EPiC Series in Built Environment
Volume 7, 2026, Pages 445–454
Proceedings of Associated Schools of Construction 62nd Annual International Conference

Investigating the Effects of Prompt Design on LLM-Generated Priors for Bayesian Calibration in Building Energy Simulation

Yangxi Bai[1], Endong Wang[1]
[1]State University of New York College of Environmental Science and Forestry (SUNY ESF)

Bayesian calibration has been increasingly used for improving the accuracy of building energy models, but its performance strongly depends on the selection of prior distributions. Large language models (LLMs) provide a new potential way to generate prior distributions by extracting domain knowledge. However, prompt design during LLM execution is critical to the reliability of priors generated by LLMs. The quality and structure of prompts determine how LLMs interpret domain knowledge and translate it into priors, which in turn could influence both prior characteristics and calibration performance. This study investigates how different prompt designs affect the statistical characteristics of LLM-generated priors in building energy simulation calibration. Using three years of monthly electricity data from a campus building, three levels of prompt information (low, medium, and high) were applied to generate priors for four typical parameters. In this study, results show that prompts with lower information levels produced more balanced priors, leading to faster convergence and higher calibration accuracy. These findings suggest that prompt design should balance informativeness and generality to achieve effective LLM-assisted Bayesian calibration, providing a new perspective for integrating LLMs into building energy simulation and modeling.

Keywords: Prompt Design, Large Language Models, Prior Generation, Bayesian Calibration, Building Energy Simulation

Introduction

Bayesian calibration is becoming an increasingly popular method for improving the reliability of building energy simulation and modeling processes (Hou et al., 2021; Kennedy & O'Hagan, 2001). Bayesian approach enables uncertainty quantification and enhances predictive performance, which is important for decision-making in building design, operation, and retrofit analysis (Attia et al., 2020; Hou et al., 2021; Wang 2025). However, the effectiveness of Bayesian calibration depends highly on the definition of prior distributions, which specify the plausible parameter ranges before calibration. Inappropriate priors can slow convergence or bias posterior inference, leading to reduced predictive accuracy. As a result, the performance of Bayesian calibration strongly depends on the definition of prior distributions (Berger, 1990; Ruggeri et al., 2021). Traditionally, most existing Bayesian studies define priors based on expert knowledge, literature reviews, or empirical rules (Calama-González et al., 2021; Chong et al., 2017; Wang 2025). Although these methods reflect engineering experience, they are subjective and difficult to reproduce. The resulting priors often vary among analysts, leading to inconsistent calibration outcomes and limited transparency. Furthermore, new practitioners and researchers may find it difficult to construct reliable priors, which limits the application of Bayesian calibration in building simulation and modelling processes.

Large language models (LLMs) provide a potential solution to address this challenge by generating priors in a more objective and repeatable way. Trained on extensive text-based knowledge, LLMs can extract domain-relevant information and translate it into quantitative assumptions for engineering parameters (Kasneci et al., 2023). This approach reduces dependence on manual expert input and makes Bayesian calibration more accessible to practitioners. Recent advances in LLMs for building energy simulation have demonstrated their capability to extract, summarize, and structure engineering knowledge from diverse sources (Hong & Zhang, 2025). For example, Jiang et al. (2024) developed EPlus-LLM, which converts natural language building descriptions into EnergyPlus input files with high accuracy and minimal human intervention. Zhang et al. (2024) further showed that LLMs can interpret EnergyPlus simulation outputs and answer human queries through natural language interaction. Some other studies have explored LLM applications in building research, such as data interpretation, knowledge retrieval, and simulation assistance (Coakley et al., 2014). However, their potential to support Bayesian calibration has rarely been explored, especially in constructing priors through LLMs. Additionally, recent studies have highlighted that the performance of large language models is sensitive to prompt formulation. Yongchao Zhou (2022) demonstrated that prompt quality notably affects task outcomes and proposed an automatic prompt engineering framework capable of generating human-level instructions for LLMs. Hatakeyama-Sato et al. (2023) further evaluated GPT-4 in scientific applications and confirmed that the structure and specificity of prompts directly influence the accuracy and interpretability of model outputs. More recently, Krishna Padmanabhan (2025) applied fine-tuned LLMs to automate Bayesian modeling in adaptive trials, illustrating that carefully designed prompts can reduce the technical barrier to Bayesian analysis. These studies suggest that prompt engineering influences how LLMs interpret contextual information and generate outputs. However, these studies have focused on generic modeling tasks rather than building energy simulation. How the amount of information provided in a prompt influences the priors generated by an LLM and the resulting calibration performance, particularly in data-limited contexts, remains unexplored. This study provides a case-based methodological test that examines how different prompt information levels influence LLM-generated priors within a Bayesian calibration process for building energy simulation and modelling. Three prompt information levels are compared using the same building model and monthly electricity data. Within this specific context, the findings provide methodological insight into the feasibility and behavior of using LLMs as prior-generation tools, and context-dependent guidance for data-limited building energy calibration.

Methodology

Model Description

This study is based on a single mixed-use building using measured monthly electricity data collected over a 48-month period from January 2018 to December 2021. The building is a three-story facility located in Syracuse, NY (climate zone 5A), with a total floor area of approximately 54,000 ft² and varied operational patterns. Such buildings are characterized by dynamic internal loads and operational variability, which limit the information content of observed data and increase sensitivity to prior specification, making them suitable for methodological testing of prior-generation strategies. This case also reflects a common scenario in practice, where only aggregated monthly electricity data and basic design documentation are available. The electricity usage was sourced from the Building Automation System (BAS) metering records, pre-processed by the facility management team to remove outliers and fill missing values prior to analysis. Weather data were obtained from National Weather Service Climate

Data for Hancock Airport. The 48-month dataset was divided into a 36-month calibration period and a subsequent 12-month evaluation period to ensure sufficient seasonal coverage for stable Bayesian inference at the monthly scale. Parameter selection was guided by previous calibration studies based on monthly utility data (Attia et al., 2020; Chong et al., 2019; Zhao et al., 2021). From the parameter sets identified in these studies, four parameters directly influencing electrical demand were selected: Equipment Power Density (EPD), Lighting Power Density (LPD), Cooling Setpoint (CSP), and Air Handling Unit Fan Total Efficiency (FTE). Other parameters, such as envelope properties and infiltration rates, were excluded as they primarily affect thermal loads and are more identifiable with higher-resolution or thermal-specific data.

In this study, the building energy model was developed using EnergyPlus as the simulation engine, with jEPlus used for parameterization and batch simulations. A surrogate model was constructed to approximate the relationships between model inputs and outputs:

$$E(t) = f_{surrogate}(CDD(t), HDD(t), EPD, LPD, CSP, FTE) + \delta(t) + \varepsilon(t)$$

Here, $f_{\text{surrogate}}$ adopts a polynomial regression form, fitted to the simulation data using the least squares method ($R^2 > 0.95$). This surrogate model replaces EnergyPlus evaluations in Bayesian calibration, reducing computation time from approximately 4 minutes to less than 1 second per simulation. $E(t)$ represents electricity usage (kWh) at month t ; CDD and HDD indicate cooling and heating degree days, respectively; and $\varepsilon(t)$is a random error term assumed to follow a normal distribution $N(0, \sigma^2)$. In this study, a prototype medium office model was adopted as a proxy of the actual building. The model structure discrepancy was addressed using the Kennedy–O'Hagan framework through the discrepancy term $\delta(t)$. The Bayesian model is formulated as follows:

$$\theta = \{EPD, LPD, CSP, FTE, \delta, \sigma\} \sim Prior(\cdot)$$
$$E_{pred}(t) = f_{surrogate}(\cdot, \theta) + \delta(t)$$
$$y(t) \mid \theta \sim Normal(E_{pred}(t), \sigma^2)$$
$$\theta \mid y \propto Likelihood \times Prior$$

The likelihood assumes that the observed monthly electricity usage $y(t)$follows a normal distribution with mean $E_{\text{pred}}(t)$ and standard deviation σ. The No-U-Turn Sampler (NUTS) was used to generate four MCMC chains, each containing 2,000 samples (including 1,000 burn-in iterations), resulting in 4,000 posterior samples for inference. Convergence diagnostics were assessed using the Gelman–Rubin statistic ($\hat{R} < 1.1$) and effective sample size (ESS > 400) criteria.

Prompt design of LLM

To investigate how prompt design influences the ability of LLMs to generate reliable prior distributions for energy calibration, this study designed a set of three structured prompts using GPT-4o, a commonly used and relatively mature LLM in recent energy simulation studies. Each prompt was tested independently in isolated chat sessions using default model settings to avoid carry-over effects. Output consistency was verified through repeated runs and additional validation in Visual Studio Code at temperature zero. The priors reported are from a single representative run for each prompt level, used directly without modification.

As shown in Table 1, each prompt follows the same structural format and consists of (a) a defined expert role assigned to the LLM, (b) a task description requesting prior distributions for four parameters (EPD, LPD, FTE, and CSP), and (c) a block of building-related information. The expected output is a markdown-style table reporting the prior distribution and a brief justification. The prompt structure was developed following the prompt engineering guidelines of OpenAI and recent applications in building energy modeling (Zhang et al., 2024). The prompts differ only in the depth of building information provided, forming three prompt levels of increasing information exposure, from basic contextual descriptors to detailed building and operational specifications. The low-information prompt (P1) includes only coarse descriptors, such as building type, climate zone, and basic operational schedule, and is intended to characterize the building at a general typological level. The medium-information prompt (P2) extends this context by adding building size, number of stories, certification level, and typical operating hours, providing a more detailed description of building operation. The high-information prompt (P3) further incorporates detailed spatial functions and usage patterns. This hierarchical prompt structure reflects how information typically becomes available in building energy modeling. This design provides a controlled way to vary prompt information content and examine how increasing contextual detail conditions the behavior of LLM-generated priors.

Table 1. Prompt design

<table>
<tr><th>No.</th><th>Content</th></tr>
<tr><td>Prompt 1
(Low)</td><td>You are a highly intelligent building energy modeling expert (IQ 160).
Task: Estimate prior probability distributions for the following four building energy simulation calibration parameters:
- Equipment Power Density (EPD) [W/m^2]
- Lighting Power Density (LPD) [W/m^2]
- AHU Fan Total Efficiency (FTE) [0–1]
- Cooling Setpoint Temperature (CSP) [°C]
Context:
- Building type: Medium-sized educational facility
- Location: Syracuse, NY (Climate Zone 5A)
- Operation: Weekdays only
Instructions:
For each parameter, provide:
- Suggested distribution type (e.g., normal, uniform)
- Estimated parameter values (e.g., min/max, mean/standard deviation)
- A brief justification
Format your response as a markdown table with four columns:
Parameter | Distribution | Values | Justification</td></tr>
<tr><td>Prompt 2
(Medium)</td><td>You are a highly intelligent building energy modeling expert (IQ 160).
Task: Estimate prior probability distributions for the following four building energy simulation calibration parameters:
- Equipment Power Density (EPD) [W/m^2]
- Lighting Power Density (LPD) [W/m^2]
- AHU Fan Total Efficiency (FTE) [0–1]
- Cooling Setpoint Temperature (CSP) [°C]
Context:
- Building type: University student center
- Certification: LEED Platinum
- Size: 54,000 ft^2, 3 stories
- Location: Syracuse, NY (Climate Zone 5A)
- Operation: Monday to Friday, 7:00 AM – 6:00 PM; closed on weekends and holidays
Instructions:
For each parameter, provide:
- Suggested distribution type (e.g., normal, uniform)
- Estimated parameter values (e.g., min/max, mean/standard deviation)
- A brief justification
Format your response as a markdown table with four columns:
Parameter | Distribution | Values | Justification.</td></tr>
</table>

<table>
<tr><td>Prompt 3
(High)</td><td>You are a highly intelligent building energy modeling expert (IQ 160), specializing in HVAC systems and internal load modeling.
Task: Estimate prior probability distributions for the following four building energy simulation calibration parameters:
- Equipment Power Density (EPD) [W/m^2]
- Lighting Power Density (LPD) [W/m^2]
- AHU Fan Total Efficiency (FTE) [0–1]
- Cooling Setpoint Temperature (CSP) [°C]
Context:
- Building type: University student center
- Certification: LEED Platinum
- Size: 54,000 ft^2, 3 stories
- Location: Syracuse, NY (Climate Zone 5A)
- Function by level:
  - Lower level: Multipurpose classroom (also for exhibitions), laboratories, mechanical room
  - First level: Bookstore, student dining area, conference and event spaces, display zones
  - Second level: Office spaces and an outdoor terrace
- Operation: Monday to Friday, 7:00 AM – 6:00 PM; closed on weekends, holidays, and academic breaks; occasional evening events
Instructions:
For each parameter, provide:
- Suggested distribution type (e.g., normal, uniform, triangular)
- Estimated parameter values (e.g., min/max, mean/standard deviation)
- Justification based on building use and industry standards (e.g., ASHRAE)
Format your response as a markdown table with four columns:
Parameter | Distribution | Values | Justification.</td></tr>
</table>

Data Analysis Procedure

The rolling-window approach was used to track the posterior distribution evolution. Starting with the first 12 months of data, the window was advanced by 1.5 months each iteration, with Bayesian calibration repeated at every step until all 36 months of training data were used. This process generated 24 sequential calibration steps. For each step, the posterior mean and standard deviation of the four parameters were recorded. Convergence was defined when the posterior standard deviation decreased to less than 50% of its prior standard deviation. This allows the quantification of the convergence step (the number of steps required to converge) and the information gained from priors to posteriors. The Kullback–Leibler (KL) divergence, Wasserstein distance, mean difference, and the standard deviation difference were used to quantify differences among priors. Model prediction accuracy was evaluated following ASHRAE Guideline 14, using CVRMSE (cross-validated coefficient of variation of the root-mean-square error) and NMBE (normalized mean bias error) as primary indicators. Convergence efficiency was resented by convergence steps and posterior standard deviation reduction rate. MCMC sampling quality was evaluated using ESS, $\hat{R}$ for residual independence. Uncertainty quantification included the 95% prediction interval width, coverage rate, and interval efficiency. Model comparison was conducted using the Watanabe–Akaike Information Criterion (WAIC) and Bayes factors.

Results

Figure 1 illustrates the priors of the four parameters generated from the three prompts (P1, P2, and P3), where the vertical axis indicates the probability density. For EPD, both P1 and P2 produced a Normal (10, 2) distribution, while P3 generated a broader Normal distribution. For LPD, P1 generated Normal (8, 1.5), whereas P2 and P3 produced Normal (7, 1.5), resulting in a mean difference of 1.0 W/m². For CSP, P1 shows a wider Uniform distribution than P2 and P3. The notable difference appears in FTE: P1 and P2 produced a Beta (2, 5) distribution (mean = 0.286, left-skewed), whereas P3 generated a Triangular (0.5, 0.7, 0.85) distribution.

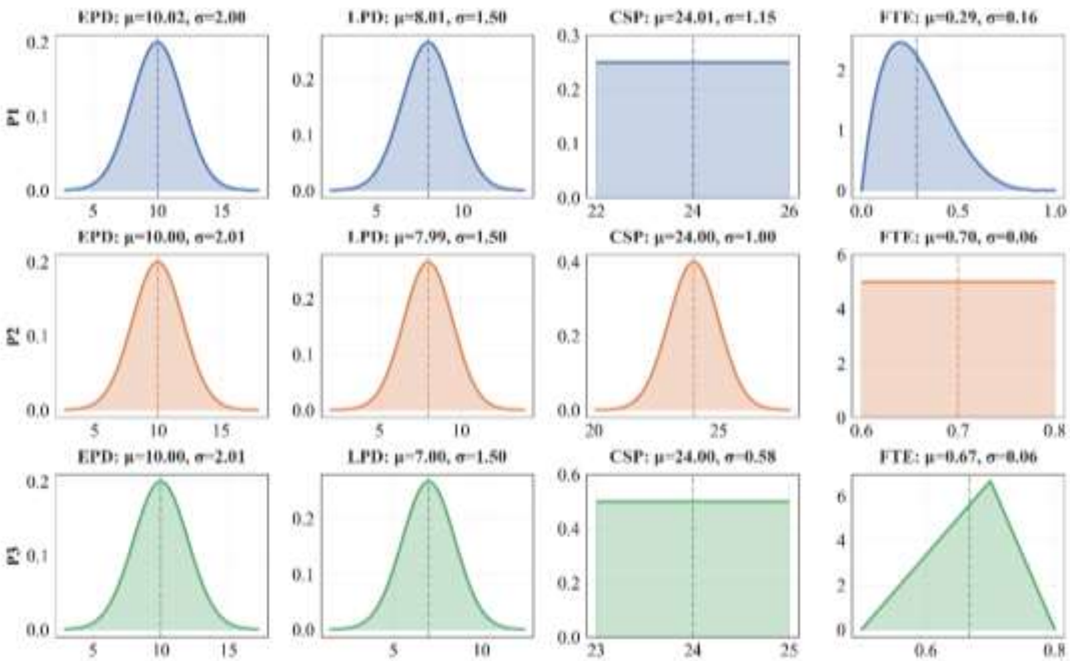

Figure 1. Prior distribution comparison

Figure 2 presents four distance metrics comparing the three prior specifications across four parameters. The KL divergence indicates that FTE exhibits the largest distributional differences between P1-P2 (12.941) and P1-P3 (11.811), while P2-P3 shows minimal divergence (0.004). The Wasserstein distance reveals that LPD has a notable geometric difference between P1 and P3 (1.017 W/m²). The standard deviation difference shows that P1 assigns higher uncertainty to CSP compared to P3 (0.570°C). Across all four metrics, P2 and P3 demonstrate consistent similarity, whereas P1 differs from both, particularly in FTE and LPD. Figure 3 further shows the time-series prediction and residual analysis results. In the main panel, the predicted curve of P1 matches the measured data more closely than P2 and P3, particularly during the high-load period. The residual distributions (inset panel) show that P1 residuals are centered near zero and exhibit the smallest spread, consistent with the lowest standard deviation.

Further quantitative analysis of the 95% prediction intervals shows that P1 produces the narrowest intervals, 22% narrower than P2 and 36% narrower than P3. The coverage rate of P1 is 91.7% (11 out of 12 months within the interval), with the only exception being February 2021. P3 achieves 100% coverage but at the cost of overly wide intervals. The interval efficiency, defined as the ratio of coverage to interval width, shows that P1 achieves the highest value, 18% higher than P2 and 47% higher than P3. The convergence analysis shows that P1 converged within approximately six iterations, compared to nine for P2 and thirteen for P3, with the difference most pronounced for FTE.

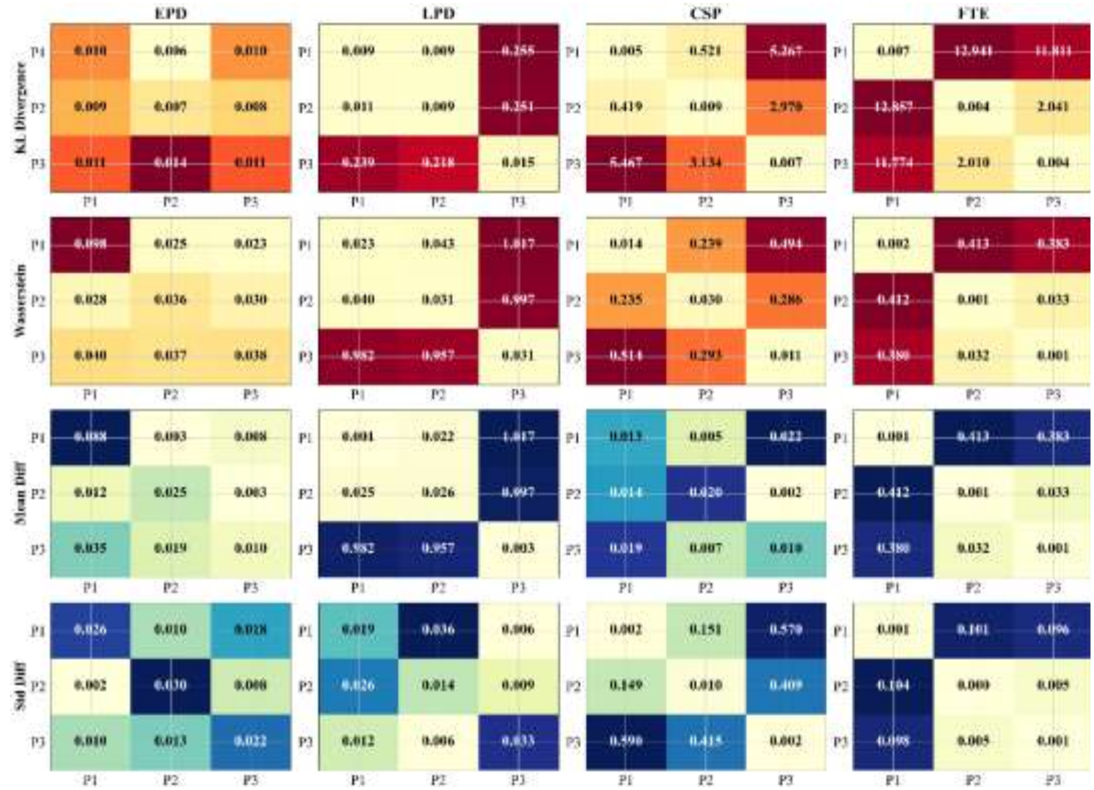

Figure 2. Distance metrics across parameters

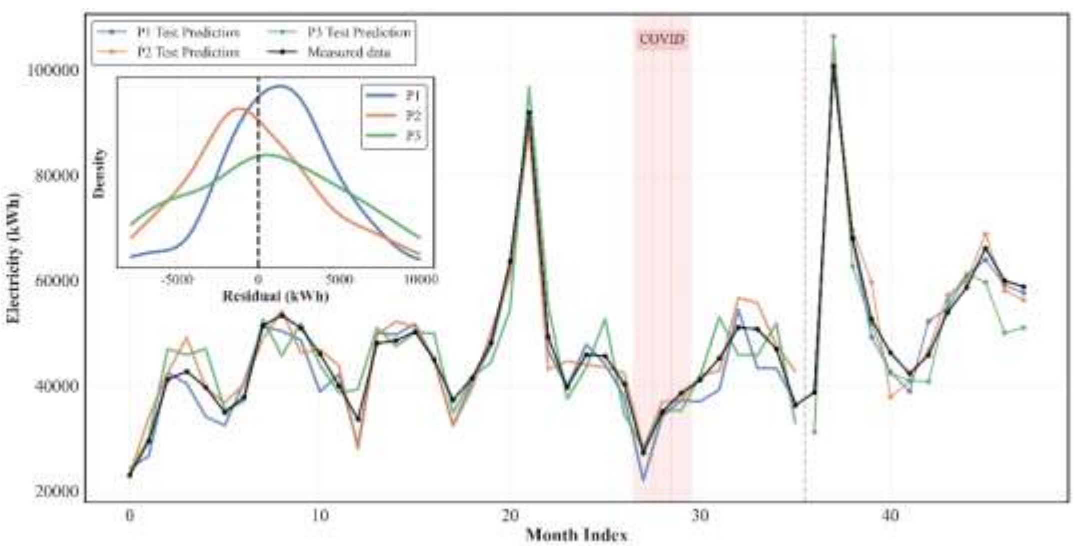

Figure 3. Prediction accuracy comparison

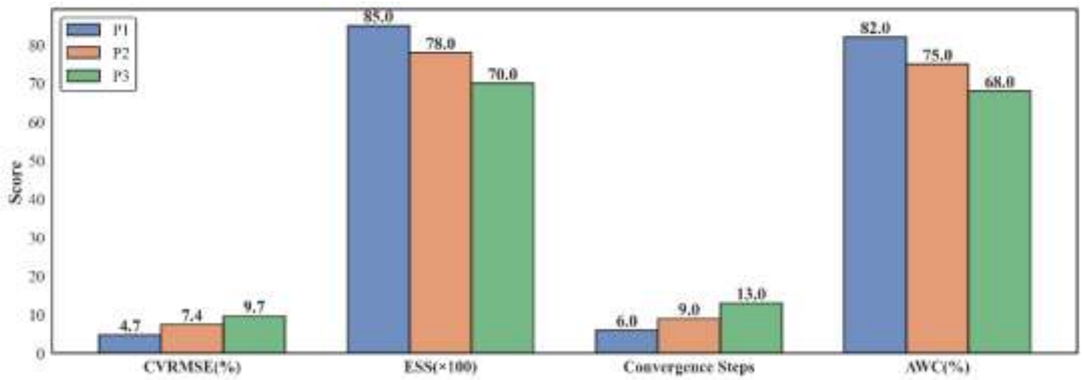

Figure 4. Key performance metrics comparison

Figure 4 shows the MCMC convergence diagnostics and model performance across the three priors. P1 shows the best overall performance among all prompts. It achieves the lowest CVRMSE (4.7%) and the highest ESS, indicating faster and more stable convergence. The convergence steps required for P1 is the smallest, confirming its superior efficiency. P1 also exhibits the greatest average width contraction (AWC), suggesting the highest reduction in parameter uncertainty. In addition to the metrics shown in Figure 4, the $\hat{R}$ and WAIC further support these findings. P1 achieves an average $\hat{R}$ of 1.010 (all parameters below 1.012), satisfying the "excellent convergence" criterion ($\hat{R} < 1.01$). Its WAIC value (512.3) is lower than P2 (528.7) and P3 (547.1), with a corresponding Bayes factor indicating decisive evidence in favor of P1. According to ASHRAE Guideline 14, P1 reaches the "Outstanding" performance level (CVRMSE < 5% and NMBE < ±1%), while P2 and P3 meet only the "Excellent" level (CVRMSE < 10%). These results confirm that P1 achieves the best trade-off between model fit and complexity, receiving the strongest Bayesian support among the three priors.

Discussion

Prompt design plays a key role in how LLM interprets the context and produces the prior distributions. In this study, three prompt levels were constructed by changing only the amount of building information, allowing an initial observation of how information richness affects prior generation ability of LLM. Results show that adding more contextual details may not always improve the quality of priors, suggesting that LLM using high-information prompt may focus too heavily on specific descriptive details and may generate priors that are overly constrained or biased. In this case, the low-information prompt leads the LLM to generate priors with less constrained, representing more general engineering knowledge learned by the model, acting to a "weakly informative prior" in Bayesian analysis. Additionally, prompt design should follow the LLM's instruction structure and linguistic logic, so that the LLM can interpret the task correctly and generate priors that reflect domain knowledge. Besides, building energy calibration often faces limited measured data, such as monthly electricity use or short monitoring periods, as in this study. Under such conditions, the priors play a much larger role in shaping the posterior results. When the dataset is small, a strong or narrow prior can easily dominate the posterior, leaving little room for the data to adjust the model. This explains why the high-information prompt, which produced more concentrated priors, performed worse in both convergence and accuracy. The low-information prompt, in contrast, provided wider priors that allowed the data to correct and refine the estimates more effectively, resulting in more stable predictive behavior.

These results highlight that, in this case study, the value of LLM lies less in producing highly detailed distributions than in providing priors that are sufficiently broad and stable to support data-driven calibration. Overly narrow priors may constrain posterior updating and lead to biased inference. This ensures that the posterior inference remains data-driven rather than prompt-driven. In practice, prompt design should be treated as a calibration step itself when applying LLMs for Bayesian calibration, with priority given to accurate and essential building information rather than maximizing descriptive detail. For early-stage modeling or projects with limited monitoring data, practitioners can start with prompts containing only essential building descriptors (e.g., building type, climate zone, and general operating schedule) to generate initial priors. If calibration performance is unsatisfactory, prompt detail can be incrementally adjusted rather than maximized.

However, the findings should be interpreted within the scope of the study's constrained design. The results do not support generalizable conclusions across different building types, climates, parameter sets, or data resolutions. The observed behaviors may vary with higher-frequency data, alternative calibration targets, or multiple buildings. Further studies are needed to assess the transferability of these observations to broader contexts. Besides, while less detailed prompts produced more adaptable priors in this context, the lower bound of prompt simplification was not tested. Prompts that exclude necessary information such as building type or climate zone may fail to elicit meaningful priors. Identifying the minimum information threshold for effective prior generation remains a direction for future research.

Conclusion

This study reports a case-based methodological examination of how prompt design influences the characteristics of LLM-generated priors within a Bayesian calibration process for building energy simulation. Within this specific modeling context, the results suggest that increasing prompt detail does not necessarily lead to more effective priors. Under the same prompt structure, overly detailed prompts may result in narrower or more biased prior distributions, limiting the capacity for data-driven updating during calibration. These observations suggest that effective prompt design in this context should consider balancing informativeness and data availability, rather than maximizing descriptive detail only. This work provides methodological insight into the behavior and feasibility of using LLM-generated priors within Bayesian calibration under limited data. The findings serve as a reference for future studies exploring LLM-generated prior formulation and highlight the need for further multi-building and multi-context investigations to assess the broader applicability of these observations.

Acknowledgment

We greatly appreciate the financial support from U.S. Department of Energy (ARPA-E Award No. DE-AR0001639) and Syracuse Center of Excellence (SyracuseCOE).

References

Attia, S., Shadmanfar, N., & Ricci, F. (2020). Developing two benchmark models for nearly zero energy schools. *Applied Energy*, *263*. https://doi.org/10.1016/j.apenergy.2020.114614

Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, *25*(3). https://doi.org/10.1016/0378-3758(90)90079-A

Calama-González, C. M., Symonds, P., Petrou, G., Suárez, R., & León-Rodríguez, Á. L. (2021). Bayesian calibration of building energy models for uncertainty analysis through test cells monitoring. *Applied Energy*, *282*. https://doi.org/10.1016/j.apenergy.2020.116118

Chong, A., Lam, K. P., Pozzi, M., & Yang, J. (2017). Bayesian calibration of building energy models with large datasets. *Energy and Buildings*, *154*. https://doi.org/10.1016/j.enbuild.2017.08.069

Chong, A., Xu, W., Chao, S., & Ngo, N. T. (2019). Continuous-time Bayesian calibration of energy models using BIM and energy data. *Energy and Buildings*, *194*. https://doi.org/10.1016/j.enbuild.2019.04.017

Coakley, D., Raftery, P., & Keane, M. (2014). A review of methods to match building energy simulation models to measured data. In *Renewable and Sustainable Energy Reviews* (Vol. 37). https://doi.org/10.1016/j.rser.2014.05.007

Hatakeyama-Sato, K., Yamane, N., Igarashi, Y., Nabae, Y., & Hayakawa, T. (2023). Prompt engineering of GPT-4 for chemical research: what can/cannot be done? *Science and Technology of Advanced Materials: Methods*, *3*(1). https://doi.org/10.1080/27660400.2023.2260300

Hong, T., & Zhang, L. (2025). AI for building energy modeling: A transformation. *Building Simulation*, *18*(9), 2219–2225. https://doi.org/10.1007/s12273-025-1329-4

Hou, D., Hassan, I. G., & Wang, L. (2021). Review on building energy model calibration by Bayesian inference. In *Renewable and Sustainable Energy Reviews* (Vol. 143). https://doi.org/10.1016/j.rser.2021.110930

Jiang, G., Ma, Z., Zhang, L., & Chen, J. (2024). EPlus-LLM: A large language model-based computing platform for automated building energy modeling. *Applied Energy*, *367*, 123431. https://doi.org/10.1016/j.apenergy.2024.123431

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. In *Learning and Individual Differences* (Vol. 103). https://doi.org/10.1016/j.lindif.2023.102274

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *63*(3). https://doi.org/10.1111/1467-9868.00294

Krishna Padmanabhan, D. B. (2025). *BACTA-GPT: An AI-Based Bayesian Adaptive Clinical Trial Architect*. *arXiv:2507.02130*. https://doi.org/10.48550/arXiv.2507.02130

Ruggeri, F., Sánchez-Sánchez, M., Sordo, M. Á., & Suárez-Llorens, A. (2021). On a New Class of Multivariate Prior Distributions: Theory and Application in Reliability. *Bayesian Analysis*, *16*(1). https://doi.org/10.1214/19-BA1191

Wang, E. (2025). Benchmarking building energy efficiency with Bayesian regression on incomplete data. *Communications in Computer and Information Science*, *2394*. https://doi.org/10.1007/978-3-031-85324-1_16

Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J. (2022). Large language models are human-level prompt engineers. *Eleventh International Conference on Learning Representations*. May 1-May 5, 2023, Kigali Rwanda.

Zhang, L., Chen, Z., & Ford, V. (2024). Advancing building energy modeling with large language models: Exploration and case studies. *Energy and Buildings*, *323*, 114788. https://doi.org/10.1016/j.enbuild.2024.114788

Zhao, T., Xu, J., Zhang, C., & Wang, P. (2021). A monitoring data based bottom-up modeling method and its application for energy consumption prediction of campus building. *Journal of Building Engineering*, *35*. https://doi.org/10.1016/j.jobe.2020.101962