Proceedings of the ASME 2025
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC-CIE2025
August 17-20, 2025, Anaheim, CA

DETC2025-168873

Attendee Read Only Copy

EVALUATING THE ROLE OF MODEL SIZE IN AGENTIC AI FOR EXPERT-LIKE MATERIAL SELECTION

Megan Y. Ying
Carnegie Mellon University
Pittsburgh, PA
mying@andrew.cmu.edu

Daniele Grandi
Autodesk Research
San Francisco, CA
daniele.grandi@autodesk.com

Allin Groom
Autodesk Research
London, UK
allin.groom@autodesk.com

Christopher McComb*
Carnegie Mellon University
Pittsburgh, PA 15232
ccm@cmu.edu

ABSTRACT

Material selection is fundamental to the design process, as it significantly affects the cost, performance, appearance, manufacturability, and sustainability of a product. It is a complex, open-ended challenge that forces designers to continuously adapt to new information, balance diverse stakeholder demands, weigh trade-offs, and navigate uncertainties to achieve the optimal outcome. Previous studies have explored the potential of large language models (LLMs) to assist in the material selection process, with findings suggesting that LLMs could provide valuable support. However, discrepancies between LLM outputs and expert recommendations indicate the need for further research. To address the limitations of standalone LLMs, particularly their lack of reasoning and action-execution capabilities, agentic AI has been developed with enhanced functionalities. These agents integrate LLMs with external search tools, allowing them to retrieve and analyze domain-specific information, iteratively refine responses, and improve decision-making alignment with experts. This study compares standalone LLMs and agentic AI frameworks, examining how search-augmented agents can more effectively emulate expert decision-making in material selection. Our findings reveal a nonlinear relationship between model size and performance, with some models demonstrating lower proximity to human survey results and struggling to follow instructions. These insights contribute to a broader understanding of AI integration in design workflows.

1 INTRODUCTION

Material selection is integral to the development cycle of engineering projects, influencing many aspects of the final product. The intrinsic properties of a material determine how it behaves in response to conditions such as stresses, temperatures, and environmental factors, impacting a design's performance and life cycle [1]. Mechanical and physical properties also affect compatibility with manufacturing processes, influencing the feasibility of achieving the desired geometries [2]. Incompatible material-process pairings can increase costs and reduce performance, while effectively optimized geometry can improve functionality, reduce material waste, and lower costs [3, 4].

Beyond functionality, a material's aesthetic properties, such as color, texture, and finish, shape user perception [5]. Overall product cost is influenced not only by raw material expenses, but also by the complexity and number of manufacturing processes required [6]. Sustainability concerns are also shaping material choices as industries balance environmental impact, re-

*Address all correspondence to this author.

Copyright © 2025 by ASME

source depletion, and consumer demand for eco-friendly solutions [4, 7, 8]. Ultimately, the interplay between material properties, performance, aesthetics, cost, sustainability, and manufacturing makes material selection a complex task.

The design process is often iterative, where initial design intent or product requirements guide material selection, which in turn refine those requirements [9]. With multiple trade-offs and stakeholder negotiations, this iterative process can become time and resource-intensive [10]. While databases and frameworks support material selection, critical insights often lie beyond datasheets, requiring expert knowledge and contextual information from suppliers, case studies, and in-house reports [11].

Advancements in artificial intelligence (AI) and natural language processing (NLP), particularly with the emergence of large language models (LLMs) trained on extensive textual datasets, have significantly enhanced the ability of computers to process and generate meaningful text [12]. Enhanced computational power and the growing availability of datasets in recent years have driven the widespread adoption of LLMs across diverse fields, with one potential application being their use as tools to recommend materials that align with the design intent [13].

Previous studies have collected material recommendations across various design cases and criteria from both human experts and LLMs to assess how well LLM-generated material suggestions aligned with expert choices [14, 15]. While LLMs were capable of generating material suggestions, many did not closely align with expert recommendations. Researchers found that adjusting factors like model type, prompting strategy, and parameter settings improved alignment with expert recommendations, highlighting the need for further research into replicating human decision-making [14].

While LLMs are capable of generating text based on input prompts, they struggle with complex reasoning and tasks in domains such as logic, spatial intelligence, and linguistic understanding [16, 17]. Additionally, previous research has identified a specific challenge in material selection, where LLMs exhibit less variability in their responses for newer materials like composites, and offer less consistent scores for well-established materials like steel. This pattern suggests that older materials may be overrepresented in model outputs due to their historical prevalence in literature rather than their relevance to modern design [14].

AI agents are one approach to overcoming some LLM limitations by incorporating additional tools for perception and interaction with environments, enabling autonomy, planning, problem-solving, and decision-making [18, 19]. In some agentic AI frameworks, the LLM acts as the core reasoning engine, while a perception module interprets external information and an action module executes tasks [20]. Together, these modules enable agentic AI to go beyond passive text generation, allowing agents to perceive task instructions, create step-by-step plans, interact with external systems, and learn from feedback to improve performance [21].

This paper seeks to build on prior research [14, 15] by exploring whether augmenting LLMs with search data can enhance their ability to emulate the decisions of human experts in material selection. Specifically, this study examines the impact of providing an LLM-driven agent with the ability to access Wikipedia and iteratively reason while doing so. We conduct a series of experiments to evaluate the role of model size and prompting techniques in aligning agent recommendations with expert decision-making.

This paper seeks to address the following research questions:

1. How does the use of an agentic AI framework influence alignment with expert material selection behavior, compared to standalone LLMs?
2. How does the agent's behavior vary with respect to model size?
3. To what extent does model size impact the ability of LLM-powered agents to select appropriate materials?

To further support development of LLM methods, our data and code are available on Github [1].

2 BACKGROUND

In this section, we review related works on the challenges of material selection, AI evaluation, and the use of AI in material selection.

2.1 Challenges in Material Selection

Material selection is crucial in product development, influencing the entire lifecycle beyond the early design stages [22,23]. With thousands of materials and processes available, filtering out unsuitable options early requires expertise. Just as material properties can be modified through various processes, impacting both performance and cost, design constraints can also be adjusted, changing the suitability thresholds of material candidates [13]. The tension between design requirements and materials necessitates a robust but flexible selection process. Researchers have developed various frameworks for material selection, all emphasizing screening and ranking as essential steps to refine material options [24].

Sustainability concerns are increasingly shaping material choices as industries respond to climate change, resource depletion, and pollution. Companies are financially incentivized to adopt sustainable materials and manufacturing methods to meet consumer demand, comply with environmental regulations, and reduce costs through more efficient processes [7]. Additionally, material shortages are making traditional materials less accessible, prompting engineers to explore alternative, less-developed

[1] https://github.com/cmudrc/Agents-for-Material-Selection

Copyright © 2025 by ASME

options [8]. This shift affects engineering decisions by prioritizing durability, recyclability, and production efficiency [4].

The Ashby chart, which compares materials by two properties, is widely used but may be limited for designs with multiple constraints [9, 24]. Other methods, such as the questionnaire method [25], case-based reasoning [26], multi-criteria decision analysis [27], and weighted properties [25], rely on measurable attributes, expert input, or prior applications. While quantitative analysis identifies material solutions quickly, it requires expertise in data analysis. Expert knowledge ensures accuracy but can limit innovation, while case-based selection depends on the quality of prior examples and may struggle with novel problems [11].

At the beginning of the design process, material selection focuses on defining core requirements and exploring materials that align with the design intent, considering factors like mechanical, physical, and manufacturing properties [28]. Trade-offs often arise, as materials may meet some but not all requirements [23]. AI is increasingly aiding material discovery by analyzing large datasets, predicting properties, and suggesting optimal combinations [9,16,29]. However, due to uncertain product requirements, designers often default to familiar or reverse-engineered materials, which may not always be optimal [29].

As the design progresses, product requirements become more refined, with broad limitations yielding too many material options and overly restrictive ones in too few. Finalizing constraints makes trade-offs more apparent, emphasizing the need for testing, simulation, and prototyping to validate material choices and address unexpected issues. Engineers must also consider manufacturing processes and system components as designs near fabrication [23]. Due to its multidisciplinary nature, material selection is often treated as an optimization problem, with designers iterating on constraints to identify the best material [10]. Advanced computational tools, such as neural networks, assist by analyzing material properties and design parameters to predict product performance under varying conditions [9, 30].

2.2 Evaluating LLMs and Agentic AI

The emergent behaviors of LLMs, and their wide applicability to various domains, has led researchers to establish numerous benchmarks to evaluate LLM reasoning abilities in many different fields. Moreover, the rapid development of LLMs also demanded an equivalent development of new evaluation methods.

Starting with static test sets such as GLUE [31] and SuperGLUE [32], researchers further developed multi-domain and open-domain benchmarks for automatic evaluation of general language understanding [33–39]. Text-based question-answer benchmarks, or reference-dependent benchmarks, require the model to interpret additional information from a selected corpus, thus allowing for the evaluation of general knowledge, factoid-type questions, questions based on long-text documents, or from multiple sources in the model's pre-trained corpus [40–44]. Multimodal benchmarks and visual question-answer benchmarks typically evaluate a model's capability to answer questions about a specific image, focusing primarily on reasoning and perception abilities [45, 46, 46–49]. Despite much development around automatic evaluation, manual evaluation remains a common and trusted method, as it can provide more comprehensive and accurate feedback on model performance [35, 50–53].

In AI research, the term 'agent' originally referred to an artificial entity capable of perceiving its surroundings through sensors, making decisions, and subsequently taking actions in response [54, 55]. Emergent LLM behaviors have further enhanced the capabilities of AI agents to plan and reason, and coupled with tool usage, modern AI agents can tackle very complex tasks [56–63]. Currently, the utility of LLM-powered agents is evaluated primarily by the success rate of task completion, as agents are typically implemented as human assistants [64, 65]. This metric evaluates whether agents achieve predetermined objectives or expected scores [66–69]. Researchers have also conducted more detailed assessments of agents' foundational capabilities, such as environmental comprehension, reasoning, planning, tool usage, and decision-making [61, 70–72]. Furthermore, the efficiency of the agent in completing the task has also been used as an evaluation criterion, as the agent should complete the task within an appropriate timeframe and with minimal resources [66]. In this work, we build on prior research that looked at material selection and evaluate the performance of an AI agent on the same task.

2.3 Automating Material Selection with AI

Over the years, the material science field has kept evolving to meet the demand to characterize and develop novel materials, such as composites, and engineer entirely new ones with unique property combinations, such as metamaterials [73, 74]. Recent advances in NLP have led the field towards the use of LLMs for material discovery [75], experimental solid mechanics [76], predictive material science [77], biologically inspired materials design [78], and various other applications in the material science field [79].

Understanding a material's mechanics plays an important role in determining the feasibility of the material for a design problem, however, as discussed in the previous section, many other factors play a role in material selection for product design. To this end, a few review papers have evaluated LLM capabilities against mechanical engineering design tasks, including material selection and other manufacturing-related tasks [80, 81]. LLMs have been applied to forward and inverse material design problems [82] and knowledge retrieval from a material failure mechanics corpus [83]. Buehler and colleagues expanded the approach by leveraging multi-agent systems to solve mechanics problems and automate alloy design [84, 85].

Copyright © 2025 by ASME

In this paper, we bridge recent work on agentic AI frameworks with the material selection task in product design, improving on prior work and reducing the bias that LLMs exhibit on this task [14].

3 METHODS

In this section, we describe the methodology used for our study. Our approach involves using a previously collected dataset of expert material selection decisions, developing a tool-augmented LLM-based agent, testing prompting strategies across model sizes, and conducting analyses to assess performance and behavior.

3.1 Expert Data Collection

To address our research questions, we used the MSEval[2] dataset collected by Jain *et al.* [15]. MSEval was obtained through a survey conducted among material selection professionals, including materials scientists, materials engineers, design engineers, and others in related fields, via the Autodesk Research Community.

The experts were asked to assess a set of nine materials (*steel, aluminum, titanium, glass, wood, thermoplastic, elastomer, thermoset, and composite*) based on their suitability for sixteen different scenarios, assigning a score between 0 and 10, with 0 representing an unsatisfactory choice and 10 representing an excellent choice. These scenarios included all possible combinations of four design applications (*kitchen utensil grip, safety helmet, spacecraft component, and underwater component*) and four design criteria (*lightweight, heat resistant, corrosion resistant, and high strength*).

The selection of these applications, criteria, and materials aimed to cover a wide range of use cases and maximize the likelihood of identifying both suitable and unsuitable design options. Only responses from participants who indicated experience in material selection were retained. Although not all participants answered every question, all responses were included, resulting in a data set collected from 138 material selection experts.

3.2 Design of LLM-Based Agent

To build our LLM-based agent, we used the Hugging Face Transformers library. To align with expert decision-making, we designed the agent to evaluate multiple factors critically and solve problems effectively. We chose the ReAct framework for its step-by-step approach, where reasoning and actions dynamically refine plans, enhancing accuracy and reducing hallucinations [86]. Additionally, research suggests that LLM-based agents are better equipped to manage complex tasks and interact with external systems when using executable code [87]. Therefore, we chose an agent that generates Python code instead of JSON blobs for tool calls.

To initialize a Hugging Face agent, key parameters must be defined, including selecting an LLM [88]. We chose the Qwen 2.5 series for its strong performance and range of model sizes, measured by the number of parameters, or adjustable values that determine how the model understands and generates language. This range, from 0.5 billion to 72 billion parameters, allows us to examine the impact of model size on replicating expert decision-making [89]. These models were trained on a dataset of 18 trillion tokens, the smallest units of text that an LLM can process, surpassing the 7 trillion used in previous versions. This led to significant advancements in text generation, mathematics, coding, instruction-following, and reasoning [89].

Furthermore, the agent requires a toolbox from which it can select tools to use. While the Hugging Face Transformers library includes default tools, we determined that our agent needed internet access to search for material characteristics, identify comparable designs, and retrieve real-time data. We developed a custom tool class that utilizes the Wikipedia API to perform searches and retrieve summaries of the top results [88].

Finally, the agent must be initialized with a system prompt that defines its input, expected behavior, available tools, and output format [88]. Clear instructions and examples within the prompt are crucial for achieving accurate results [90]. We customized the default system prompt to include examples of using the Wikipedia search tool to evaluate material effectiveness in various design cases, avoiding survey data to prevent bias. An example of these different components interacting with each other in the agentic framework is shown in Figure 1.

3.3 LLM Experiments

In order to address our research questions we perform an experiment in which we compare our agentic framework using zero-shot prompting to standalone LLMs using zero-shot, few-shot, parallel, and chain-of-thought prompting methods. Zero-shot prompting provides only task instructions, offering high interpretability, the absence of examples, and a more flexible prompt [91]. In contrast, few-shot prompting involves one or more examples for better context, which helps optimize model performance [92]. For the few-shot evaluation, we used two examples not included in our survey.

A parallel approach directs the model to generate all nine scores in a single query, providing faster inference, reduced computational costs, and better memory efficiency. However, it requires additional post-processing and may increase the risk of model hallucinations [93, 94]. Additionally, research has shown that generating a sequence of intermediate reasoning steps, or a chain of thought, can enhance LLM performance in reasoning tasks [95]. We evaluate a two-step chain-of-thought approach,

[2] https://github.com/grndnl/llm_material_selection_jcise

Copyright © 2025 by ASME

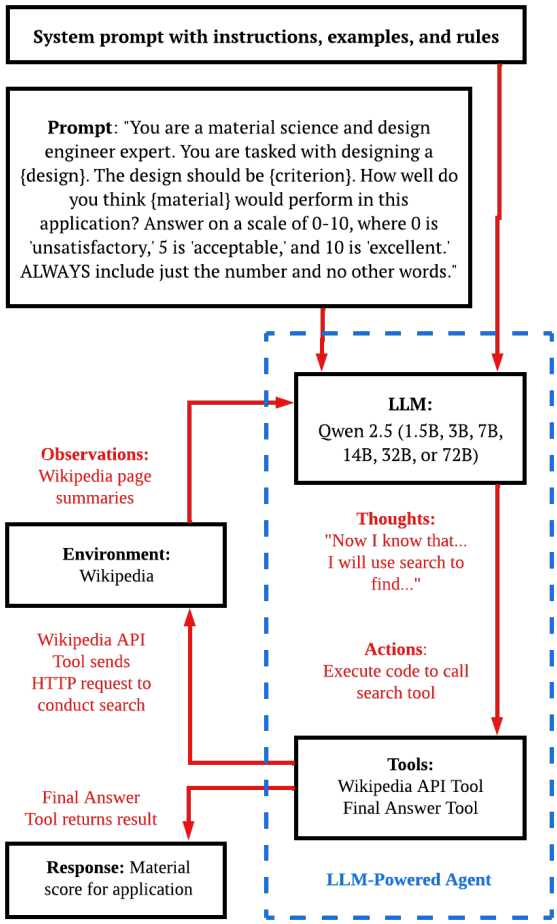

FIGURE 1. A visualization depicting the dynamic workflow of the agentic AI used in this work, illustrating how the prompt, LLM, tools, and environment interact through an iterative process of thoughts, actions, and observations to produce a final response.

first prompting the model to explain how a material would perform in a specific design scenario, then integrating that explanation into the final query where the model provides a score. The exact prompts for each approach (*zero-shot, few-shot, parallel, and chain-of-thought*) can be found in the paper Grandi *et al.* [14].

In our agentic experiments, we retained only responses where the final answer tool was successfully called and the last search query did not produce an error, ensuring that the results accurately reflected the agent's reasoning process. We also excluded responses that did not return an answer within the specified 0-10 range. If a failure occurred during reasoning, we reran the prompt up to five times with the same design, material, and criterion combination before excluding it from analysis. The number of reruns, completion tokens, and types of search queries made by the agent were logged for further analysis of the agent's internal behavior.

We tested these five methods (*agentic, zero-shot, few-shot, parallel, and chain-of-thought*) across six different Qwen 2.5 model sizes (*1.5B, 3B, 7B, 14B, 32B, and 72B*) to assess the effects of both model size and prompting type. The results were then saved for further analysis[3].

3.4 Evaluation

We compute the average z-score and mean absolute error (MAE) of the generated score relative to the to distribution of survey data to measure alignment with material selection professionals. We assess how these metrics vary across model sizes and prompting strategies.

We compute the z-score for each LLM-generated score to measure its deviation from survey responses for the same material. The z-score is calculated as

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the score, μ is the mean of survey responses evaluating that material, and σ is the standard deviation of survey responses for that material.

The MAE is calculated as

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2}$$

where n is the number of responses, y_i is the ith expert score, and $\hat{y}_i$ is the ith predicted score.

To explore the agent's internal behavior for our second research question, we analyzed its performance across different model sizes using several metrics. For each model size, we measured the number of reruns per prompt, both overall and for successfully retained responses. We also examined search behavior by calculating the total number of search queries per design, criterion, and material combination, along with the proportion of unique queries to assess search diversity. Lastly, we quantified token usage by computing the number of completion tokens generated by the agent. We averaged these metrics across all test cases.

To determine the similarity between search query terms, we used sentence embeddings, which represent words as continuous-valued vectors capturing their syntactic and semantic properties and thereby enabling effective measurement of textual

[3]Results from the 14B model are excluded from most analyses due to a very high error rate, which is discussed further in Section 4.3

Copyright © 2025 by ASME

similarity [96, 97]. We used Hugging Face's Sentence Transformers, a class of transformer models designed for generating sentence embeddings and selected MiniLM as the most effective model for clustering our data points. To prevent clustering based solely on specific design, criterion, or material, we filtered queries by replacing design applications with '{design},' criteria with '{criterion},' and materials with '{material}' before representing the queries as word embeddings.

To mitigate computational costs associated with the high dimensionality of word embeddings, we apply Principal Component Analysis (PCA), a linear method that reduces dimensionality by projecting the data onto principal components while preserving overall variance [98]. After performing dimensionality reduction, we visualize the word embeddings of agent queries alongside system prompt queries to evaluate whether the agent queries cluster around the prompt queries.

To answer our third research question, we used multiple linear regression to model the relationship between the predicting factors and the results. We ran several regression models, with MAE or z-score as the dependent variable, and prompting strategy and model size as independent variables. Since our independent variables include both categorical and numerical data, we applied one-hot encoding to convert categorical data into a binary format suitable for linear regression. Each prompting method (*agentic, zero-shot, few-shot, parallel, and chain-of-thought*) is represented by a separate column, with a value of 1 indicating the presence of that method and 0 indicating its absence. We then assessed the statistical relationship between the predictor and response variables using the coefficients and p-values.

A coefficient in our regression model represents the change in the dependent variable for a unit increase in an independent variable, with all other predictors held constant. By setting the agentic approach as the baseline with a coefficient of 0, we interpret other coefficients as the change in the dependent variable relative to this baseline. The p-value indicates the likelihood of obtaining a result as extreme as the one observed, assuming the independent variable has no effect on the dependent variable. A low p-value indicates statistical significance, allowing us to reject the null hypothesis and conclude that the predictor impacts the response, while a high p-value suggests no significant relationship between the variables.

4 RESULTS AND DISCUSSION

In this section, we present results from our experiments, focusing on the inference process of the LLM-based agent, the impact of model size and prompting strategy on alignment with expert decisions, and behavioral trends observed across agentic AI.

4.1 Example of LLM-Based Agent Inference Process

The agent's inference process begins by determining what information to search for and executing the necessary search queries. This may involve searching for key material properties and essential characteristics required for the target application. After gathering this information from Wikipedia, the agent evaluates the material's compatibility with the design requirements. Finally, based on its findings, it assigns a numerical rating.

Figure 2 illustrates an example of the agent's inference process in evaluating composites for lightweight spacecraft components. It performs several unique, targeted Wikipedia searches based on the prompt and newly acquired information, refining its assessment through multiple iterations. By synthesizing material properties and engineering constraints, the agent ensures that its final conclusion aligns with key requirements specified in the prompt, such as strength-to-weight ratio. This approach brings the agent closer to replicating the reasoning process of a material selection expert, who may draw on domain knowledge and prior research to iteratively refine conclusions.

However, some of the Wikipedia results retrieved by the agent are repeated or irrelevant, suggesting that Wikipedia may not always provide the necessary information. Additionally, in this example, the agent primarily focuses on weight and strength, while other critical factors, such as thermal stability and environmental resistance, should also be considered when evaluating the suitability of a material for a given application and criterion.

4.2 Effects of Model Size and Prompting Strategy on LLM Performance

Regression analysis is used to develop models that relate experimental parameters to key LLM performance metrics. Specifically, we aim to predict proximity to survey results, measured by z-score and MAE, based on model size and prompting strategy. This relationship is mathematically expressed as $\hat{z} = f(C,S)$ and $\hat{MAE} = f(C,S)$, where $\hat{z}$ is the predicted z-score, $\hat{MAE}$ is the predicted MAE, S is the model size, and C denotes prompt type. S is treated as a continuous variable. We represent C using several binary categorical variable; C_A represents the use of an agentic framework, C_Z represents zero-shot prompting, C_F represents few-shot prompting, C_P represents parallel prompting, and C_C represents chain-of-thought prompting. We consider both main effects and secondary effects in the regression model.

The coefficients for each of these variables from our linear regression model for predicting z-scores are shown in Figure 3, while those for predicting MAE metrics are shown in Figure 4. Our results indicate that the interaction between model size and prompt type has minimal impact on performance for both z-score and MAE, as they do not exhibit statistical significance with a p-value greater than 0.1.

Coefficient values for C are relative to the effect of C_A. We observe that zero-shot, few-shot, and chain-of-thought prompt-

Copyright © 2025 by ASME

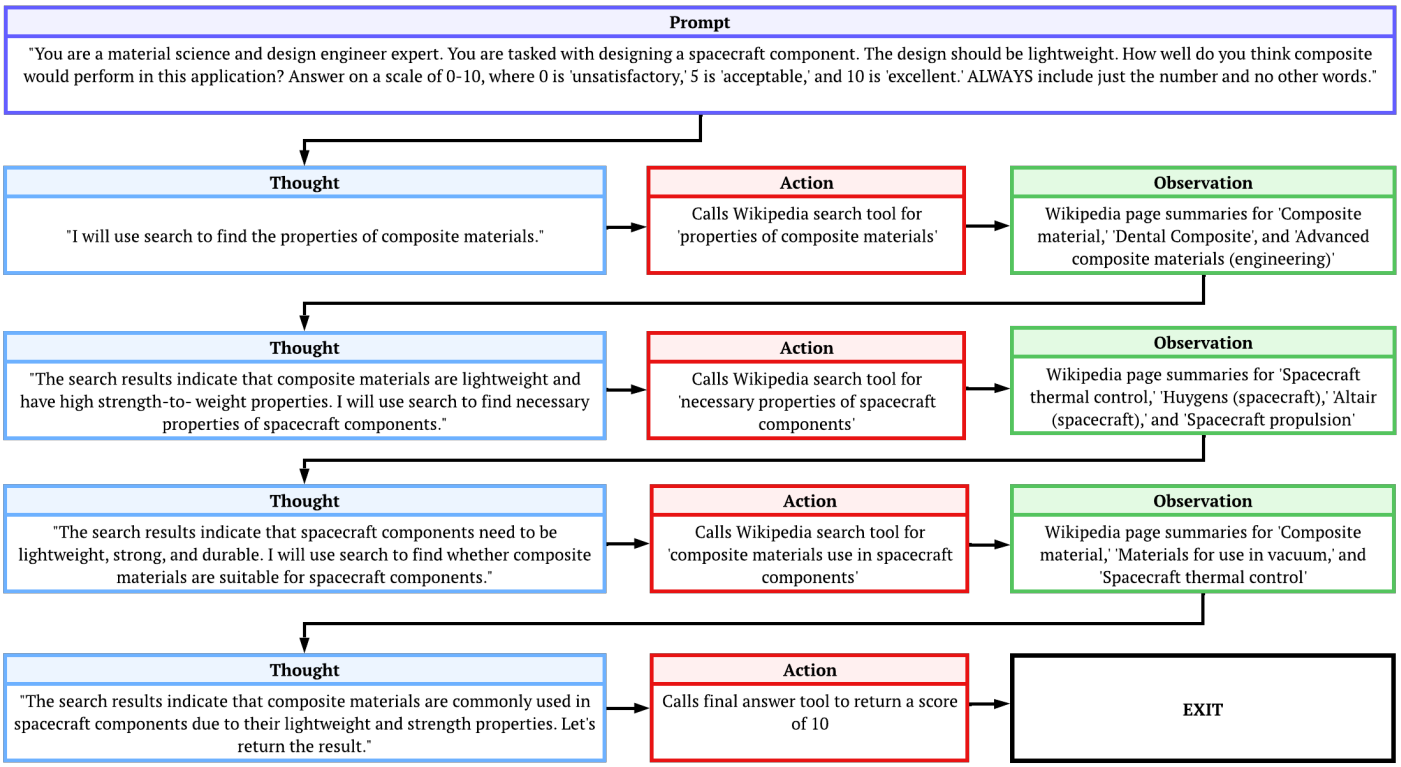

FIGURE 2. A flowchart of the LLM-based agent's reasoning process when asked to assess the suitability of composite materials for a lightweight spacecraft component. The agent can interact with Wikipedia, determine when and what to search, and use findings to draw conclusions.

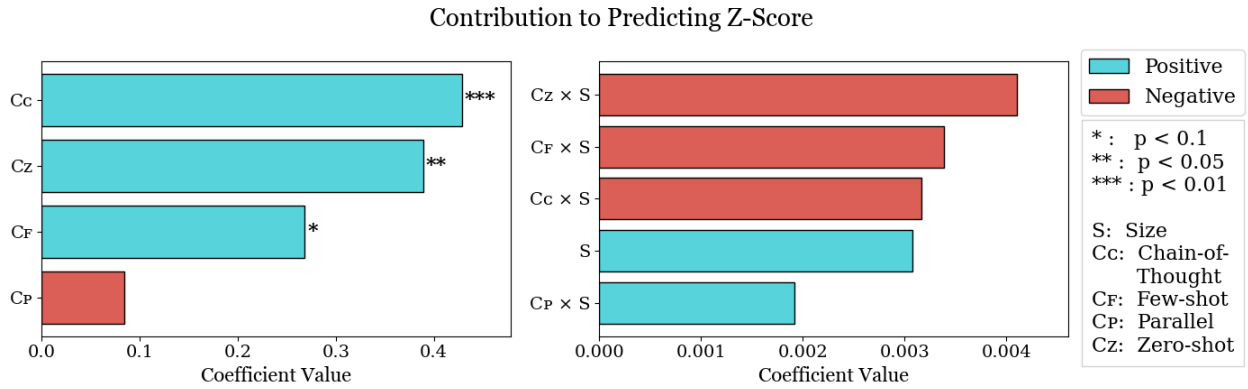

FIGURE 3. The contributions of model size and prompt type to predicting z-score are represented as coefficients in the linear regression model. The p-values and coefficients indicate that chain-of-thought, zero-shot, and few-shot prompting with standalone LLMs lead to a significant increase in z-score compared to the agentic framework.

ing methods lead to a statistically significant increase in the z-score compared to the agentic framework, suggesting that these strategies worsen model performance. On the other hand, while parallel prompting appears to decrease the z-score, its effects are not statistically significant. This indicates that the benefit of an agentic framework over parallel prompting is negligible. Additionally, there appears to be minimal change in the z-score as model size increases, demonstrating a non-linear effect of size on proximity to survey results.

In contrast, all prompting methods using standalone LLMs

Copyright © 2025 by ASME

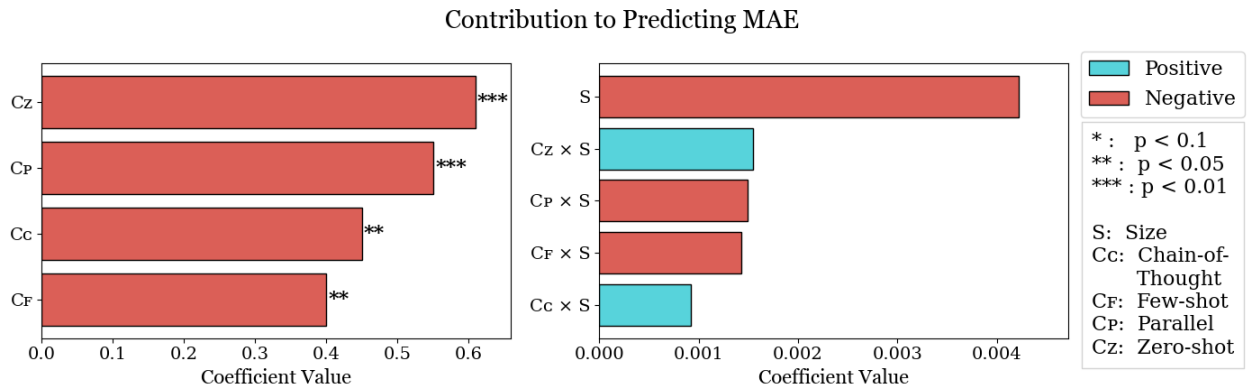

FIGURE 4. The contributions of model size and prompt type to predicting MAE relative to survey data are represented as coefficients in the linear regression model. The p-values and coefficients indicate that all prompting strategies used with standalone LLMs result in a significant decrease in MAE compared to the agentic framework.

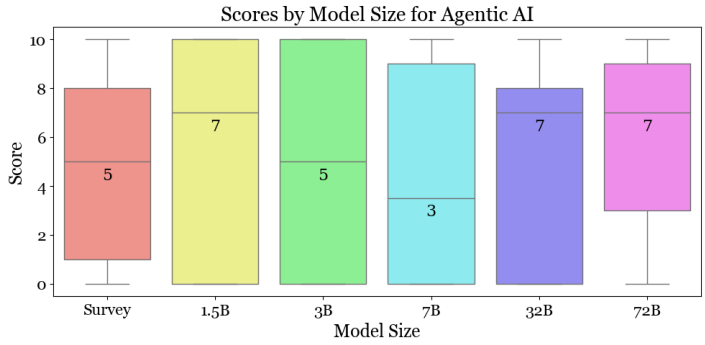

FIGURE 5. The scores from our agentic experiments analyzed by model size, with median values labeled.

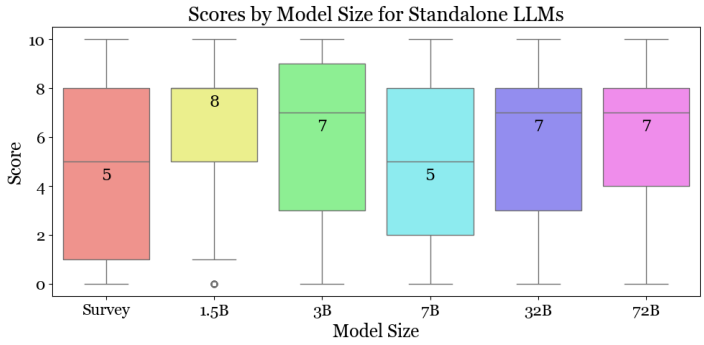

FIGURE 6. The scores from our standalone LLM experiments analyzed by model size, with median values labeled.

result in a significant reduction in MAE compared to the agentic framework, with zero-shot prompting yielding the largest decrease in MAE. We do not observe a statistically significant decrease in MAE as model size increases, further suggesting that larger models do not necessarily lead to more accurate predictions. This indicates that the choice of prompting strategy has a more pronounced effect on performance than model size.

4.3 LLM-Based Agent Behavior

Figure 5 illustrates the distribution of scores by model size for our agentic experiments, while Figure 6 presents the distribution for standalone LLM experiments. The distribution of scores across all model sizes, categorized by prompt type, is shown in Figure 7.

The results indicate that standalone LLMs generate scores with lower variability than material selection experts, as indicated by the smaller interquartile range. However, the agentic framework produces the highest variance, surpassing survey

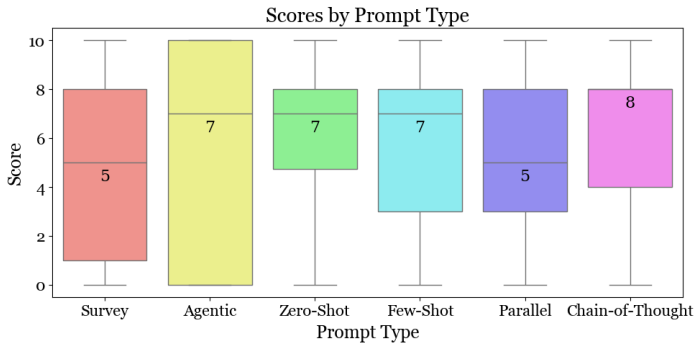

FIGURE 7. The scores from our experiment analyzed by prompt type, with median values labeled.

data and suggesting that LLM-based agents better capture expert decision-making variability. For smaller and mid-sized models (1.5B, 3B, and 7B), agentic AI generally rates materials as more

Copyright © 2025 by ASME

1.5B	3B	7B	14B	32B	72B
6.9%	0.0%	0.0%	94.4%	3.5%	34.7%

TABLE 1. The percentage of responses omitted by our program, categorized by model size, based on whether the final answer tool was called and whether the last agent iteration resulted in an error. Nearly all responses were omitted for the 14B model, while no responses were omitted for the 3B and 7B models.

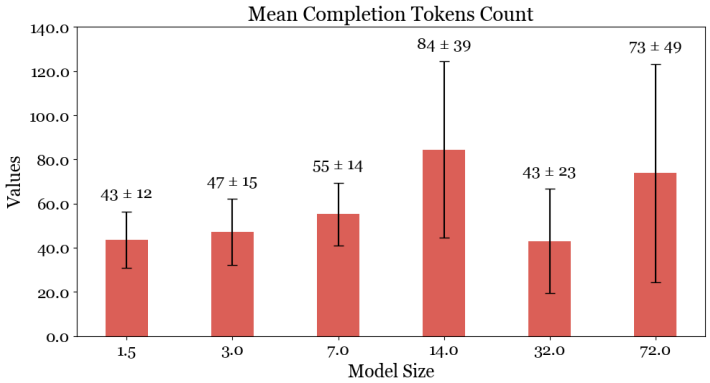

FIGURE 8. The mean number of completion tokens used for each iteration of the agent workflow, categorized by model size, with error bars representing standard deviation. The 14B model required the most tokens, followed by the 72B model, with the 32B model using the fewest tokens.

suitable than standalone LLMs. Across all plots, AI consistently assigns higher suitability scores than human professionals.

Upon reviewing the agent logs, we found that the LLM-powered agent occasionally encountered repeated code parsing errors. After reaching the maximum number of query iterations, it defaulted to returning a response based on the error-handling mechanism rather than the intended reasoning process. Further analysis revealed that the final answer tool was not called, and an error was reported in the last iteration before a response. During development, we modified our code to rerun a prompt up to five times in an attempt to achieve successful execution before omitting it.

We measured prompt omissions, total and successful reruns, total and unique queries, and completion tokens across model sizes. Table 1 shows the percentage of omitted responses, while Figure 8 presents the average completion token count. Summaries of the remaining metrics can be found in Figures 10 and 11 in Appendix A.

Our results revealed non-linear effects of model size on performance. We found that while smaller models (1.5B and 3B) exhibited some omissions and reruns, the 7B model demonstrated the most stability, with no omissions, nearly zero reruns for successful prompts, and a 100% unique query proportion. However, at 14B, performance dropped drastically, with nearly all results being omitted, leading us to exclude the 14B data from further analysis. Larger models (32B and 72B) also exhibited more omissions, increased reruns, and a lower proportion of unique queries compared to the 7B model. The average number of completion tokens initially increased with model size, then dropped at 32B, before rising again at 72B.

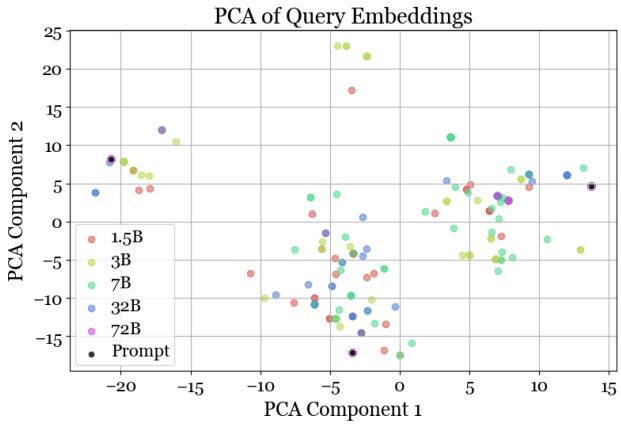

FIGURE 9. The results of the word embeddings analysis after PCA, color-coded based on model size, reveal three main clusters around the system prompt embeddings. This suggests that the models generally generate query responses that align with the system prompt.

In the agent logs, we observed that many of the agent's queries closely followed the predefined query formats in the system prompt, such as *properties of* {*material*}, *necessary properties of* {*design*}, and {*material*} *use in* {*design*}.This finding aligns with our embedding analysis, shown in Figure 9 after applying PCA to reduce the dimensionality. The data points primarily group into three main clusters near the embeddings associated with the example searches provided in the system prompt (which are represented by black dots). This indicates that the models often generate similar terms or structures. However, there is still significant variability around these points, suggesting variability in query construction.

Upon reviewing the queries in each cluster, we find that the clusters correspond to distinct themes: one focuses on material properties related to a criterion, another centers on design-related properties and requirements, and the third examines how materials perform within specific applications and criteria. Furthermore, model sizes tend to cluster differently. For instance, the left cluster includes only prompts from smaller and larger models, while the right cluster is primarily composed of smaller and mid-sized models, suggesting that model size influences the types of queries generated by the LLM-powered agent.

Copyright © 2025 by ASME

4.4 Additional Observations

A review of the agent logs revealed that the 14B and 72B models often provided longer explanations, leading to greater completion token usage. The 14B model typically included its justification within the final response, not fully adhering to the instruction to return only a single integer. This behavior is consistent with its higher omission and rerun rates. In contrast, the 72B model expressed its reasoning through more extensive code when calling the final answer tool, incorporating comments to justify the assigned score. For the remaining models (1.5B, 3B, 7B, and 32B), we observe that as model size increases, the length of generated thoughts also tends to increase, resulting in higher token consumption. Unlike the 14B and 72B models, these models do not exhibit behaviors such as returning invalid final answers or generating longer code.

The average z-score and MAE across prompting strategies and model sizes are presented in Figs. 12 and 13 in Appendix A. We assume that a closer proximity to the survey data indicates greater accuracy, with z-scores and MAEs closer to 0 being ideal. Our z-score analysis reveals a non-linear relationship between model size and alignment with expert decision-making, consistent across all prompting strategies, with the z-score minimized at 7B and highest at 1.5B and 72B.This aligns with prior findings that mid-sized models require fewer reruns and improve query efficiency, as well as our linear regression results, which found no statistically significant effect of model size on the z-score.

When assessed by MAE, parallel prompting with the 72B model performs best. However, zero-shot and parallel prompting with the 3B model yield comparable results, suggesting that smaller models can achieve similar accuracy. The agentic system, despite being the second-best strategy based on z-score, shows the least alignment in terms of MAE. Beyond a certain model size, MAE remains steady, contrasting with the z-score behavior, where performance initially increases and then decreases with size. These differences stem from their calculation methods. The z-score measures direction and magnitude, normalizing for scale, while the MAE focuses solely on the magnitude of the distance between points. As a result, when using z-scores, negative and positive deviations can cancel each other out, indicating that the agent produces results that do not consistently deviate above or below the survey data.

The non-linear trends observed in our study align with previous research showing that performance in instruction-following, math, and coding among Qwen-2.5 models does not scale predictably with size. Specifically, the Qwen-2.5 14B model performed worse on several benchmarks compared to the 7B and 32B models [99]. This behavior may be due to the uneven scaling of architectural factors, such as layers, attention heads, and context length, with model size [89].

5 LIMITATIONS AND FUTURE WORK

While our analysis provides valuable insights into the effects of prompting strategies and model size on LLM performance, several limitations must be considered. Our data was collected exclusively from Qwen 2.5B models, which limits the generalizability of our findings to other model series. Different architectures may exhibit varying performance patterns depending on size and prompt type. Future research should investigate a broader range of models to better understand how these factors interact across diverse LLMs.

Our study also focused solely on zero-shot prompting with an LLM-based agent, leaving the potential effects of other prompting strategies within an agentic framework unexamined. Future research should explore how few-shot, parallel, and chain-of-thought prompting influence an agentic system's ability to replicate expert decision-making in comparison to standalone LLMs.

We equipped our LLM-based agent with a Wikipedia search tool. While Wikipedia provides general information, it is not considered a scholarly resource and has a limited knowledge base, as evidenced by its often uninformative page summaries. In practice, material selection professionals often rely on peer-reviewed journals, industry reports, and past project data. Future research could explore retrieval-augmented generation (RAG) tools for accessing specialized databases or alternative search APIs to enhance result quality.

Additionally, during our experiments, we observed unexplained behavior in our agentic framework, particularly with the 14B model, which frequently encountered errors and failed to execute the final answer tool correctly. Investigating the root causes of these errors and identifying potential mitigation strategies would be valuable. Future research could also explore how variations in prompting techniques and phrasing may influence the frequency and nature of these failures.

We observed slight differences in the behavior of the LLM-based agent, leading to variations in the final scores generated across different runs. Similar fluctuations were noted in standalone LLMs as well. Since our analysis was based on a single run for each method, future work should explore the stability of results across multiple trials to assess the reproducibility and variance of model outputs.

6 CONCLUSION

The agentic framework effectively enables LLMs to execute actions and interact with external environments using tools. Our research demonstrates that this structured approach allows LLMs to break down complex tasks and generate informed responses, facilitating a reasoning process that more closely resembles how a material selection expert might approach decision-making. Our LLM-based agent tends to replicate search queries similar to those provided in the system prompt, emphasizing the impor-

Copyright © 2025 by ASME

tance of strong examples in guiding AI behavior for complex design tasks.

Our findings reveal a non-linear relationship between model size and agentic AI performance, particularly in omission rate, token usage, and alignment with survey data. The 14B model exhibited erratic behavior, frequently failing to follow instructions and consuming the most tokens, while the 7B model had z-scores closest to zero. These findings align with prior research on Qwen-2.5 models, which show that performance in instruction-following, mathematical reasoning, and code generation does not consistently improve with size, likely due to uneven scaling of architectural components. However, there appeared to be a more linear trend between model size and MAE relative to survey data, suggesting that mid-sized LLMs may produce more balanced results without consistent directional bias.

Our linear regression results demonstrated that model size had little impact on alignment to expert decision-making, while the choice of prompting style played a larger role in predicting both z-score and MAE. We also found that the effects of model size within each prompting strategy were minimal. This suggests that prompt engineering has a greater influence than model size in driving accurate agentic behavior across various design applications. By carefully refining prompts, designers can optimize AI's reasoning and decision-making abilities, often achieving superior performance with mid-sized models.

ACKNOWLEDGMENT

We would like to thank the anonymous survey participants in the Autodesk Research Community for their contribution.

REFERENCES

[1] Wiederhorn, S., Fields, R., Low, S., Bahng, G.-W., Wehrstedt, A., Hahn, J., Tomota, Y., Miyata, T., Lin, H., Freeman, B., Aihara, S., Hagihara, Y., and Tagawa, T., 2018. "Mechanical properties". In *Springer Handbook of Materials Measurement Methods*. Springer, pp. 283–397.

[2] William D. Callister, J., and Rethwisch, D. G., 2007. *Materials Science and Engineering: An Introduction*, Vol. 7. John Wiley Sons, Inc., New York.

[3] Giachetti, R., 1997. "Manufacturing process and material selection during conceptual design". In Proceedings of the 6th Industrial Engineering Research Conference.

[4] Clark, G., Kosoris, J., Hong, L. N., and Crul, M., 2009. "Design for sustainability: Current trends in sustainable product design and development". *Sustainability,* ***1***(3), August, pp. 409–424.

[5] Karana, E., and Kesteren, I. V., 2008. "Materials affect: The role of materials in product experience". In *Design and Emotion Moves*, P. Desmet, J. van Erp, and M. Karlsson, eds. Cambridge Scholars Publishing, UK, ch. 11, pp. 221–245.

[6] Farag, M. M., 2020. *Materials and Process Selection for Engineering Design*. CRC Press, Boca Raton.

[7] Bissler, A., and Ghahestani, S., 2018. "Key factors in driving sustainability initiatives in the supply chain: A multiple case study of manufacturing companies". *Journal of Supply Chain Management,* ***45***(2), pp. 78–92.

[8] Alonso, E., Gregory, J., Field, F., and Kirchain, R., 2007. "Material availability and the supply chain: Risks, effects, and responses". *Environmental Science Technology,* ***41***(19), pp. 6649–6656.

[9] Brechet, Y., Bassetti, D., Landru, D., and Salvo, L., 2001. "Challenges in materials and process selection". *Progress in Materials Science,* ***46***(3–4), pp. 407–428.

[10] Jahan, A., Ismail, M., Sapuan, S., and Mustapha, F., 2010. "Material screening and choosing methods – a review". *Materials Design,* ***31***(2), February, pp. 696–705.

[11] Ashby, M., Bréchet, Y., Cebon, D., and c, L. S., 2004. "Selection strategies for materials and processes". *Materials Design,* ***25***(1), February, pp. 51–67.

[12] Chen, K., Fei, C., Bi, Z., Liu, J., Peng, B., Zhang, S., Pan, X., Xu, J., Wang, J., Yin, C. H., Zhang, Y., Feng, P., Wen, Y., Wang, T., Li, M., Ren, J., Niu, Q., Chen, S., Hsieh, W., Yan, L. K. Q., Liang, C. X., Xu, H., Tseng, H.-M., Song, X., and Liu, M., 2024. Deep learning and machine learning – natural language processing: From theory to application. arXiv, October. URL`https://arxiv.org/abs/2411.05026`.

[13] Chiner, M., 1988. "Planning of expert systems for materials selection". *Materials Design,* ***9***(4), July–August, pp. 195–203.

[14] Grandi, D., Jain, Y. P., Groom, A., Cramer, B., and McComb, C., 2024. "Evaluating large language models for material selection". *Journal of Computing and Information Science in Engineering,* ***25***(2), September, pp. 1–33.

[15] Jain, Y. P., Grandi, D., Groom, A., Cramer, B., and McComb, C., 2025. "Mseval: A dataset for material selection in conceptual design to evaluate algorithmic models". *Journal of Mechanical Design,* ***147***(4), April.

[16] Huang, J., and Chang, K. C.-C., 2023. "Towards reasoning in large language models: A survey". *Findings of the Association for Computational Linguistics,* ***ACL 2023***, July, pp. 1049–1065.

[17] Williams, S., and Huckle, J., 2024. Easy problems that llms get wrong. arXiv. URL`https://arxiv.org/abs/2405.19616`.

[18] Wang, Q., Wang, Z., Su, Y., Tong, H., and Song, Y., 2024. "Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?". In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), L.-W. Ku, A. Martins,

Copyright © 2025 by ASME

and V. Srikumar, eds., Association for Computational Linguistics, pp. 6106–6131.

[19] Mukherjee, A., and Chang, H. H., 2025. Agentic ai: Autonomy, accountability, and the algorithmic society. arXiv. URLhttps://arxiv.org/abs/2309.07870.

[20] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., and Gui, T., 2023. The rise and potential of large language model based agents: A survey. arXiv, September. URLhttps://arxiv.org/abs/2309.07864.

[21] Ruan, J., Chen, Y., Zhang, B., Xu, Z., Bao, T., Du, G., Shi, S., Mao, H., Li, Z., Zeng, X., and Zhao, R., 2023. Tptu: Large language model-based ai agents for task planning and tool usage. arXiv, August. URLhttps://arxiv.org/abs/2308.03427.

[22] Zarandi, M. H. F., Mansour, S., Hosseinijou, S. A., and Avazbeigi, M., 2011. "A material selection methodology and expert system for sustainable product design". *The International Journal of Advanced Manufacturing Technology,* ***57***(9–12), May, pp. 885–903.

[23] Deng, Y.-M., and Edwards, K., 2007. "The role of materials identification and selection in engineering design". *Materials Design,* ***28***(1), January, pp. 131–139.

[24] Ashby, M. F., 2010. *Materials Selection in Mechanical Design (Fourth Edition)*. Butterworth-Heinemann.

[25] Rahim, A. A., Musa, S. N., Ramesh, S., and Lim, M. K., 2020. "A systematic review on material selection methods". *Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications,* ***234***(7), pp. 1032–1059.

[26] Amen, R., and Vomacka, P., 2001. "Case-based reasoning as a tool for materials selection". *Materials Design,* ***22***(5), pp. 353–358.

[27] Emovon, I., and Oghenenyerovwho, O. S., 2920. "Application of mcdm method in material selection for optimal design: A review". *Results in Materials,* ***7***, October.

[28] Karandikar, H. M., and Mistree, F., 1992. "An approach for concurrent and integrated material selection and dimensional synthesis". *Journal of Mechanical Design,* ***114***(4), December, pp. 633–641.

[29] Rao, R. V., 2008. "A decision making methodology for material selection using an improved compromise ranking method". *Materials Design,* ***29***(10), December, pp. 1949–1954.

[30] Chandrasekhar, A., Sridhara, S., and Suresh, K., 2022. "Integrating material selection with design optimization via neural networks". *Engineering with Computers,* ***38***(2), September, pp. 4715–4730.

[31] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R., 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

[32] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R., 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems, Feb. arXiv:1905.00537 [cs].

[33] Lin, Y.-T., and Chen, Y.-N., 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models, May. arXiv:2305.13711 [cs].

[34] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J., 2021. Measuring Massive Multitask Language Understanding, Jan. arXiv:2009.03300 [cs].

[35] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y., 2023. Holistic Evaluation of Language Models, Oct. arXiv:2211.09110 [cs].

[36] Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Lei, J., Fu, Y., Sun, M., and He, J., 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models, Nov. arXiv:2305.08322 [cs].

[37] Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N., 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models, Sept. arXiv:2304.06364 [cs].

[38] Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B., 2024. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback, Jan. arXiv:2305.14387 [cs].

[39] Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I., 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, Mar. arXiv:2403.04132 [cs].

[40] Richardson, M., Burges, C. J., and Renshaw, E., 2013. "Mctest: A challenge dataset for the open-domain machine comprehension of text". In Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 193–203.

[41] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P., 2016. "Squad: 100,000+ questions for machine comprehension of

Copyright © 2025 by ASME

text". *arXiv preprint arXiv:1606.05250*.

[42] Yang, Y., Yih, W.-t., and Meek, C., 2015. "Wikiqa: A challenge dataset for open-domain question answering". In Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 2013–2018.

[43] Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N. A., and Gardner, M., 2021. "A dataset of information-seeking questions and answers anchored in research papers". *arXiv preprint arXiv:2105.03011*.

[44] Shaham, U., Ivgi, M., Efrat, A., Berant, J., and Levy, O., 2023. "Zeroscrolls: A zero-shot benchmark for long text understanding". *arXiv preprint arXiv:2305.14196*.

[45] Doris, A. C., Grandi, D., Tomich, R., Alam, M. F., Ataei, M., Cheong, H., and Ahmed, F., 2025. "Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation". *Journal of Computing and Information Science in Engineering,* ***25***(2), p. 021009.

[46] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al., 2023. "Mme: A comprehensive evaluation benchmark for multimodal large language models". *arXiv preprint arXiv:2306.13394*.

[47] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al., 2023. "Mmbench: Is your multi-modal model an all-around player?". *arXiv preprint arXiv:2307.06281*.

[48] Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., and Chang, M.-W., 2023. "Can pre-trained vision and language models answer visual information-seeking questions?". *arXiv preprint arXiv:2302.11713*.

[49] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A., 2022. "Learn to explain: Multimodal reasoning via thought chains for science question answering". *Advances in Neural Information Processing Systems,* ***35***, pp. 2507–2521.

[50] Ma, K., Grandi, D., McComb, C., and Goucher-Lambert, K., 2023. "Conceptual design generation using large language models". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 116838, American Society of Mechanical Engineers, p. V006T06A021.

[51] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y., 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4, Apr. arXiv:2303.12712 [cs].

[52] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P., 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, Nov. arXiv:2302.04023 [cs].

[53] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D., 2024. Can Large Language Models Transform Computational Social Science?, Feb. arXiv:2305.03514 [cs].

[54] Brewka, G., 1996. "Artificial intelligence—a modern approach by stuart russell and peter norvig, prentice hall. series in artificial intelligence, englewood cliffs, nj.". *The Knowledge Engineering Review,* ***11***(1), pp. 78–79.

[55] Wooldridge, M., and Jennings, N. R., 1995. "Intelligent agents: Theory and practice". *The knowledge engineering review,* ***10***(2), pp. 115–152.

[56] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al., 2025. "The rise and potential of large language model based agents: A survey". *Science China Information Sciences,* ***68***(2), p. 121101.

[57] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S., 2023. "Generative agents: Interactive simulacra of human behavior". In Proceedings of the 36th annual acm symposium on user interface software and technology, pp. 1–22.

[58] Liu, R., Yang, R., Jia, C., Zhang, G., Zhou, D., Dai, A. M., Yang, D., and Vosoughi, S., 2023. "Training socially aligned language models on simulated social interactions". *arXiv preprint arXiv:2305.16960*.

[59] Sumers, T., Yao, S., Narasimhan, K., and Griffiths, T., 2023. "Cognitive architectures for language agents". *Transactions on Machine Learning Research*.

[60] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al., 2024. "A survey on large language model based autonomous agents". *Frontiers of Computer Science,* ***18***(6), p. 186345.

[61] Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Zhou, X., Huang, Y., Xiao, C., et al., 2024. "Tool learning with foundation models". *ACM Computing Surveys,* ***57***(4), pp. 1–40.

[62] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al., 2023. "Augmented language models: a survey". *arXiv preprint arXiv:2302.07842*.

[63] Ren, R., Wang, Y., Qu, Y., Zhao, W. X., Liu, J., Tian, H., Wu, H., Wen, J.-R., and Wang, H., 2023. "Investigating the factual knowledge boundary of large language models with retrieval augmentation". *arXiv preprint arXiv:2307.11019*.

[64] Liu, B., Jiang, Y., Zhang, X., Liu, Q., Zhang, S., Biswas, J., and Stone, P., 2023. "Llm+ p: Empowering large language models with optimal planning proficiency". *arXiv preprint arXiv:2304.11477*.

[65] Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J. B., Shu, T., and Gan, C., 2023. "Building cooperative embodied agents modularly with large language models". *arXiv preprint arXiv:2307.02485*.

[66] Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu,

Copyright © 2025 by ASME

Z., and Sun, M., 2023. "Communicative agents for software development". *arXiv preprint arXiv:2307.07924,* ***6***(3).

[67] Mehta, N., Teruel, M., Sanz, P. F., Deng, X., Awadallah, A. H., and Kiseleva, J., 2023. "Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback". *arXiv preprint arXiv:2304.10750*.

[68] Aher, G. V., Arriaga, R. I., and Kalai, A. T., 2023. "Using large language models to simulate multiple humans and replicate human subject studies". In International Conference on Machine Learning, PMLR, pp. 337–371.

[69] Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al., 2023. "Agentbench: Evaluating llms as agents". *arXiv preprint arXiv:2308.03688*.

[70] Lin, J., Tomlin, N., Andreas, J., and Eisner, J., 2024. "Decision-oriented dialogue for human-ai collaboration". *Transactions of the Association for Computational Linguistics,* ***12***, pp. 892–911.

[71] Liang, Y., Zhu, L., and Yang, Y., 2023. "Tachikuma: understading complex interactions with multi-character and novel objects by large language models". *arXiv preprint arXiv:2307.12573*.

[72] Xu, B., Liu, X., Shen, H., Han, Z., Li, Y., Yue, M., Peng, Z., Liu, Y., Yao, Z., and Xu, D., 2023. "Gentopia: A collaborative platform for tool-augmented llms". *arXiv preprint arXiv:2308.04030*.

[73] Sciammarella, C. A., and Sciammarella, F. M., 2012. *Experimental Mechanics of Solids*. John Wiley & Sons, Mar. Google-Books-ID: aVP6tpNXAf4C.

[74] Jenkins, C., and Khanna, S., 2005. *Mechanics of Materials: A Modern Integration of Mechanics and Materials in Structural Design*. Academic Press, Mar. Google-Books-ID: fVZjtfpbE9wC.

[75] Zaki, M., Jayadeva, Mausam, and Krishnan, N. M. A., 2024. "MaScQA: investigating materials science knowledge of large language models". *Digital Discovery,* ***3***(2), Feb., pp. 313–327. Publisher: RSC.

[76] Jin, H., Zhang, E., and Espinosa, H. D., 2023. "Recent Advances and Applications of Machine Learning in Experimental Solid Mechanics: A Review". *Applied Mechanics Reviews,* ***75***(6), Nov., p. 061001. arXiv:2303.07647 [cs].

[77] Hu, Y., and Buehler, M. J., 2023. "Deep language models for interpretative and predictive materials science". *APL Machine Learning,* ***1***(1), Feb., p. 010901.

[78] Luu, R. K., and Buehler, M. J., 2024. "BioinspiredLLM: Conversational Large Language Model for the Mechanics of Biological and Bio-Inspired Materials". *Advanced Science,* ***11***(10), p. 2306724. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202306724.

[79] Lei, G., Docherty, R., and Cooper, S. J., 2024. "Materials science in the era of large language models: a perspective". *Digital Discovery,* ***3***(7), July, pp. 1257–1272. Publisher: RSC.

[80] Picard, C., Edwards, K. M., Doris, A. C., Man, B., Giannone, G., Alam, M. F., and Ahmed, F., 2024. From Concept to Manufacturing: Evaluating Vision-Language Models for Engineering Design, Dec. arXiv:2311.12668 [cs].

[81] Mustapha, K. B., 2025. "A survey of emerging applications of large language models for problems in mechanics, product design, and manufacturing". *Advanced Engineering Informatics,* ***64***, Mar., p. 103066.

[82] Buehler, M. J., 2023. "MeLM, a generative pretrained language modeling framework that solves forward and inverse mechanics problems". *Journal of the Mechanics and Physics of Solids,* ***181***, Dec., p. 105454.

[83] Buehler, M. J., 2024. "MechGPT, a Language-Based Strategy for Mechanics and Materials Modeling That Connects Knowledge Across Scales, Disciplines, and Modalities". *Applied Mechanics Reviews,* ***76***(021001), Jan.

[84] Ni, B., and Buehler, M. J., 2023. MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge, Nov. arXiv:2311.08166 [cs].

[85] Ghafarollahi, A., and Buehler, M. J., 2024. Rapid and Automated Alloy Design with Graph Neural Network-Powered LLM-Driven Multi-Agent Systems, Oct. arXiv:2410.13768 [cond-mat].

[86] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y., 2022. React: Synergizing reasoning and acting in language models. arXiv, October. URL`https://arxiv.org/abs/2210.03629`.

[87] Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., and Ji, H., 2024. "Executable code actions elicit better llm agents". URL`https://arxiv.org/abs/2402.01030`.

[88] Zhou, W., Jiang, Y. E., Li, L., Wu, J., Wang, T., Qiu, S., Zhang, J., Chen, J., Wu, R., Wang, S., Zhu, S., Chen, J., Zhang, W., Tang, X., Zhang, N., Chen, H., Cui, P., and Sachan, M., 2023. Agents: An open-source framework for autonomous language agents. arXiv, September. URL`https://arxiv.org/abs/2309.07870`.

[89] Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z., 2025. "Qwen2.5 technical report". URL`https://arxiv.org/abs/2412.15115`.

[90] Marvin, G., Hellen, N., Jjingo, D., and Nakatumba-Nabende, J., 2024. "Prompt engineering in large language models". In Data Intelligence and Cognitive Informatics,

Copyright © 2025 by ASME

I. J. Jacob, S. Piramuthu, and P. Falkowski-Gilski, eds., Algorithms for Intelligent Systems, Springer Nature, Springer Singapore, pp. 387–402.

[91] Jalham, I. S., 2006. "Decision-making integrated information technology (iit) approach for material selection". *International Journal of Computer Applications in Technology,* ***25***(1), January, pp. 65–71.

[92] Sinhababu, N., Parry, A., Ganguly, D., Samanta, D., and Mitra, P., 2024. "Few-shot prompting for pairwise ranking: An effective non-parametric retrieval model". URL`https://arxiv.org/abs/2409.17745`.

[93] Chen, H. M., 2024. "Parallel prompt decoding: A cost-effective and memory-efficient approach for accelerating llm inference". MS Thesis, Imperial College London, London, England.

[94] Rawte, V., Sheth, A., and Das, A., 2023. "A survey of hallucination in large foundation models". URL`https://arxiv.org/abs/2309.05922`.

[95] Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E., Le, Q. V., and Zhou, D., 2022. "Chain-of-thought prompting elicits reasoning in large language models". In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2022), S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., Advances in Neural Information Processing Systems, Neural Information Processing Systems (NeurIPS), Curran Associates Inc., pp. 24824–24837.

[96] Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013. "Efficient estimation of word representations in vector space". In Proceedings of the International Conference on Learning Representations (ICLR 2013).

[97] Senel, L. K., Utlu, I., Yucesoy, V., Koc, A., and Cukur, T., 2018. "Semantic structure and interpretability of word embeddings". *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* ***26***(10), October, pp. 1769–1779.

[98] Raunak, V., Gupta, V., and Metze, F., 2019. "Effective dimensionality reduction for word embeddings". In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, and M. Rei, eds., Association for Computational Linguistics, pp. 235–243.

[99] Hui, T., Zhao, L., Dong, G., Zhang, Y., Zhou, H., and Su, S., 2024. Smaller language models are better instruction evolvers. arXiv. URL`https://arxiv.org/abs/2412.11231`.

Appendix A: Additional Analysis

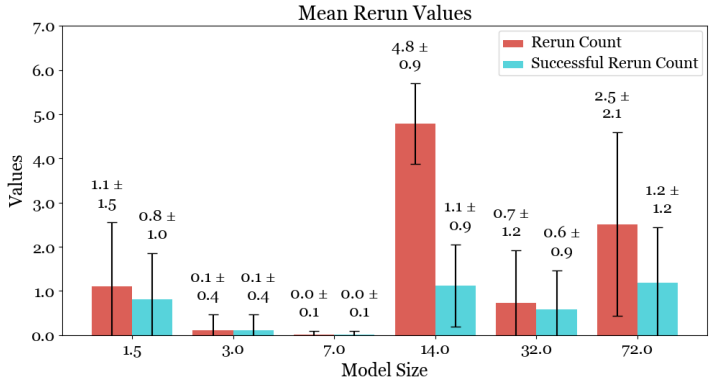

FIGURE 10. The mean number of reruns for all prompts and successful prompts, categorized by model size. Both the overall mean and the average number of reruns for successful prompts are lowest for the 7B model.

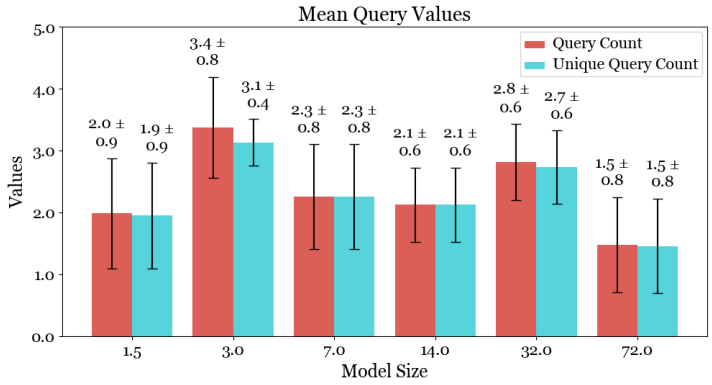

FIGURE 11. The mean number of queries and unique queries for successful prompts, broken down by model size. All queries were unique for the 7B and 14B models, whereas the 1.5B, 32B, and 72B models had some repeated queries.

Copyright © 2025 by ASME

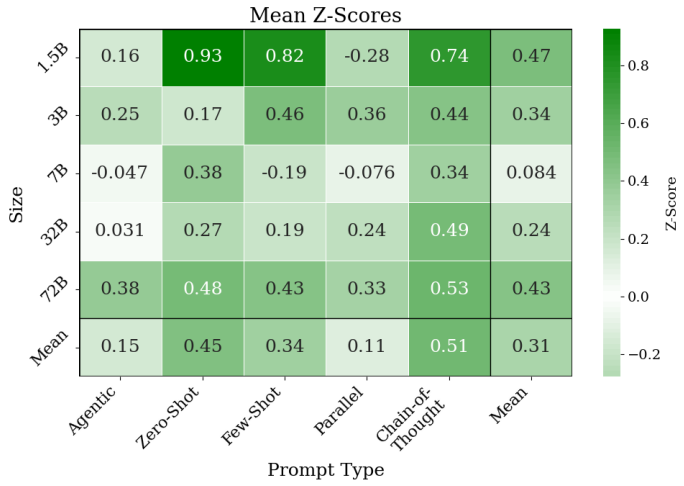

FIGURE 12. The mean z-scores from experiments across all prompting strategies and model sizes. On average, the lowest z-score is observed for the 7B and 32B agentic models, while the highest z-score is seen for the 1.5B model using zero-shot and few-shot prompting.

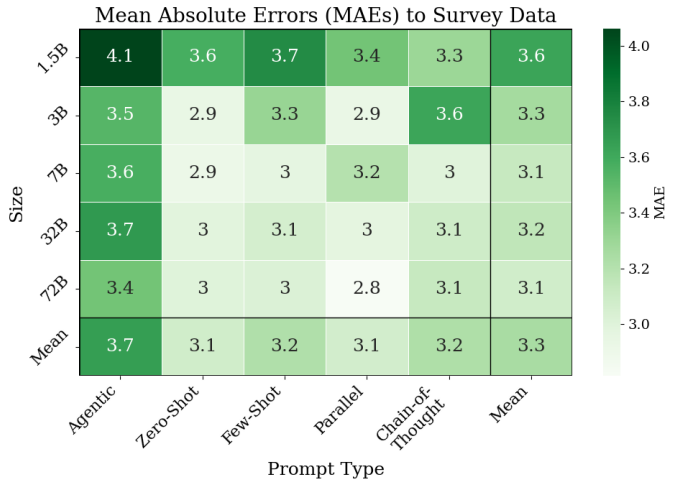

FIGURE 13. The MAEs to survey data from experiments across all prompting strategies and model sizes. The greatest error is observed for the 1.5B agentic model, while the least is recorded for the 72B model using parallel prompting.

Copyright © 2025 by ASME