LLMs can Design Sustainable Concrete –a Systematic Benchmark

Preprint · January 2024

DOI: 10.13140/RG.2.2.33795.27686

Christoph Völker
Bundesanstalt für Materialforschung und -prüfung
33 PUBLICATIONS **167** CITATIONS
SEE PROFILE

Sabine Kruschwitz
Bundesanstalt für Materialforschung und -prüfung
93 PUBLICATIONS **1,201** CITATIONS
SEE PROFILE

LLMs can Design Sustainable Concrete – a Systematic Benchmark

Christoph Völker*[a], Tehseen Rug[b] , Kevin Maik Jablonka[c] and Sabine Kruschwitz[a,d]

Received 26th January 2024,

In the context of a circular building material economy, the complexity of resource flows and the variability of material composition pose significant challenges. This study demonstrates how Large Language Models (LLMs) can advance material design by adopting a Knowledge-Driven Design (KDD) approach that outperforms traditional Data-Driven Design (DDD) methods. Our focus is on designing alkali-activated concrete (AAC) mix designs, an environmentally friendly alternative to conventional Portland cement-based concrete. GPT-3.5 Turbo and GPT-4 Turbo enable using fuzzy design knowledge as previously untapped input data modality. A key aspect of our research is to improve the performance of the LLMs in post-training. We use strategies such as refining prompt context, extending test time, and including a verifier.The study's systematic benchmarks are based on 240 AAC formulations extracted from the literature. The target was on achieving maximum compressive strength through an adaptive design approach over multiple development cycles. We compare these results to the traditional DDD baseline methods. KDD outperforms conventional methods by providing robust initial predictions and demonstrating effective adaptability informed by laboratory validation data, culminating in the development of high-quality AAC formulations. These results provide valuable insight into the capabilities of chat-based LLMs in managing complex material formulations, which are particularly beneficial in situations where traditional DDD is impractical due to data collection issues. With natural language as the basis the KDD is intuitively accessible to domain experts. The methodology and results of this study have significant implications for the field of materials science, particularly in the context of a circular economy, and pave the way for innovative applications of LLMs in various scientific fields.

1. Introduction

The environmental impact of traditional concrete production has become a pressing issue, highlighting the need for increased circularity and significant CO_2 reduction [1]. The development of new cementitious binders low in calcium offers a promising solution, potentially reducing carbon dioxide emissions by 40-80 percent while retaining structural properties comparable to conventional cement [2]. However, the development of alternative materials such as alkali-activated concretes (AAC) or geopolymers, presents unique challenges.

AACs stand out due to their ability to be synthesized from a wide array of aluminosilicate feedstocks and various compositions of activator solutions, offering various opportunities for customization of properties to meet specific needs. The use of diverse, often heterogeneous waste streams and secondary raw materials as feedstocks for AAC leads to a wide range of possible compositions [3, 4, 5]. Yet this diversity in composition introduces significant uncertainties, complicating the process of prescriptive material design. This complexity necessitates innovative solutions to standardize AAC production, ensuring both environmental benefits and structural reliability.

Traditional development methods for cementitious materials, such as AAC, are often inefficient and limited due to their empirical and prescriptive nature, struggling to effectively handle the material's extensive range of compositional variations [6]. The introduction of Data-Driven Design (DDD) methods, such as Sequential Learning (SL) and Bayesian Optimization (BO) marked a significant improvement [7, 8, 9, 10, 11], yet they are dependent on initial data collection: The high variability of the precursor materials, where each batch can have different properties, means that it is simply not practical to use pre-existing data. Consequently, this approach often necessitates a preliminary phase of re-establishing fundamental relationships through experimental data, delaying the onset of novel formulation development.

Foundational Large Language Models (LLMs) are emerging as an alternative solution. While traditional data-driven approaches often struggle with the dynamic and complex nature of AACs, LLMs, with their extensive knowledge base, offer a way to overcome these limitations by adaptively tailoring predictions to evolving contexts [12]. A key enabler in this process is In-Context Learning (ICL) within LLMs [13]. This capability allows LLMs to effectively interpret and apply fuzzy rules and instructions. Crucially, this enables navigation of inherently ambiguous material designs by using domain knowledge, expressed in natural language, to guide design predictions. This approach introduces a novel modality of input data that is sharp in contrast to traditional methods that rely on explicitly measured data. Unlike the fixed and explicit nature of traditional data, which directly describes the materials tested, knowledge in the context of LLMs can be implicit. This implicit nature allows for a more flexible adaptation to new and diverse material compositions, through known but uncertain relationships between volatile resources.

Furthermore, the ability of LLMs to make zero-shot predictions is transformative. It eliminates the need to collect extensive initial training data—a significant bottleneck in traditional

[a] Bundesanstalt für Materialforschung und -prüfung, Unter den Eichen 87, Berlin, Germany
[b] Iteratec GmbH, St.-Martin-Str. 114, 81669 Munich, Germany
[c] Helmholtz Institute for Polymers in Energy Applications, Jena, Germany
[d] Technische Universität Berlin, Institute of Civil Engineering, Non-destructive building material testing, Gustav-Meyer-Allee 25, 13355 Berlin, Germany

*Corresponding Author (Christoph.Voelker@bam.de)

methods. In practice, this could significantly optimize the use of resources in laboratories and streamline the process of developing new materials. By integrating these advanced capabilities, LLMs not only provide a solution to the current challenges in materials science but also open the door to innovative design possibilities that were previously unattainable.

As the construction industry moves towards sustainable alternatives to traditional cement, effectively managing the complexities of materials such as AAC will be critical for their wider adoption. Shortening development cycles is key to making AACs more attractive to the market. However, it is equally important to ensure its safety and reliability in engineering applications is equally paramount. Our study addresses this dual need by proposing the use of systematic benchmarks to evaluate the effectiveness of AAC design. This methodology facilitates rigorous statistical evaluation, ensuring that technological advances in AAC are consistent with the stringent safety and reliability standards required in materials design. The findings and methodologies presented in this study, centred on AAC, have considerable potential for application to different classes of materials facing similar challenges.

1.1 Novelty and Scope

This study presents KDD, a novel approach to material design using LLMs. KDD is characterised by its virtual independence from traditional training data, which represents a breakthrough in material development with variable and volatile starting materials. The study focuses on the design of AAC, demonstrating not only a practical use case of KDD, but also providing a comprehensive development framework for researchers wishing to adopt this methodology for other material types.

A key innovation of our KDD approach is its use of fuzzy domain knowledge and lab feedback expressed in natural language, a largely untapped data modality in AI-driven material design. This utilisation of natural language simplifies complex AI processes, making KDD more accessible to a diverse range of researchers and practitioners.

Moreover, through our systematic benchmarking framework, we identify key factors that influence the performance of LLMs in material design, offering insights into how these models' performance can be improved in post-training. This process showcases the adaptability and efficiency of LLMs, not only in generating initial design predictions but also in refining these predictions based on empirical lab feedback.

The scope of our research extends to a wide audience, targeting stakeholders in construction chemistry as well as researchers exploring solutions for volatile precursor materials in a circular economy. Our findings offer a novel perspective in utilizing AI for sustainable material development, contributing to the broader goal of environmental sustainability while ensuring safety and reliability in engineering applications.

In summary, our research highlights the potential of AI, particularly LLMs, in transforming material design processes. By integrating fuzzy domain knowledge and practical lab feedback into our KDD approach, we pave the way for a more intuitive and accessible application of AI in material science, setting a benchmark for future research and implementation in this dynamic field.

2. Points of departure

This chapter provides an overview of the existing literature and previous research, outlines the research gap, and aims, hypotheses and research questions explored in this contribution.

2.1 Literature Research and Knowledge Gap

LLMs such as GPT-3.5 and GPT-4, based on the Transformer architecture, have shown remarkable ability to perform various 'downstream tasks' without the need for specialized training [12]. These models efficiently manage diverse tasks by exploiting their extensive pre-training on diverse datasets, which provides them with a comprehensive understanding of language and context.

Boyko et al. [16] have gathered insights from different scientific disciplines in their interdisciplinary view of LLMs in scientific research. Their review covers a range of applications, including creative brainstorming, information evaluation, programming, data analysis, and writing. They conclude that LLMs are driving a new era in research by enabling the efficient processing of large and heterogenous information sources.

Similarly, Microsoft Research [17] investigated the capabilities of GPT-4 in five scientific domains, including materials design and computational chemistry. The study is based on a wide range of domain-specific tasks that were evaluated by the authors. Specifically, complex scientific knowledge is queried, and GPT-4's responses are evaluated by subject matter experts. The authors conclude that GPT-4 could be a promising tool for materials science, although there is still room for improvement, for example through more domain-specific training data.

Further research confirms these capabilities. A study by Jablonka et al. [18] demonstrates that LLMs can outperform traditional machine learning models in predicting molecular material properties, and chemical reactions, especially in scenarios with limited data. The authors tested a wide range of hyperparameters giving a thorough insight into the effects of finding the right LLM setting. Another study by Ramos et al. [19] shows the success of LLMs in predicting the solubility of drug-like molecules and reaction yields, including uncertainty estimates, using ICL, which performed on par with conventional models.

Jablonka et al. [20] use 14 examples to show potential applications in materials research - going well beyond a prompt-answer scheme. These examples, created during a one-day hackathon, demonstrate the rapid innovation and novel solutions that also result from integrating LLMs into software solutions in areas such as predictive modelling, workflow automation, knowledge extraction and education. Projects such as configuring the Materials Project API (MAPI-LLM), developing a Bayesian optimization (BOLLAMA) experimental setup, and constructing scientific knowledge graphs (InsightGraph) illustrate this. Additionally, LLMs have demonstrated their ability to make quantitative predictions, such as forecasting the compressive strength of AAC (Text2Concrete). Here, the authors show an improvement in performance through the incorporation of fuzzy design rules (e.g. higher water content reduces strength), outperforming even a state-of-the-art prediction model, the M5-Random Forest [21].

Research by Boiko et al. [22]and Bran et al. [23] suggests that LLMs could potentially perform research autonomously by integrating multiple tools and incorporating long-term planning and task execution. These systems can utilize a range of expert-designed tools, from web and literature searches to specific molecular and reaction utilities, and even operate laboratory hardware, improving performance in chemistry-related tasks. Their impact on materials research represents a significant shift, as LLMs are poised to transform much of the scientific workflow.

This research has demonstrated that LLMs can not only make accurate predictions in the field of materials science but can be used to build knowledge-generating systems that mirror an AI-driven scientist. The excitement surrounding the current paradigm shift in scientific research is rooted in the newfound ability to automate tasks that traditionally required deep domain expertise. This expertise, previously hidden behind institutional and educational barriers, is now not only more readily available, but also much less expensive - with the added benefit of being seamlessly embedded in software. Importantly, these examples go beyond technical skills, such as mastering API syntax, and include the ability to reason about a given topic, and on the basis of new information.

As the capabilities of LLMs in scientific contexts become more apparent, it has also become clear that further development does not depend solely on scaling LLMs during training (as in [24, 25, 26]). Jones' research [27] indicates that investing in Test Time (TT) computation – allowing a fully trained model to generate a wider range of solutions – has proved remarkably effective: in the board game "Go" equal performance was achieved either by training a deep learning model for longer or by generating more solutions during TT. Cobe et al. [28] applied this approach to LLMs for solving mathematical text problems, showing that a smaller model using a TT strategy outperformed a 30 times larger model fine-tuned on task-specific data. Subsequently, Lightman et al. [29] highlighted the potential gains from more effective verifiers (the models that evaluate the TT-generated solutions). They found significant performance gains when TT was combined with an appropriate verifier strategy. Davidson et al. [30] review methods for post-training performance improvement and show that in some cases a factor of 20 in performance improvement is possible at a fraction of the cost. The authors noted that in addition to improving specific LLM skills, e.g. through tool integration or fine-tuning, more general approaches are also highly effective. These include improving prompts (e.g. by using few-shot examples) and generating and voting on multiple solutions.

This leads to the key knowledge gap that this study aims to address:

Is the effectiveness of current chat models, such as GPT-3.5 and GPT-4, sufficient to position them as viable alternatives to established tools in the field of Alkali-Activated Concrete (AAC) formulation? Furthermore, what key factors drive their performance, particularly in post-training applications?

This study seeks to explore this gap by implementing and evaluating different computational strategies and configurations in LLMs, with the aim of achieving a new level of efficiency and effectiveness in AAC design. Addressing this gap is crucial for advancing the application of LLMs in materials science and potentially transforming the way scientific research is conducted in this domain.

2.2 Hypothesis and Research Questions

This study hypothesizes that the strategic application of LLMs in designing eco-friendly building materials, exemplified by AAC, can optimize formulations by utilizing existing knowledge, thereby minimizing the need for extensive training data. This optimization approach is expected to save resources in initial data collection and improve the discovery of novel material properties and relationships. To investigate this hypothesis, the study is structured around several key research questions and methodologies:

1) **Performance Evaluation**: How does KDD compare in efficiency and effectiveness against traditional DDD baseline methods in AAC design?
2) **Contextual Quality**: How does the level of specificity and detail in the context provided to LLMs affect the effectiveness of the AAC design process?
3) **Standard Feedback Design Loop (SFDL) vs. Testing and Verification Design Loop (TVDL)**: How do outcomes of a standard implementation compare against an extended TT strategy?
4) **Model Comparison**: What are the differential impacts of different LLM versions (GPT 3.5-Turbo vs. GPT 4-Turbo [15]) on AAC design outcomes?

The first research question investigates the capabilities of KDD as an alternative to traditional DDD methods in the design of AACs. This exploration is highly consequential in that KDD's effective use of natural language (as opposed to vector representations in DDD) could significantly simplify the integration of inductive biases. This opens the possibility of seamlessly incorporating fuzzy concepts such as intentions or fuzzy design rules into material design, providing researchers and practitioners with an entirely new, versatile set of tools. This has the potential to unlock numerous innovative applications with far-reaching impacts.

The following questions look in more detail at the specific requirements for applying KDD. We examine how the quality and level of detail of the context provided to LLMs affects the effectiveness of the design process. Furthermore, we compare different implementations of design loops - SFDL and TVDL - to evaluate how extended TT strategies perform in comparison to standard methods. Finally, we contrast the performance of different LLM generations (GPT 3.5-Turbo vs. GPT 4-Turbo) in AAC design to contrast the insights gained on the suitability of higher efficiency through model scaling.

Essentially, these research questions collectively aim to provide a comprehensive understanding of the potential of KDD in fundamentally reshaping material design. We provide insight to the questions: can KDD set a new benchmark in this field and what future advances in AI-driven materials science are conceivable?

3. Methodology

This chapter delineates the methodology employed in our study, structured around the following three key components. First, we introduce a chat-based interface for material design, using ICL to tailor LLM responses to specific AAC design scenarios. Second, we provide a brief overview of established DDD methodologies, setting the stage for a comparative analysis. We detail our approach for assessing LLM performance in comparison to conventional DDD approaches, with a particular emphasis on the evaluation criteria and benchmarks used. Thirdly, we present the AAC formulation design task and the dataset used to validate the results, outlining the criteria for selecting the dataset and the parameters for measuring success.

3.1 Implementing the Materials Design with a Chat Model

The chat-based approach of this study aims to optimize Alkali-Activated Concrete (AAC) formulations for maximum 28-day compressive strength through an iterative design loop. In this process, the chat model generates formulation predictions through ICL and refines them based on user feedback. The following sections will first explore the basic workflows and roles of prompting, and then delve into the specific prompt design strategies employed in this work.

Workflows and Roles

In this study, two different reasoning strategies of the chat-based material design are implemented to optimize AAC formulation.

The SFDL is the most basic implementation, consisting of a feedback loop between a user and a Design Assistant (DA) chat model, as shown in Figure 1. In this approach, the DA is tasked with generating material designs that meet specified requirements, a method known as inverse design. This process involves suggesting designs to achieve predetermined properties, rather than deriving properties from an existing design [31]. The DA recommendations are then validated through laboratory experiments. The results of these validations are fed back to the DA to refine and improve its future material design proposals based on empirical evidence.

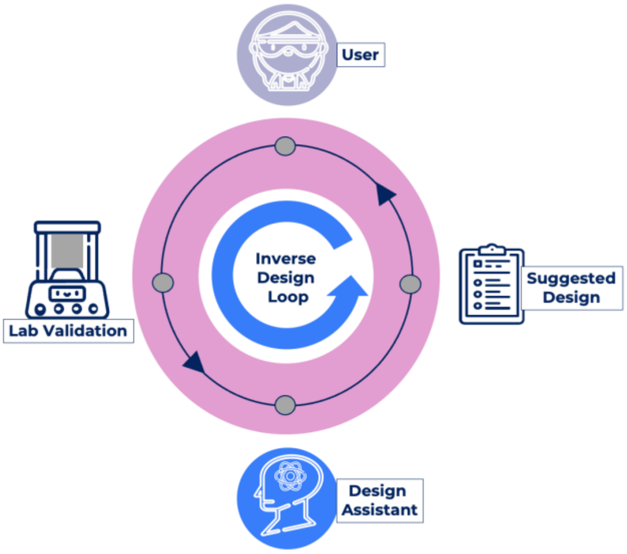


Figure 1: Standard Feedback Design (SFDL) workflow: Feedback loop between the user's lab validation data and the Design Assistant's suggested design.

The second variant shown in Figure 2, the TVDL, uses a TT-verifier workflow as proposed by Cobe et al. [28] and Lightman et al. [29]. This workflow extends the SFDL by generating additional test designs and a VM that scores each proposal in a second forward prediction loop.

The fundamental difference between the DA and the VM lies in the direction of their predictions. The DA generates multiple design parameters, such as weight proportions and processing steps, inversely, i.e. assuming that they will collectively meet the specified design requirements. In contrast, the VM makes a forward prediction of the expected compressive strength for the proposed designs. This estimate forms the basis for ranking the different formulations according to how well their estimated strengths match with the target strength. The most promising formulation, determined by the highest predicted strength, is then selected for further empirical validation by the user.

This score-based system is a significant departure from other frameworks such as Reflexion [32], Resolving Agents [33], or Chain of Thought [34] methods, which our preliminary studies found to be less effective. In these frameworks, chat models generate arguments for and against initial designs, often leading to random and selective “improvements”. In contrast, scalar scoring by the VM in the TVDL provides a more definitive and effective method for identifying the most promising designs.

A key difference between this work and the implementations by Cobe et al. and Lightman et al. is the non-fine-tuned nature of the VM. GPT-3.5-Turbo was found to be sufficiently effective for this application even without specific fine-tuning. The straightforward structure of both the SFDL and TVDL, together with their generic components, makes these approaches easily adaptable to other domains, demonstrating their versatility and potential for broader applications.

In summary, the SFDL implementation is based on a simple prompt-answer format, whereas the TVDL implementation uses a multi-stage process with additional test designs and a verifier model. Depending on the variant there are two or three main roles, respectively, which are illustrated in Figure 3 and described below.

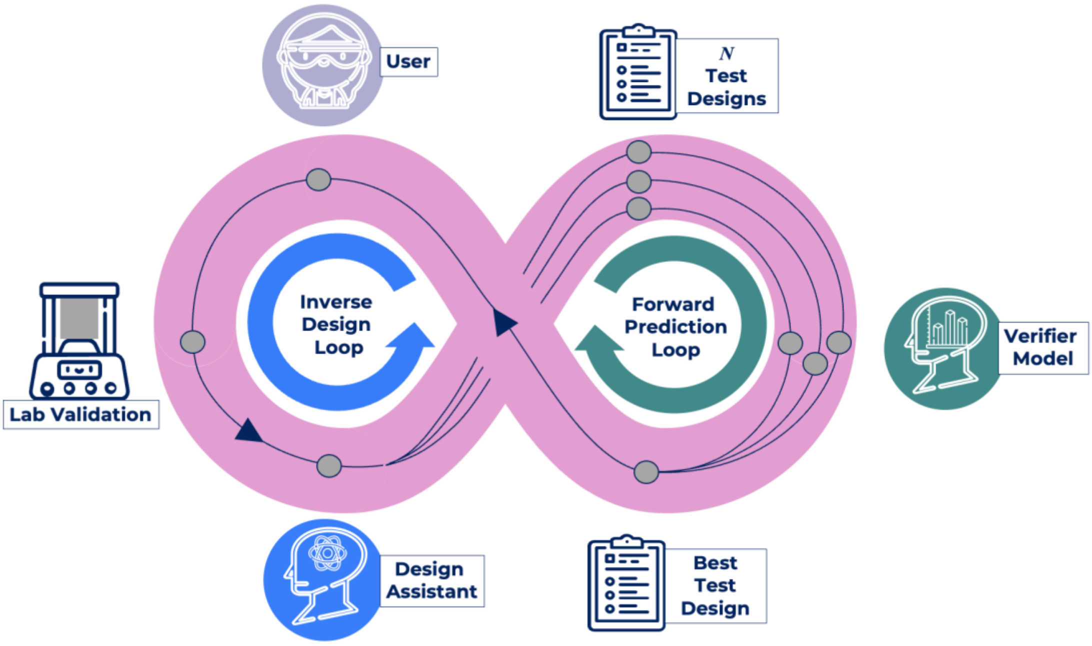


Figure 2: Test and Verification Design Loop (TVDL): Inverse design loop and forward prediction loop. Implementation with test time equal to three. Task instructions and design knowledge are provided to the DA and the VM respectively via the system message.

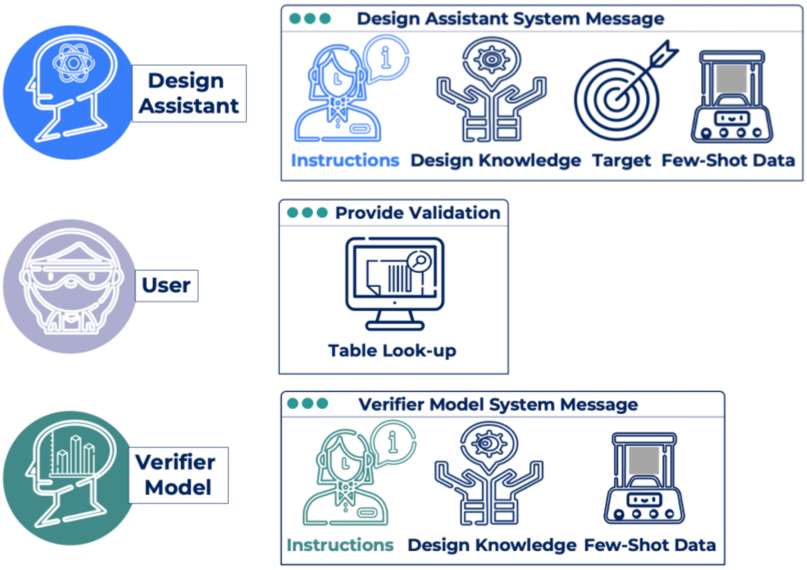


Figure 3: Overview of roles in chat-based design and implementation details.

The **DA** (see Figure 3, top) is tasked with formulating concrete mix designs. It uses the system message and few-shot data from user feedback to guide its recommendations. The system message contains detailed instructions, feasible material combinations, and design knowledge that is critical for the DA to effectively use its knowledge base effectively. This knowledge, generated at different levels of quality using GPT-4, provides the DA with the context needed to optimize its design proposals (see section 3.1.2). The system message also explicitly states the objective of achieving maximum compressive strength and the requirements for the output format.

The **user** (see Figure 3, center) provides laboratory validation feedback to the Assistant's recommendations. To streamline this process for systematic benchmarking, a table look-up has been employed in this work using established test results from the literature. These validation outcomes are provided to the Assistant as laboratory validation in the form of few-shot examples, i.e. a list of previously suggested designs and validation result pairs.

The **VM**, (see Figure 3, bottom) in the extended TVDL implementation, is tasked with analyzing and predicting the expected compressive strength of each AAC formulation proposed by the DA. The VM also has access to design knowledge provided by the system message, to make informed predictions. In addition, the VM uses few-shot data from the user to ground its predictions in the real-world performance of past formulations. The VM`s predictions allow formulations to be ranked and selected based on how closely their estimated strength matches the target strength, which in this study is the formulation with the highest estimated strength. This is selected by the user for further validation.

This study extends previous efforts in the use of Large Language Models (LLMs) for materials design. Ramos et al. [19] implemented an approach where a set of predefined formulations were evaluated using forward predictions by LLMs, essentially replacing the predictive model in traditional DDD with an LLM. Our work extends this method by introducing the DA for generative formulation prediction. This addition effectively reduces the huge formulation space to the test designs generated by the DA, leading to a significant reduction in both operational costs and computational time.

Although Jablonka et al. [18] have used LLMs for generative material design similar to the SFDL, a key difference between their approach and this study lies in how design knowledge is integrated into the predictive processes of both the DA and the VM. The way in which this contextual framework is created is discussed in the following section.

Prompt Design

The prompt design approach followed established guidelines for clarity and specificity in task definition, response formats, and parameter guidelines [12]. This included iterative refinement and consideration of error modes commonly encountered in LLM interactions [35]. For example, word sensitivity, where small differences in wording lead to large variations in model response was mitigated by creating three synonymous but structurally distinct versions of each prompt. This strategy effectively reduces the risk of divergent results due to subtle language variations. The model instructions and feedback were designed to be straightforward, outlining the task objectives and reporting the actual material properties achieved in the laboratory. The design knowledge, however, is at the heart of the KDD approach. It provides the DA and the VM with the necessary context to formulate or evaluate designs in terms of the compressive strength.

In order to investigate different granularities of the knowledge used, AAC design rules were created using GPT-4 in two different ways. The main difference between the two is in the approach to creating the context, which affects the depth and usefulness of the information obtained:

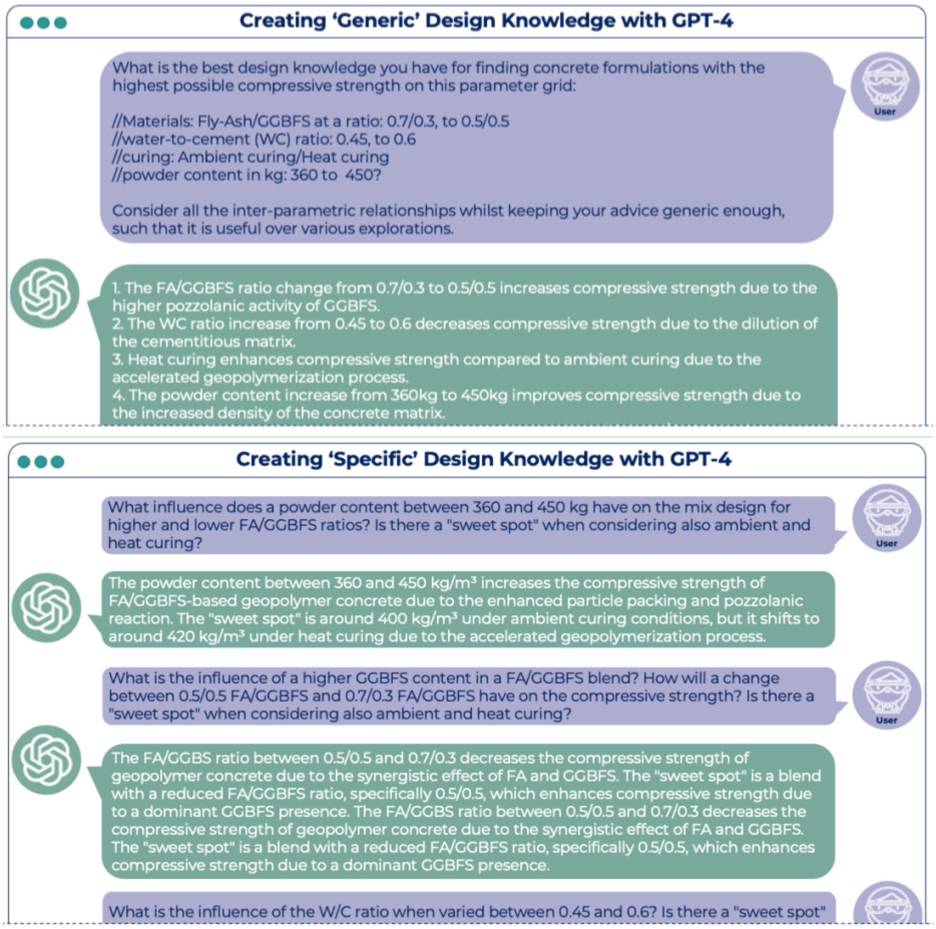
In seeking general design knowledge, GPT-4's inquiry was focused on broader design guidance. Although it referenced the parameter grid, the question was formulated without specific directives (see Figure 4, top).

For specific design knowledge, GPT-4 was consulted on each design parameter individually, including specific initial estimates and concrete instructions on how to change parameters for optimal results (see Figure 4, bottom).

The generation of generic design knowledge is straightforward and can be easily automated in software applications. The method provides a broad overview of the design parameters, suitable for gaining a general understanding of how each parameter might influence the compressive strength of AAC. However, the method does not delve into specific relationships between parameters or provide detailed guidance on how to optimize formulations. For example, in the case of powder blending ratio and the water/cement (W/C) ratio, the general advice identified the directional influence of these parameters on strength but refrained from specifying optimal value.

In contrast, the specific design knowledge required a more complex and iterative approach. Each parameter was evaluated individually by GPT-4, and the model was also asked to elucidate the relationships between different parameters. This method aimed to provide optimal results by providing detailed insights and specific "sweet spot" estimates for each parameter. For example, it could specify an ideal powder content for different curing conditions, taking into account how this parameter interacts with others such as the FA/GGBFS ratio or the W/C ratio. The specific design knowledge thus goes beyond the generic advice and provides the concrete instructions and nuanced understanding needed to fine-tune AAC formulations to achieve optimum compressive strength. While the specific approach is more time consuming and complex, potentially requiring manual intervention for each parameter, it provides precise design knowledge.

The interview prompts required some iterative refinement to ensure that GPT-4 understood the request correctly. The responses were slightly curated by removing redundant statements that often appeared in the final summaries.



Creating 'Generic' Design Knowledge with GPT-4

What is the best design knowledge you have for finding concrete formulations with the highest possible compressive strength on this parameter grid:

//Materials: Fly-Ash/GGBFS at a ratio: 0.7/0.3, to 0.5/0.5
//water-to-cement (WC) ratio: 0.45, to 0.6
//curing: Ambient curing/Heat curing
//powder content in kg: 360 to 450?

Consider all the inter-parametric relationships whilst keeping your advice generic enough, such that it is useful over various explorations.

1. The FA/GGBFS ratio change from 0.7/0.3 to 0.5/0.5 increases compressive strength due to the higher pozzolanic activity of GGBFS.
2. The WC ratio increase from 0.45 to 0.6 decreases compressive strength due to the dilution of the cementitious matrix.
3. Heat curing enhances compressive strength compared to ambient curing due to the accelerated geopolymerization process.
4. The powder content increase from 360kg to 450kg improves compressive strength due to the increased density of the concrete matrix.

Creating 'Specific' Design Knowledge with GPT-4

What influence does a powder content between 360 and 450 kg have on the mix design for higher and lower FA/GGBFS ratios? Is there a "sweet spot" when considering also ambient and heat curing?

The powder content between 360 and 450 kg/m³ increases the compressive strength of FA/GGBFS-based geopolymer concrete due to the enhanced particle packing and pozzolanic reaction. The "sweet spot" is around 400 kg/m³ under ambient curing conditions, but it shifts to around 420 kg/m³ under heat curing due to the accelerated geopolymerization process.

What is the influence of a higher GGBFS content in a FA/GGBFS blend? How will a change between 0.5/0.5 FA/GGBFS and 0.7/0.3 FA/GGBFS have on the compressive strength? Is there a "sweet spot" when considering also ambient and heat curing?

The FA/GGBS ratio between 0.5/0.5 and 0.7/0.3 decreases the compressive strength of geopolymer concrete due to the synergistic effect of FA and GGBFS. The "sweet spot" is a blend with a reduced FA/GGBFS ratio, specifically 0.5/0.5, which enhances compressive strength due to a dominant GGBFS presence. The FA/GGBS ratio between 0.5/0.5 and 0.7/0.3 decreases the compressive strength of geopolymer concrete due to the synergistic effect of FA and GGBFS. The "sweet spot" is a blend with a reduced FA/GGBFS ratio, specifically 0.5/0.5, which enhances compressive strength due to a dominant GGBFS presence.

What is the influence of the W/C ratio when varied between 0.45 and 0.6? Is there a "sweet spot"

Figure 4: Extending design knowledge with GPT-4. Top: Generic design knowledge querying GPT-4 for relationships for a given design parameter grid [36]. Bottom: Specific design knowledge, where GPT-4 is asked about the relationships of each design parameter individually [37].

3.2 Sample Population

The data used was derived from an AAC design study by Rao et al. [14]. The study included 240 AAC formulations using a blend of fly ash (FA) and ground granulated blast-furnace slag (GGBFS) as binders. Each formulation in the dataset was extracted from the publication and includes component descriptions and corresponding 28-day compressive strengths. The dataset includes design variations as detailed in Table 1. The key parameters in each formulation are the amount of binder powder used, the water to cement (W/C) ratio, the blending ratios of FA and GGBFS and the curing method used.

Table 1: Material formulation grid.

Constituent/Process	Considered Parameter Range
Powder Content in kg	360, 370, 380, 390, 400, 410, 420, 430, 440, 450
W/C Ratio	0.45, 0.5, 0.55, 0.6
Powder Blend (FA/GGBFS Ratio)	70/30, 60/40, 50/50
Curing Method	Ambient curing/Heat curing

Figure 5 illustrates our dataset with the distribution of compressive strength represented by a color gradient. The latter shows a large variation with compressive strengths ranging from 20.8 MPa to 65.3 MPa. The 99th percentile strength is 64.9 MPa and the mean strength is 47 MPa.

The strength distribution within the material categories appears to be uniform, with clear trends towards their respective local maxima, suggesting a robust correlation between the design parameters and material properties. However, the analysis of the data shows that it is not trivial to identify exceptional formulations.

Formulations with an equal blend of FA and GGBFS, a W/C ratio of 0.5, a low powder content, and heat curing have the highest compressive strengths (see Figure 5). Conversely, formulations with a higher W/C ratio, a predominance of FA and ambient curing are generally associated with lower compressive strengths.

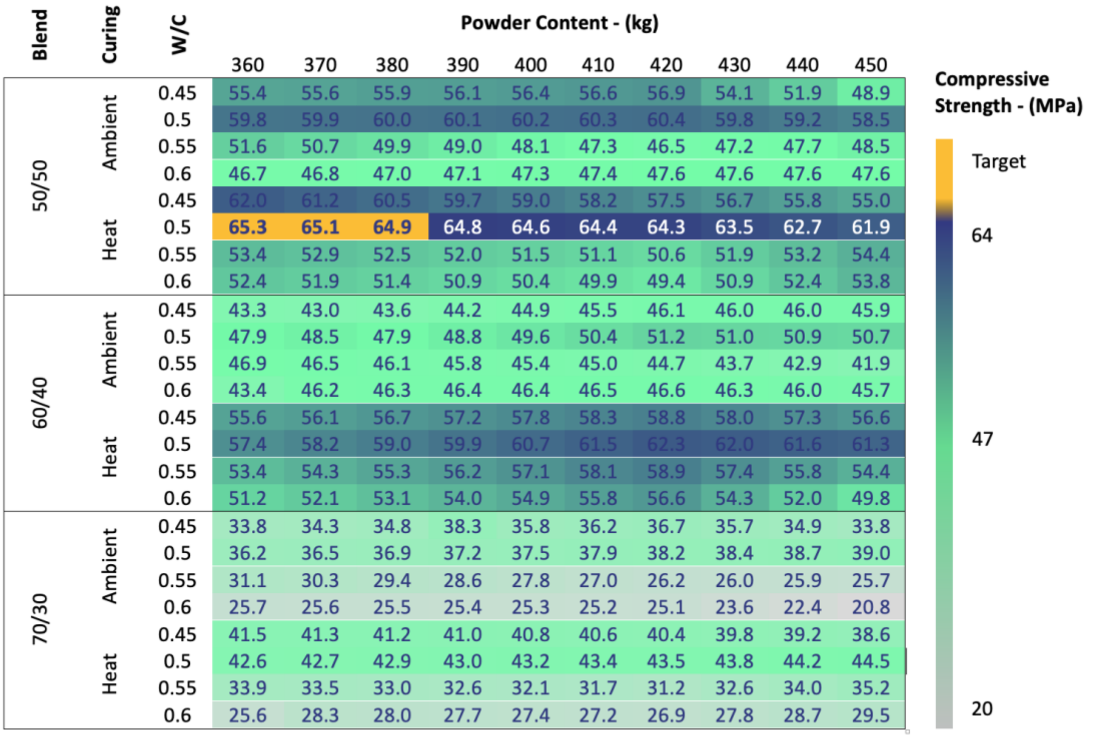
A critical observation is the counter-intuitive relationship between W/C ratio, powder content and compressive strength. Contrary to the conventional wisdom that a very low W/C ratio and higher powder content are associated with higher strength, the highest strength formulations in this study had a much more nuanced composition. In particular, the role of the powder content is ambiguous. In general, higher powder content tends to result in higher strength. However, in the case of the 50/50

FA/GGBFS blend it is the lower powder content that achieves the highest strength. Identifying these anomalies is the key challenge for the LLM as it tests its adaptability in scenarios that deviate from established concrete mix design principles. This dataset illustrates the complexity of sustainable concrete design. The variability in the composition of GGBFS and FA, coupled with the intricate multi-phase chemistry and reaction kinetics over different length and time scales, makes a prescriptive approach challenging in a production process environment. Nevertheless, the dataset demonstrates that exceptional mechanical properties can be achieved, highlighting the potential for discovering optimal formulations.

To integrate this data into the LLM, the LIFT (Language Interface for Text) framework [38] was used to convert the data into a text-based format. For example, a sample formulation is verbalized as:

"The formulation is Powderkg = 380, wc = 0.55, materials = 0.5/0.5, curing = Heat curing."

Such verbalizations facilitate the LLM in generating predictions based on similar descriptions, with the compressive strength of each formulation serving as the target metric in a few-shot learning context.



Blend	Curing	W/C	360	370	380	390	400	410	420	430	440	450
50/50	Ambient	0.45	55.4	55.6	55.9	56.1	56.4	56.6	56.9	54.1	51.9	48.9
		0.5	59.8	59.9	60.0	60.1	60.2	60.3	60.4	59.8	59.2	58.5
		0.55	51.6	50.7	49.9	49.0	48.1	47.3	46.5	47.2	47.7	48.5
		0.6	46.7	46.8	47.0	47.1	47.3	47.4	47.6	47.6	47.6	47.6
	Heat	0.45	62.0	61.2	60.5	59.7	59.0	58.2	57.5	56.7	55.8	55.0
		0.5	**65.3**	**65.1**	**64.9**	64.8	64.6	64.4	64.3	63.5	62.7	61.9
		0.55	53.4	52.9	52.5	52.0	51.5	51.1	50.6	51.9	53.2	54.4
		0.6	52.4	51.9	51.4	50.9	50.4	49.9	49.4	50.9	52.4	53.8
60/40	Ambient	0.45	43.3	43.0	43.6	44.2	44.9	45.5	46.1	46.0	46.0	45.9
		0.5	47.9	48.5	47.9	48.8	49.6	50.4	51.2	51.0	50.9	50.7
		0.55	46.9	46.5	46.1	45.8	45.4	45.0	44.7	43.7	42.9	41.9
		0.6	43.4	46.2	46.3	46.4	46.4	46.5	46.6	46.3	46.0	45.7
	Heat	0.45	55.6	56.1	56.7	57.2	57.8	58.3	58.8	58.0	57.3	56.6
		0.5	57.4	58.2	59.0	59.9	60.7	61.5	62.3	62.0	61.6	61.3
		0.55	53.4	54.3	55.3	56.2	57.1	58.1	58.9	57.4	55.8	54.4
		0.6	51.2	52.1	53.1	54.0	54.9	55.8	56.6	54.3	52.0	49.8
70/30	Ambient	0.45	33.8	34.3	34.8	38.3	35.8	36.2	36.7	35.7	34.9	33.8
		0.5	36.2	36.5	36.9	37.2	37.5	37.9	38.2	38.4	38.7	39.0
		0.55	31.1	30.3	29.4	28.6	27.8	27.0	26.2	26.0	25.9	25.7
		0.6	25.7	25.6	25.5	25.4	25.3	25.2	25.1	23.6	22.4	20.8
	Heat	0.45	41.5	41.3	41.2	41.0	40.8	40.6	40.4	39.8	39.2	38.6
		0.5	42.6	42.7	42.9	43.0	43.2	43.4	43.5	43.8	44.2	44.5
		0.55	33.9	33.5	33.0	32.6	32.1	31.7	31.2	32.6	34.0	35.2
		0.6	25.6	28.3	28.0	27.7	27.4	27.2	26.9	27.8	28.7	29.5

Figure 5: Design Space (DS) visualization showing six material clusters, categorized by FA/GGBFS powder blend and curing method. The materials are color coded to reflect the 28-day compressive strength with the target materials above the 99% strength quantile highlighted in orange.

3.3 Baseline Methods and Benchmarking

In the field of data-driven design (DDD) for material formulation, this study evaluates three basic methods: Bayesian Optimization (BO) with Gaussian Process Regression (GPR), Sequential Learning (SL) with Random Forest (RF), and Random Draw (RD) as a stochastic control method. Each method represents a unique strategy for tackling the complexities of material design.

BO with GPR is a widely used technique in materials science [39]. It uses a probabilistic model, Gaussian Process Regression, to predict material properties, facilitating an informed and efficient exploration of the design space. This method is particularly adept at understanding the nuances of the underlying data distribution, which is critical in identifying optimal material formulations.

SL with RF employs the Random Forest algorithm in a sequential learning framework [40]. This approach is designed to progressively improve the accuracy of material property predictions with each iteration. By refining the selection process of candidate materials, SL with RF aims to optimize the accuracy of predictions, making it a robust method in the iterative DDD process.

A comparison is often made between DDD and RD. In RD, experiments are conducted randomly without predictive modelling, highlighting their relative effectiveness in navigating the material design space. The probability of success of RD follows a hypergeometric distribution.

A common framework for demonstrating its capabilities is through simulated experiments, where the outcomes of all data points are predetermined. The benchmarking procedure is as follows:

- A subset of the available data is initially provided to the DDD algorithms (applicable to RF and GPR only) as a basis for training.
- In each development cycle, the most promising data point predicted by the DDD is integrated. This process

is repeated for ten cycles or until the target strength is reached for all three baseline methods. Within our KDD approach we consider ten cycles in the case of GPT-3.5-Turbo, while restricting to five cycles for GPT-4-Turbo. If one of the models reaches the design target before the last cycle, we stop the iteration.

- The success of the optimization is quantified by the highest quality material property identified at each iteration.

The benchmarking results are presented as a distribution of the cumulative performance achieved in each development cycle. This performance distribution, which may vary depending on the initial conditions of the experiment, is statistically evaluated. For example, using the 10th percentile as an indicator provides insight into the efficiency of the methods. This percentile marks the lower bound performance limit that was achieved in 90% of the design runs, indicating a high robustness of the method in attaining such compressive strengths at a given number of development cycles. The methodologies underlying the benchmarking approach and the baseline methods are described in more detail in [9].

4. Experiments

This chapter outlines the experimental program carried out in this study. It begins with a detailed description of the experimental parameters, followed by an explanation of the procedures used.

4.1 Experimental parameters

The varied experimental parameters and their configurations are summarized in Table 2 and are explained below.

Table 2: Parameters used and/or varied in this study.

	Parameter	Description
1	Chat Models	GPT-3.5-Turbo / GPT-4-Turbo
2	Baseline Models	RF/ BO/ RD
3	Target Strength	64.9 Mpa (99% quantile)
4	Design Knowledge	None / Generic / Specific
5	Reasoning Strategy	SFDL (TT=1) / TVDL (TT=3, GPT-3.5-Turbo only)
6	Number of Development Cycles	GPT-3.5-Turbo, Baseline Models: 10 / GPT-4-Turbo: 5
7	Number of runs Chat Model / Baseline Methods	15 / 30

Two versions of the chat model were evaluated: GPT-3.5-Turbo and GPT-4-Turbo. The latter serves as a baseline for the impact of model scaling compared to the implementation of a TT approach. Based on our preliminary investigations, the model temperature set to zero provided slightly better performance than a higher setting and was therefore chosen for this study.

Additionally, three baseline methods were used for comparative analysis: SL with RF, BO with GPR, and RD. These provided a diverse range of strategic approaches to materials design, from stochastic to data-driven predictive methods. For this work the scikit-learn implementation of GPR [41] and the Lolo RF [21] were implemented.

The target strength for AAC design was set at the 99% quantile of the strengths provided in the dataset. This high threshold ensured that the formulations proposed by the models were not only feasible but also optimized for peak performance.

The design knowledge built into the context of the chat models varied in three categories: generic, specific, and none. The latter simply refers to no design knowledge being provided to the DA and VM. The aim of this variation was to assess the impact of the richness of contextual information on the chat-based design predictions.

Two reasoning strategies were tested: SFDL and TVDL. The latter was only used in the case of GPT-3.5-Turbo. The number of development cycles differed between the chat models: 10 cycles for GPT-3.5-Turbo and a reduced number of 5 cycles for GPT-4-Turbo due to the significantly higher costs.

Each KDD was run 15 times (consisting of five runs for three rephrased versions of each design knowledge). Multiple runs allowed for a statistical analysis of the performance of the models. The runs for the baseline methods are on the one hand much cheaper and faster, but more prone to variation due to randomness, so 30 repetitions were performed for each approach.

4.2 Experimental Procedure

The initialization phase for the models included a preliminary dataset relevant to AAC formulations. For the chat models, GPT-3.5-Turbo and GPT-4-Turbo, this phase included providing context in three variations and design knowledge at three levels of specificity, resulting in nine unique operating conditions (Figure 6, left).

The GPT-3.5-Turbo model underwent ten development cycles for both the Standard Feedback Design Loop (SFDL) and the Testing and Verification Design Loop (TVDL), while the GPT-4-Turbo was limited to five cycles in the SFDL only (Figure 6, center and right). Each model configuration was replicated 15 times for statistical evaluation. This setup resulted in the GPT-3.5-Turbo model generating 900 formulations, with the TVDL requiring four API calls per formulation (1800 total prompts for 450 formulations). The GPT-4-Turbo model, used exclusively in the SFDL, generated 225 formulations. In total, 2475 prompts were used to create 1125 formulations for user validation.

In comparison, the baseline models were initialized differently. Both GPR and RF started with four random samples, excluding any formulations that already met the target requirement. The most promising sample from these initial sets was then added over ten development cycles. In contrast, the Random Draw (RD) method did not require initial samples. Instead, a random sample was added in each of the ten development cycles.

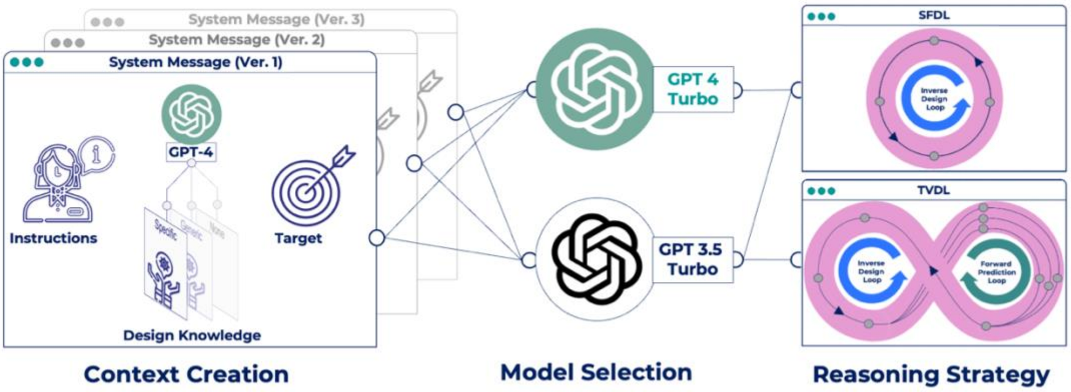


Figure 6: Benchmarking the KDD approach in different configurations. Left: Variations of the system message including three different levels of design knowledge, each in three rephrased versions, respectively. Center: Model selection; Right: Implemented reasoning strategies SFDL and TVDL

5. Results

The benchmarking results, presented in Figures 7 to 9, show the mean achieved strength, along with the 10% and 90% upper- and lower limits of each development cycle. The quantitative results are summarized in Table 3 in terms of the lower bound limit compressive strength achieved. For the baseline and GPT-3.5-Turbo models the results are shown at three key intervals: The first, fifth and tenth development cycles, which allow the assessment of zero-shot performance, few-shot performance, and final performance, respectively. GPT-4-Turbo only iterates over five cycles, i.e. the results at the fifth development cycle also mark the final performance of the models. Table 4 also shows the relative performance gains that were achieved by choosing the ideal strategy over the worst decision.
In our study, the lower bound value is a crucial performance indicator, reflecting the lower bound rule used in the approval of civil engineering materials, particularly in Europe [42]. This benchmark ensures that materials meet essential safety and reliability standards, focusing on worst-case performance rather than average results. Applying this criterion to the design methods aligns the evaluation with industry practice. It provides a realistic assessment of the practical applicability of each method, particularly in safety-critical applications. This focus becomes critical when considering these methodologies as potential replacements for traditional prescriptive approaches. Traditionally, the lower bound criteria apply directly to materials. However, with innovative, adaptive design methods, the criteria shift towards assessing the reliability of the design process itself. Therefore, our study's emphasis on the lower bound value aims to demonstrate the effectiveness of new methodologies not just in theory, but as viable, robust alternatives to traditional material design approaches.

The primary observation from the results is the consistent ability of the KDD approach to outperform the baseline models in terms of the achieved lower bound compressive strength achieved in each development cycle when the correct configuration is applied (see results in Table 3).

The quality of the context proved to be the most significant factor, with an 8% improvement with more specific design knowledge and an 8.4% decrease in performance when no design knowledge was used (see Table 4). This effect was particularly pronounced in the first round, where the average gain was 22% and the average performance loss was 41%. The SFDL implementation of GPT-3.5-Turbo, when equipped with specific design knowledge outperforms the baseline methods but shows a tendency to plateau in performance improvement after the third round (see Figure 7).

The TVDL showed a continuous improvement over previous proposals, resulting in an average performance gain of 3.6 % across all development cycles. This improvement was even more significant in later iterations, reaching an average relative improvement of 7.5 % in the final round. Although these gains seem relatively small compared to the quality of the context, the TVDL performance curve clearly shows that a lack of design knowledge can be compensated by an increase in TT
as the design progresses: While especially the design processes without design knowledge start with a very low performance, the TT allows a successive improvement of the design proposals based on the few-shot validations from the lab. Even if it ultimately fails to catch up with generic or specific design knowledge, a longer development time could enable further improvements.

In the TVDL, using either generic or specific design knowledge, the chat-based design consistently outperformed the baseline methods in every iteration and even outperformed the baselines in the final cycle without design knowledge.

The use of GPT-4-Turbo (see Figure 8) resulted in an average 2 % higher strength result compared to GPT-3.5-Turbo in the SFDL. However, in the later stages the TVDL implementation outperformed GPT-4-Turbo. This is except for scenarios with no design knowledge and few validation examples. Here a weak DA was not effectively mitigated by a similarly weak VM, highlighting the importance of capable verifiers.

Table 3: Benchmarking results. Left: Comparison of the 10% lower limits of the 28-day compressive strength: Values recorded at 1st, 5th, and 10th development cycles, along with the average value calculated over all development cycles (Note: **bold** = best, underline = second best, *italic* = third best).

Context	Model	TT	Compressive Strength at 1st dev. cycle (MPa)	Compressive Strength at 5th dev. cycle (MPa)	Compressive Strength at 10th dev. cycle (MPa)
none	GPT-3.5 Turbo	1	33.8	58.1	58.1
		3	35.8	49.9	*61.2*
	GPT-4 Turbo	1	39.5	58.0	
generic	GPT-3.5 Turbo	1	*55.0*	55.8	55.8
		3	*55.0*	*61.5*	**64.3**
	GPT-4 Turbo	1	*55.0*	57.3	
specific	GPT-3.5 Turbo	1	60.2	*62.3*	*62.3*
		3	**64.3**	**64.3**	**64.3**
	GPT-4 Turbo	1	60.2	60.2	
Gaussian Process Regression			53.9	59.8	59.9
Random Forrest			44.7	57.1	60.2
Random Draw			29.9	52.4	55.9

Table 4: Relative performance influence of parameters settings in terms % of lower limit compressive strength.

		Performance gains in (%)			
Parameter	**Parameter Setting vs. Alternative**	1st dev. cycle	5th dev. cycle	10th dev. cycle	Ø all dev. cycles
Context	*Generic vs. None*	-41.3	-3.6	-3.6	-8.4
	Generic vs. Specific	22.0	6.9	4.8	8
Model	*GPT-3.5 Turbo (SFDL) vs. GPT-4 Turbo (SFDL)*	3.7	1.8	(-)	2.0
	GPT-3.5 Turbo (SFDL)) vs. GPT-3.5 Turbo (TVDL)	-2.1	3.1	7.5	3.6

In summary, chat-based design methods have proven to outperform strong baseline methods in this analysis. The most significant performance gains are achieved through higher context quality, especially in the early rounds. The TVDL strategy has also been shown to improve performance, especially in the later rounds when there are enough few-shot examples. Interestingly, more TT and higher context quality seem to offer complementary benefits: while higher context quality provides a higher initial performance offset, TT leads to steeper performance gains over development cycles. Surprisingly, the use of GPT-4, a more advanced chat model, in the SFDL implementation did not yield significant benefits. Consequently, the most effective result after five and ten development cycles was achieved by the GPT-3.5-Turbo model in the TVDL implementation. At the tenth development cycle both, generic and specific knowledge, achieved a lower bound strength of 64.3 MPa, narrowly missing the design target of 64.9 MPa by only 0.6 MPa (see Figure 5). The best baseline model, RF, gave a final strength of 60.2 MPa in the final round. The best zero-shot result was also achieved by GPT-3.5-Turbo in the TVDL implementation with specific design knowledge, reaching 64.3 MPa. In contrast, the best baseline result in the first development cycle was achieved by BO with 53.9 MPa.

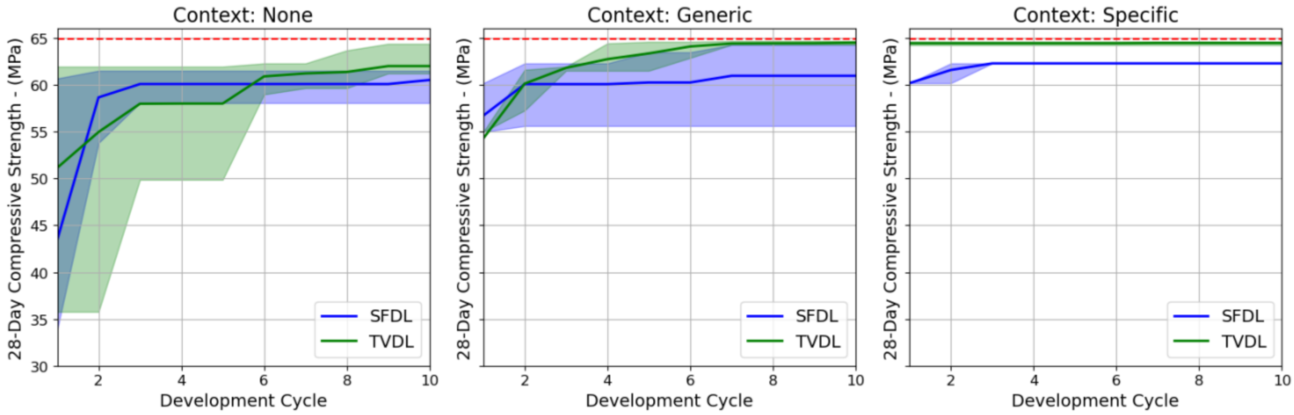


Figure 7: Performance of the chat-based materials design using GPT-3.5-Turbo in terms of the achieved 28-day compressive strength in MPa at each development cycle as the mean value (solid line) and the 10% and 90% lower and upper bounds (areas) for the SFDL implementation (purple) and the TVDL implementation (green). F.l.t.r. Design knowledge None/Generic/Specific.

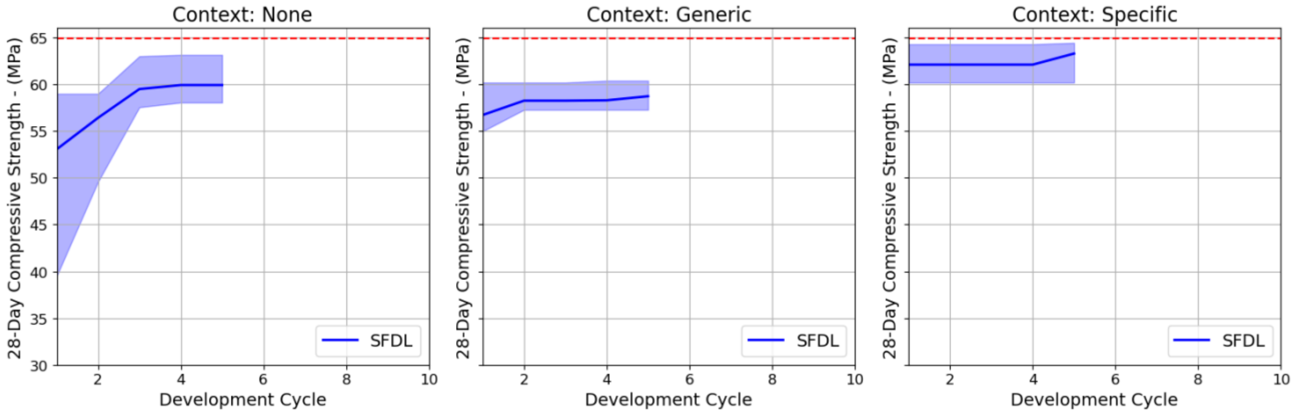


Figure 8: Performance of chat-based materials design using GPT-4-Turbo in terms of the achieved 28-day compressive strength in MPa at each development cycle as the mean value (solid line) and the 10% and 90% lower and upper bounds (areas) for the SFDL implementation. F.l.t.r. Design knowledge None/Generic/Specific.

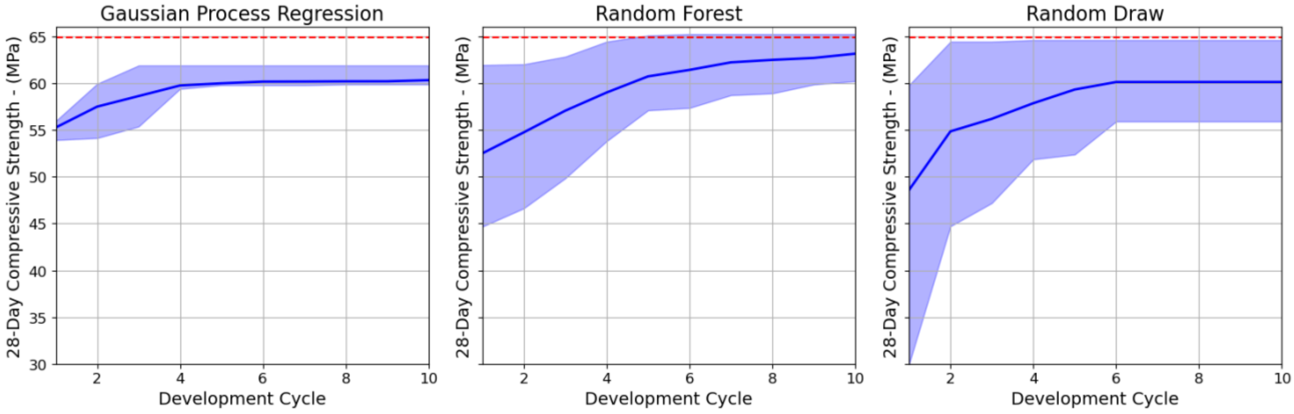


Figure 9: Performance of the DDD baseline methods in terms of the achieved 28-day compressive strength in MPa at each development cycle as the mean value (solid line) and the 10% and 90% lower and upper bounds (areas). F.l.t.r. BO with GPR, SL with RF, and RD.

6. Conclusion and Discussion

The aim of this study was to address a pressing challenge in the field of materials science: the optimization of material formulations through a KDD approach with a focus on waste-based materials. The urgency of this challenge lies in the need to increase the use of secondary, recycled materials, thereby contributing to increased resource efficiency and reducing the significant carbon emissions associated with traditional cementitious materials.

The results of this research confirm the central hypothesis that a KDD approach using LLMs and incorporating fuzzy design knowledge, is indeed capable of producing high quality material formulations. This conclusion, drawn from a comprehensive benchmarking and analysis, demonstrates the effective application and potential of LLMs in the field of sustainable materials design.

Here, natural language knowledge represents an entirely new modality of input data that has so far remained untapped in computational materials design. This represents a significant departure from prescriptive design rules or the more traditional DDD. In the benchmarking study presented, this approach enabled the rapid and highly reliable design of materials from scratch which were based on highly variable resource flows.

Four main conclusions can be drawn from our systematic benchmarking:

1) In a direct comparison the chat-based design approach with established baseline methods, we observed a superior performance in terms of the lower bound strength achieved, even without any initial training data. This was a remarkable achievement, especially considering that methods such as RF and GPR initially had a slight advantage due to their dependence on initial training data. This was achieved through a significantly higher robustness of the design result - which is surprising given the apparently more solid input data of traditional DDD compared to the relatively fuzzy design knowledge in KDD.

2) A critical finding of our study was the paramount importance of the quality of contextual information provided to LLMs in enhancing their performance. There were significant performance improvements when specific design knowledge and sweet spot estimates were incorporated into the model, demonstrating that accurate and contextually rich information is key to optimizing AI-driven design methodologies. Interestingly, much of this core design knowledge was autonomously generated using GPT-4, highlighting the potential for AI integration in design software to generate flexible and accurate information specific to the task at hand.

3) Another important aspect of our study was the adoption of scaling strategies, inspired by applications in text-based mathematical problems, and applied here to materials design tasks. While we have shown that incorporating of domain knowledge into the context is a very effective approach to improving predictive performance, it is not universally applicable. For example, domain knowledge may not always be available and may sometimes be incorrect. Therefore, we investigated an additional mechanism to improve the performance of LLMs. We focused on strategies that can be applied at test time as they do not require access to the model weights and can be applied without costly training steps. By increasing the TT and using GPT-3.5-Turbo as the verifier model, we observed significant performance improvement. The 'effectiveness of this strategy may be attributed to the inherent complexity involved in predicting multiple design parameters in AAC formulations, a task that challenges the linear, forward-looking approach of transformer models. The inverse nature of material design queries, requiring inference from a result to potential causes, is arguably more complex than the forward predictions typically made by verifier models [43]. Thus, the integration of verifier models complemented and compensated for the limitations of the DA, leading to more efficient use of the feedback loop in the design process.

4) This study demonstrated that GPT-3.5-Turbo with the TVDL strategy enabled the smaller model to surpass the performance of its larger counterpart GPT-4-Turbo. This was achieved at a two and a half times lower cost, highlighting the effectiveness of using post-training enhancements such as test-time strategies and verifier models.

In summary, this study demonstrates the of LLMs in AAC design and provides a novel path for future AI applications in materials science. Our findings highlight how LLMs can advance sustainable materials design, contributing to environmental goals and enriching the role of AI in materials engineering.

6.1 Future Research

While this study focuses on a specific design case in AAC, it opens the door to a variety of promising research avenues in materials science and the application of LLMs. The concept of inverse design, combining initial suggestions with systematic iterative improvements, stands out as a key area for evaluating the capabilities of LLMs. The challenge here is to achieve both, a robust initial performance and continuous improvement of subsequent proposals.

A critical takeaway from our research is the significant potential for enhancing the domain-specific capabilities of LLMs in materials science through relatively generic post-training measures. This enhancement involves refining the context quality and extending TT in conjunction with a VM. While improving context quality may present practical challenges due to the constant pursuit of optimal quality, scaling up TT appears to be a cost-effective and feasible strategy. This has as recently been taken up by Yuan et al. [44] to improve the overall performance of general LLMs (Llama2). However, as highlighted in an earlier paper [29], the performance of the VM can have a critical impact on performance. Identifying ideal VMs for specific domain tasks is a fascinating research challenge.

These exploratory paths are particularly relevant for applications in less constrained, real-world design problems. Traditional DDD methods often require limiting the solution space to be constrained. However, even in such constrained domains, the solution space on a coarsely sampled grid can contain millions of possibilities, making it difficult to identify optimal solutions. Generative design approaches do not require constraints, increasing the chances of identifying solutions not previously considered. In this context, scalable TT methods offer a promising direction, although their effectiveness warrants further investigation and the development of methodological benchmarking frameworks.

Systematic benchmarks, involving a wider range of examples, are essential for these future studies. Such benchmarks should take into account the dynamic nature of the tasks assigned to LLMs. Implementing the automated benchmarking approach from this study requires datasets that are large enough to cover a significant design space and structured enough (e.g., on a regular design grid) to facilitate effective communication with LLMs. This will ensure that proposed formulations can be validated from the literature. However, most of the available data are compiled from many scattered, small-scale, heterogeneous laboratory investigations (e.g. [45]). An effective restriction of the solution space for benchmarking the generative capabilities of the DA is not feasible with currently available models - but perhaps by increasing the usable context window of future LLM generations. Here, research tasks that enable rapid and automatable design processes, either through simulation or accelerated laboratory validation, are currently highly desirable alternatives (e.g. [46]).

In summary, this study not only highlights the potential of LLMs in materials science, but also lays the groundwork for future explorations in three main strategies: improving context quality, using enhanced TT, and employing capable VMs. Future explorations could range from improving the core reasoning capabilities of LLMs to adapting them for more diverse and complex scientific applications, thereby broadening their utility and impact in the scientific community.

Declaration of interests and data availability

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The data and source code necessary to reproduce the results will be made available according to F.A.I.R. standards upon the acceptance of this contribution. The code, data and results required to reproduce this work are available under: https://github.com/BAMcvoelker/LLM-s-can-Design-Sustainable-Concrete-a-Systematic-Benchmark-Code-

CRediT author statement

Christoph Völker: Conceptualization, Methodology, Software, Formal analysis, Visualization, Supervision, Project administration, Funding acquisition, Writing - Original Draft, Tehseen Rug: Conceptualization, Methodology, Software, Formal analysis, Writing- Original Draft, Kevin Maik Jablonka: Writing - Review & Editing, Sabine Kruschwitz: Writing - Review & Editing, Project administration, Funding acquisition

Acknowledgements

We gratefully acknowledge support from the European Union's Horizon Europe research and innovation program under grant agreement No. 101056773 for funding this work under the Reincarnate project.

References

[1] U. Environment, K. L. Scrivener, V. M. John and E. M. Gartner, “Eco-efficient cements: Potential economically viable solutions for a low-CO2 cement-based materials industry,” *Cement and Concrete Research,* vol. 114, pp. 2-26; DOI: https://doi.org/10.1016/j.cemconres.2018.03.015, 2018.

[2] J. L. Provis and J. van Deventer, Alkali Activated Materials, State-of-the-Art Report, RILEM TC 224-AAM, Springer, Dordrecht, Heidelberg, New York, London, 2014.

[3] H. S. Gökçe, M. Tuyan, K. Ramyar and M. L. Nehdi, “Development of Eco-Efficient Fly Ash–Based Alkali-Activated and Geopolymer Composites with Reduced Alkaline Activator Dosage,” *Journal of Materials in Civil Engineering,* vol. 32, no. 2, pp. 04019350; DOI: 10.1061/(ASCE)MT.1943-5533.0003017, 2020.

[4] J. He, Y. Jie, J. Zhang, Y. Yu and G. Zhang, “Synthesis and characterization of red mud and rice husk ash-based geopolymer composites,” *Cement and Concrete Composites,* vol. 37, pp. 108-118; DOI: http://dx.doi.org/10.1016/j.cemconcomp.2012.11.010 , 2013.

[5] J. L. Provis, A. Palomo and C. Shi, “Advances in understanding alkali-activated materials,” *Cement and Concrete Research,* pp. 110-125, 2015.

[6] R. Firdous, M. Nikravan, R. Mancke, M. Vöge and D. Stephan, “Assessment of environmental, economic and technical performance of geopolymer concrete: a case study,” *J Mater Sci,* no. 57, pp. 18711-18725; DOI: https://doi.org/10.1007/s10853-022-07820-6 , 2022.

[7] C. Völker, R. Firdous, D. Stephan and S. Kruschwitz, “Sequential learning to accelerate discovery of alkali-activated binders,” *Journal of materials science,* pp. 15859-15881, 2021.

[8] C. Völker, S. Kruschwitz, B. Moreno Torres, R. Firdous, G. A. Zia and D. Stephan, “Accelerating the search for alkali-activated cements with sequential learning,” in *fib International Congress*, Oslo, Norway, 2022.

[9] C. Völker, B. M. Torres, T. Rug, R. Firdous, G. A. J. Zia, S. Lüders, H. L. Scaffino, M. Höpler, F. Böhmer, M. Pfaff, D. Stephan and S. Kruschwitz, “Data driven design of Alkali-activated concrete using Sequential Learning,” *Journal of cleaner Production ,* pp. 1-40; DOI: 10.1016/j.jclepro.2023.138221, 2023.

[10] S. Ament, A. Witte, N. Garg and J. Kusuma, “Sustainable Concrete via Bayesian Optimization,” *Arxiv (Preprint),* pp. 1-10; DOI: arXiv:2310.18288v3, 2023.

[11] E. Saleh, A. Tarawneh, M. Naser, M. Abedi and G. Almasabha, “Gaussian Process-Batch Bayesian optimization framework for mixture design of ultra high performance concrete,” *Construction and Building Materials,* no. 330, p. 127270; https://doi.org/10.1016/j.conbuildmat.2022.127270, 2022.

[12] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro and Y. Zhang, “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” *Arxiv (Preprint),* pp. 1-155; DOI: arXiv:2303.12712v5, 2023.

[13] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li and Z. Sui, “A Survey on In-context Learning,” *Arxiv,* pp. 1-21; DOI: arXiv:2301.00234v3, 2023.

[14] G. M. Rao and T. D. G. Rao, “A quantitative method of approach in designing the mix proportions of fly ash and GGBS-based geopolymer concrete,” *Australian Journal of Civil Engineering,* vol. 16, no. 1, pp. 53-63; DOI: 10.1080/14488353.2018.1450716, 2018.

[15] OpenAI, “Documentation - Models ('gpt-4-1106-preview' and 'gpt-3.5-turbo-1106'),” [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5. [Accessed 1 January 2024].

[16] J. Boyko, J. Cohen, N. Fox, M. Han Veiga, J. I.-H. Li, J. Liu, B. Modenesi, A. H. Rauch, K. N. Reid, S. Tribedi, A. Visheratina and X. Xie, “An Interdisciplinary Outlook on Large Language Models for Scientific Research,” *Arxiv (Pre-Print),* pp. 1-27; DOI: arXiv:2311.04929v1, 2023.

[17] M. R. AI4Science and M. A. Quantum, “The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4,” *Arxiv (Preprint),* pp. 1-230; arXiv:2311.07361v2, 2023.

[18] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, “Leveraging Large Language Models for Predictive Chemistry,” *ChemRxiv (Preprint),* pp. 1-22; DOI: https://doi.org/10.26434/chemrxiv-2023-fw8n4-v3, 2023.

[19] M. C. Ramos, S. S. Michtavy, M. D. Porosoff and A. D. White, “Bayesian Optimization of Catalysts With In-context Learning,” *Arxiv (Preprint),* p. DOI: 10.48550/ARXIV.2304.05341, 2023.

[20] K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans and N. Gastellu, “14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon,” *Digital Discovery,* vol. 2, no. 5, pp. 1233-1250; DOI: 10.1039/D3DD00113J, 2023.

[21] M. Hutchinson, “lolopy,” [Online]. Available: https://pypi.org/project/lolopy/. [Accessed 14 01 2024].

[22] D. Boiko, R. MacKnight, B. Kline and G. Gomes, “Autonomous chemical research with large language models,” *Nature,* vol. 624, pp. 570-578;https://doi.org/10.1038/s41586-023-06792-0, 2023.

[23] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, “ChemCrow: Augmenting large-language models with chemistry tools,” *Arxiv (Preprint),* pp. 1-38; DOI: arXiv:2304.05376 , 2023.

[24] G. Suriya, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang and S. Bubeck, “Textbooks Are All You Need,” *Arxiv (Preprint),* pp. 1-26; DOI: arXiv:2306.11644v2 , 2023.

[25] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu and D. Amodei, “Scaling Laws for Neural Language Models,” *Arxiv (Preprint),* pp. 1-30; DOI: arXiv:2001.08361v1, 2020.

[26] N. Muennighoff, A. M. Rush, B. Barak, T. Le Scao, A. Piktus, N. Tazi, S. Pyysalo, T. Wolf and C. Raffel, “Scaling Data-Constrained Language Models,” *Arxiv (Preprint),* pp. 1-50; DOI: arXiv:2305.16264v4, 2023.

[27] . L. Jones, “caling Scaling Laws with Board Games,” *Arxiv (Preprint),* pp. 1-8; DOI: arXiv:2104.03113v2 , 2021.

[28] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse and J. Schulman, “Training Verifiers to Solve Math Word Problems,” *Arxiv (Preprint),* pp. 1-22; DOI: arXiv:2110.14168v2, 2021.

[29] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever and K. Cobbe, “Let’s Verify Step by Step,” *Arxiv (Preprint),* pp. 1-29; DOI: arXiv:2110.14168v2 , 2023.

[30] T. Davidson, J.-S. Denain, P. Villalobos and G. Bas, “AI capabilities can be significantly improved without expensive retraining,” *Arxiv,* pp. 1-30; DOI: arXiv:2312.07413v1, 2023.

[31] A. Zunger, “Inverse design in search of materials with target functionalities,” *Nature Reviews Chemistry,* vol. 2, pp. 1-16; https://doi.org/10.1038/s41570-018-0121, 2018.

[32] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan and S. Yao, “Reflexion: Language Agents with Verbal Reinforcement Learning,” *Arxiv (Preprint),* pp. 1-19; DOI: arXiv:2303.11366 , 2023.

[33] V. Nair, E. Schumacher, G. Tso and A. Kannan, “DERA: Enhancing Large Language Model Completions with Dialog-Enabled Resolving Agents,” *Axiv (Preprint) ,* pp. 1-38; DOI: arXiv.2303.17071v1, 2023.

[34] K. Hebenstreit, R. Praas, L. P. Kiesewetter and M. Samwald, “An automatically discovered chain-of-thought prompt generalizes to novel models and datasets,” *Arxiv (Preprint),* pp. 1-10; DOI: arXiv:2305.02897v2, 2023.

[35] I. R. McKenzie, A. Lyzhov, M. Pieler, A. Parrish, A. Mueller, A. Prabhu, E. McLean, A. Kirtland, A. Ross, A. Liu, A. Gritsevskiy, D. Wurgaft, D. Kauffman, G. Recchia, J. Liu and Cavanagh, “Inverse Scaling: When Bigger Isn't Better,” *Arxiv (Preprint),* pp. 1-36; DOI: arXiv:2110.14168v2 , 2023.

[36] OpenAI, “Playground Web App: Augmenting generic design knowledge,” [Online]. Available: https://platform.openai.com/playground/p/R3ni8dAjFhbp6k0xdrVdlmJ6?model=gpt-4&mode=chat. [Accessed 15 January 2024].

[37] OpenAI, “Playground Web App: Augmenting specific design knowledge,” [Online]. Available: https://platform.openai.com/playground/p/FNn2n778h7nPTI1aV7tBpxOD?model=gpt-4&mode=chat. [Accessed 15 January 2024].

[38] T. Dinh, Y. Zeng, R. Zhang, Z. Lin, M. Gira, S. Rajput, J.-y. Sohn, D. Papailiopoulos and K. Lee, “LIFT: Language-Interfaced Fine-Tuning for Non-Language Machine Learning Tasks,” *Arxiv (Preprint),* pp. 1-54; DOI: arXiv:2206.06565v4, 2022.

[39] B. Rohr, H. S. Stein, D. Guevarra, Y. Wang, J. A. Haber, M. Aykol, S. K. Suram and J. M. Gregoire, “Benchmarking the acceleration of materials discovery by sequential learning,” *Chem. Sci.,* vol. 11, no. 10, pp. 696-2706; doi 10.1039/C9SC05999G, 2020.

[40] J. Ling, M. Hutchinson, E. Antono, S. Paradiso and B. Meredig, “High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates,” *Integrating Materials and Manufacturing Innovation,* pp. 207-217, 2017.

[41] Scikit Learn, “Scikit Learn,” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html. [Accessed 15 December 2022].

[42] EN 1990:2002+A1, “Basis of structural design,” European Committee for Standardization, Brussels, Belgim, 2005.

[43] Z. Wu, L. Qiu, A. Ross, E. Akyürek, B. Chen, B. Wang, N. Kim, J. Andreas and Y. Kim, “Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks,” *Arxiv (Preprint) ,* pp. 1-49; DOI: arXiv:2307.02477v2, 2023.

[44] W. Yuan, R. Y. Pang, K. Cho, S. Sukhbaatar, J. Xu and J. Weston, “Self-Rewarding Language Models,” *Arxiv (Preprint),* pp. 1-15; DOI: arXiv:2401.10020v, 2024.

[45] B. Moreno Torres, C. Völker and R. Firdous, “Concreting a sustainable future: A dataset of alkali-activated concrete and its properties,” *Data in Brief,* vol. 50, pp. 1-9; DOI: https://doi.org/10.1016/j.dib.2023.109525, 2023.

[46] E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya and E. Olivetti, “Autonomous experimentation systems for materials development: A community perspective,” *Matter,* vol. 4, no. 9, pp. 2702-2726; DOI: https://doi.org/10.1016/j.matt.2021.06.036, 2021.