annai/chousa/rippou_chousa/backnumber/20230802.html

[25] Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27(5):1413–1432. https://doi.org/10.1007/s11222-016-9696-4, URL https://doi.org/10.1007/s11222-016-9696-4

[26] Wang P, Li L, Chen L, et al (2024) Large Language Models are not Fair Evaluators. In: Ku LW, Martins A, Srikumar V (eds) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, pp 9440–9450, https://doi.org/10.18653/v1/2024.acl-long.511, URL https://aclanthology.org/2024.acl-long.511/

[27] W.K. Kellogg Foundation (2004) Logic Model Development Guide. URL https://www.naccho.org/uploads/downloadable-resources/Programs/Public-Health-Infrastructure/KelloggLogicModelGuide_161122_162808.pdf

[28] Ye J, Wang Y, Huang Y, et al (2025) Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In: The Thirteenth International Conference on Learning Representations, URL https://openreview.net/forum?id=3GTtZFiajM

[29] Zheng L, Chiang WL, Sheng Y, et al (2023) Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook,