Bike-Bench: A Bicycle Design Benchmark for Generative Models with Objectives and Constraints

Lyle Regenwetter
MIT
regenwet@mit.edu

Yazan Abu Obaideh
ProgressSoft
yazan.amer@protonmail.com

Fabien Chiotti
Utrecht University
fabien.chiotti@gmail.com

Ioanna Lykourentzou
Utrecht University
i.lykourentzou@uu.nl

Faez Ahmed
MIT
faez@mit.edu

Abstract

We introduce Bike-Bench, an engineering design benchmark for evaluating generative models on problems with multiple real-world objectives and constraints. As generative AI's reach continues to grow, evaluating its capability to understand physical laws, human guidelines, and hard constraints grows increasingly important. Engineering product design lies at the intersection of these difficult tasks, providing new challenges for AI capabilities. Bike-Bench evaluates AI models' capability to generate designs that not only resemble the dataset, but meet specific performance objectives and constraints. To do so, Bike-Bench quantifies a variety of human-centered and multiphysics performance characteristics, such as aerodynamics, ergonomics, structural mechanics, human-rated usability, and similarity to subjective text or image prompts. Supporting the benchmark are several datasets of simulation results, a dataset of 10K human-rated bicycle assessments, and a synthetically-generated dataset of 1.4M designs, each with a parametric, CAD/XML, SVG, and PNG representation. Bike-Bench is uniquely configured to evaluate tabular generative models, LLMs, design optimization, and hybrid algorithms side-by-side. Our experiments indicate that LLMs and tabular generative models fall short of optimization and optimization-augmented generative models in both validity and optimality scores, suggesting significant room for improvement. We hope Bike-Bench, a first-of-its-kind benchmark, will help catalyze progress in generative AI for constrained multi-objective engineering design problems. Code, data, and other resources are published at decode.mit.edu/projects/bikebench/.

1 Introduction

Generative AI has recently captured widespread attention for its general-purpose problem-solving capabilities [1, 2, 3]. Despite a wealth of exploratory work [4, 5, 6, 7], generative AI has not seen widespread adoption [8] in the trillion-dollar engineering design industry [9, 10]. Engineering design can be generally characterized as the methodical decision-making process needed for the physical realization of products, systems, or other real-world solutions [11, 12, 13]. Many of engineering design's requisite skills pose significant challenges for current Generative AI models [14, 15], including exact constraint satisfaction, adherence to quantitative and qualitative design guidelines, and intimate knowledge of multidisciplinary physical laws. To name explicit examples, in ship hull design, generative models have been found to extensively violate geometric, performance, and safety constraints, sometimes over 95% of the time [16]. In large language model (LLM) benchmarks, models fail to extract precise design regulations from engineering standards [17]. Finally, in free-form structural design, generative models fall far short of optimization algorithms [18] due to their inability

Preprint.

to learn generalizable physics rules. To excel in engineering design, generative models must overcome their imprecise constraint satisfaction, their struggle to understand both objective and subjective design guidelines, and their ignorance of physical laws. To help the community assess progression, we introduce a benchmark (Figure 1) focused on bicycle design, a problem that prominently features each of these challenges.

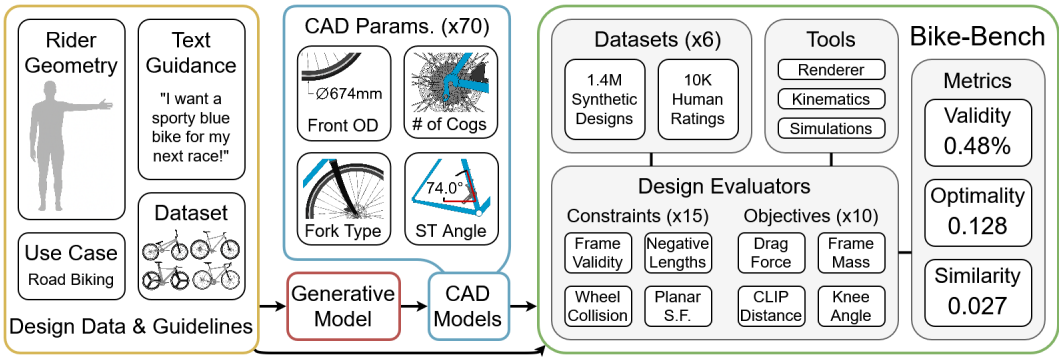

Figure 1: Bike-Bench is a set of evaluators and metrics, built to benchmark generative models' capability to synthesize parametric bicycle designs satisfying a variety of objectives and constraints.

As a constrained, multiphysics-guided, and human-centered design benchmark, Bike-Bench fills several important voids in the existing space of generative modeling benchmarks. Unlike many image, sketch, or 3D model-based benchmarks [19, 20] used for design, Bike-Bench evaluates exclusively parametric designs. This forces benchmarked models to synthesize designs with an exact mapping to a Computer-Aided-Design (CAD) file, essentially guaranteeing a precise, ready-to-manufacture bicycle model, rather than a more abstract image, sketch, or point cloud with less-clear downstream value. Bike-Bench also differs from CAD datasets, which rarely have features to evaluate multiphysics objectives, hard constraints, or human-centered preferences [21, 22, 23]. In contrast, Bike-Bench is comprised of 10 multidisciplinary design objectives and 15 design constraints, all of which revolve around a rich set of design evaluators. These evaluators leverage datasets of physics simulations, a geometry engine and renderer, and even a dataset of real human-sourced design assessments.

Unlike classic generative modeling benchmarks, Bike-Bench is more than just an exercise in the maximization of distributional similarity. To measure the practical design capabilities of a generative model, Bike-Bench also calculates aggregate validity and optimality scores over sample sets – metrics that are frequently overlooked, despite their outsized importance in engineering design [24]. For a model to simultaneously succeed in all of Bike-Bench's metrics, it must strategically deviate from the data manifold to improve performance and constraint satisfaction in targeted ways. Bike-Bench supports tabular models, LLMs, direct design optimization algorithms, and hybrid algorithms that transcend boundaries, providing a unique opportunity to benchmark previously disparate algorithms side-by-side. Our benchmarking results suggest that generative models, and particularly LLMs, have extensive room for growth and improvement in constrained engineering design problems. Key contributions of this paper are summarized as follows:

- We introduce a synthetic dataset of 1.4M bicycle designs represented as tabular data, SVG, PNG, and XML files for Computer-Aided-Design (CAD) software. This dataset supports a variety of design generation tasks such as text-to-cad, image-to-cad, and parametric-to-image generation.
- We introduce a dataset of 10K human-sourced ratings of bicycle designs. These ratings model subjective human assessments of designs' usability.
- We introduce Bike-Bench, a benchmark evaluating validity, optimality, and similarity of generative-model-synthesized bicycle designs. The benchmark features 25 design evaluations spanning aerodynamics, structural mechanics, ergonomics, aesthetics, geometric feasibility, and human perception of usability.
- We benchmark a variety of design generation techniques, including a large language model (OpenAI o4-mini), multiple tabular generative models, optimization-augmented generative models, and both gradient-based and heuristic design optimization algorithms.

2 Background

We provide brief overview of engineering design benchmarks for generative models, as well as benchmarks constructed to evaluate multiple classes of algorithms. A background on data-driven bicycle design is included in Appendix A.1, as well as a motivation for the selection of bicycle design as a benchmark problem.

Engineering Design Benchmarks for Generative Models: Bike-Bench joins a limited set of engineering design benchmarks designed primarily for evaluation and comparison of generative models. DesignQA [17], for example, evaluates the ability of LLMs to answer questions about engineering standards and requirements. Engineering design is highly data-limited [25], and most datasets are not configured as benchmarks [26]. Regardless, several engineering design datasets can be configured to support performance-aware generative model training [27, 28, 29, 30]. While a few of these datasets feature numerous objectives and constraints [31, 32], most only evaluate one principal objective. Bike-Bench is the first dedicated generative modeling benchmark focused on design synthesis under numerous multidisciplinary objectives and constraints.

Benchmarks for multiple classes of algorithms: Few benchmarks directly compare generative models to entirely different classes of algorithms, because such comparisons may superficially seem 'unfair.' For example, even though generative models more and more frequently compete with optimization algorithms [33], the comparison may not seem fair because optimization algorithms are allowed to call evaluators, while purely data-driven models must infer from datasets. We contend that such comparisons are rigorous, important, and particularly timely. In practice, generative AI models are regularly placed in direct competition with other classes of algorithms. Direct competition across classes of models can be easily appreciated in the historical development of deep-learning-based computer vision over algorithmic methods [34] or LLMs over classical language models [35, 36]. Benchmarks that embrace this direct competition will help practitioners understand shortcomings and capacity for growth. Such benchmarking is gaining traction [37] particularly in fields like engineering design [38], where the adoption of generative AI has faced considerable resistance, due to the strength of alternative methods [18]. Bike-Bench helps fill this need for engineering design benchmarks that accommodate different classes of algorithms.

3 Datasets

Bike-Bench introduces several new datasets, many of which are used to train the predictive models that Bike-Bench uses to evaluate many facets of bicycle design. Bike-Bench also consolidates and adapts two existing datasets: The **BIKED dataset** [39], comprised of 4500 human-design bikes, serves as the basis for Bike-Bench's distribution-modeling component. Bike-Bench uses a custom 70-parameter subset of salient design features from BIKED's full representation, which are described in Appendix A.5. The **FRAMED dataset** [40], focusing on structural mechanics of bike frames, features nearly 15K designs simulated using Finite Element Analysis (FEA) under multiple loading cases. FRAMED supports Bike-Bench's structural mechanics evaluators and frame validity analysis. We refer readers to the respective papers for more details on BIKED and FRAMED.

3.1 Dataset: 1.4M Synthetically-Generated Bicycle Designs

We introduce a new dataset of 1.4 million synthetically-generated bicycle designs. For each design, we provide parametric data, images, XML files and, for convenience, CLIP embeddings of the images. We generate this dataset using a **customized renderer** running parallelized bare-bones instances of the BikeCAD software.

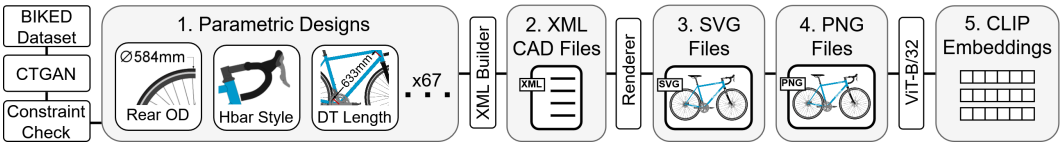

Figure 2: Overview of the synthetic data generation pipeline.

Synthetic datapoints are generated using CTGAN [41], a tabular GAN model specializing in realistic synthetic data generation. Each design is procedurally mapped to to a Computer-Aided-Design

(CAD) file. Though CAD file structure varies widely from software to software, BikeCAD files adopt an XML format and typically consist of several thousand key-value pairs. Our custom renderer is then invoked to render the CAD-XML files into Scalable Vector Graphic (SVG) files, which are then converted to rasterized Portable Network Graphics (PNG) files. Finally, we calculate the CLIP embedding for each generated image using OpenAI's 'clip-vit-base-patch32', a ViT-B/32 model provided by HuggingFace. The intermediates (XML, SVG, PNG) are retained and included in the dataset. Bike-Bench primarily uses this dataset to train a direct embedding model from Bike-Bench's parametric design space to the CLIP embedding space, allowing for aesthetics-based design evaluation. However, the dataset can conceivably support a variety of different predictive modeling and design generation tasks, such as text-to-CAD, image-to-CAD, parametric-to-image, etc. Using the parallelized pipeline, the full 1.4M designs took approximately four days to generate on an ordinary workstation and occupy approximately 750GB of storage, compressed. Users can easily run our pipeline to generate more data at their discretion.

3.2 Dataset: 10K Human-Sourced Bicycle Ratings

We also introduce a dataset of 10,350 human ratings of perceived bicycle usability. These ratings were collected through a rigorous selection procedure. Rather than spreading ratings evenly across the 4500 designs in BIKED, we instead opted to focus on a subset of 200 designs and collect approximately 50 ratings per bike, allowing us to more easily perform statistical significance tests and calculate rater agreement. This selection process for the 200 designs aimed to achieve a uniform distribution across Bike-Bench's 70 key parameters. To target users with cycling experience and familiarity with bike usability, only countries with at least a 35% weekly riding frequency were included, limiting eligibility to 14 countries based on a survey across 28 countries in 2022 [42]. Prolific was used to crowdsource participants for rating bicycle designs, where participants answer "Yes" or "No" to the question: "Does this bicycle look easy to use?". Binary (yes/no) assessments were selected to avoid the flaws of continuous ratings [43]. The 200 sampled bikes were divided into four groups of 50, ensuring diversity within each group. This reduced rater fatigue and helped maintain consistent rating quality throughout the session. After the completion of the rating process, a minimum rating time threshold of 90 seconds was set to remove participants who swiped too quickly, giving little time to fully evaluate each bike design. This reduced the number of valid respondents to an average of 50.75 per group, resulting in 10,350 valid ratings. Bike-Bench configures this dataset as a regression problem (predicting the proportion of raters who would consider a bike 'easy to use'). However, a classification dataset of consensus assessments was also extracted through statistical testing, as described in Appendix A.2.2.

3.3 Dataset: 4K Cyclist Aerodynamics Simulations

We finally introduce a new dataset comprised of 4,000 3D models of cyclists in various poses and their steady-state drag force under a $10\frac{m}{s}$ relative headwind, as evaluated by a computational fluid dynamics (CFD) simulation. The 3D cyclist models are parameterized by six anthropometric measurements sampled from an approximate model of published population statistics [44, 45, 46, 47, 48], and five parameters defining the interface points between the cyclist and the bike, which are sampled from BIKED.

4 Evaluation Criteria

In this section, we discuss Bike-Bench's design evaluation criteria. Design problems almost always feature competing objectives. Befittingly, Bike-Bench's many objectives and constraints are challenging to optimize in synchrony. For example, the most trivial ways to optimize performance objectives will typically involve violating constraints. In this section, we discuss how Bike-Bench's various datasets and tools are combined to evaluate a medley of functional attributes for bicycle design.

Geometric Feasibility: In data-driven parametric design, highly expressive parametrization schemes often have a drawback in 'allowing' invalid configurations. Bike-Bench's 70 design parameters enable considerable expressive flexibility, but can be selected or generated in ways that result in a variety of geometrically-infeasible designs – a common pattern among generative models trained on BIKED [49]. These designs commonly feature overlapping or disconnected components, parts with negative dimensions, and frames that violate the triangle inequality, for example. This infeasibility

must therefore be identified and evaluated through constraint checks. We have compiled a set of 12 closed-form geometric constraint checks (listed in the full list of evaluations in Appendix A.4), which is significantly expanded from a set released in BIKED [39]. We supplement these closed-form checks with a data-driven feasibility predictor trained to identify bike frames with more complex geometric issues using FRAMED's binary classification data.

Structural Soundness The structural soundness of a bike's frame plays an important role in its comfort, power-efficiency, and safety. Bike frames are generally preferred to be as rigid as possible, which limits dissipation of energy due to the flexure of the frame when pedaling [50]. We calculate several 'composite' structural performance indicators – planar, transverse, and eccentric compliance, as well as frame weight – which are all considered objectives to minimize. We also calculate planar and eccentric safety factor constraints, stating that the yield stress of the frame material must be at least 1.5 times the maximum stress incurred during planar or eccentric loads. We use a predictive model trained on FRAMED [40] to these structural attributes. Similar predictive models trained on FRAMED have been shown to closely align with real-world experimental simulations [40].

Aerodynamics Aerodynamic drag reduction is a principal consideration in competitive cycling, but is generally beneficial in all cycling settings. In general, the drag force directly incurred by the bicycle is much smaller than the drag incurred by the cyclist's body. Thus, drag is primarily minimized through repositioning of the cyclist, which is a function of the rider's anatomical geometry and the positioning of bicycle components that the rider interfaces with (saddle, handlebars, and pedals). Therefore, aerodynamic performance is a factor of both a bicycle and an associated rider. To quantify aerodynamics, we first calculate the interface points between the bike and the cyclist based on the rider, then call a predictive model trained on the aerodynamics data for a drag force estimate.

Ergonomics In addition to playing a significant role in aerodynamics, cyclist geometry also plays an important role in ergonomics. Examining the range of angles experienced by certain key joints during regular cycling activity [51] is a simple indicator for ergonomic fit. We develop a kinematics solver which calculates a rider's maximum knee angle, hip angle and shoulder angle and compares them to published target values for various types of cycling [51]. This solver yields three ergonomic objective scores which are calculated as a function of the bicycle design, the cyclist's anthropomorphic measurements, and the cyclist's use case (road biking, mountain biking, or commuting).

Human Perception of Usability Consumers frequently use subjective criteria to evaluate products. Increasingly, products are marketed as 'user-friendly,' with usability transitioning from a historic perception as a 'bonus' feature to a requisite expectation for many modern products [52]. Additionally, manufacturers emphasize their products as 'people-oriented,' particularly for items that involve direct body contact or require manual operation [53]. Thus, it is important to focus not only on the performance of the bicycle but also on its perceived usability. Accounting for human-centered criteria allows insights gathered from a broad population to help align the design of the bicycles, evolving them to meet growing consumer expectations. We evaluate the perceived usability of bikes using a regression model trained on our new dataset of 10K human-sourced usability scores.

Aesthetics Bicycles are regularly used for fashion, lifestyle, and other cultural statements, elevating the bicycle from a mere transportation product to a versatile tool for nuanced individual expression [54]. As such, the bicycle design industry is heavily influenced by aesthetics, individual preferences, and customization. Since individualized design customization requires significant design throughput, customization is an area where data-driven design methods can particularly excel. Thus, generative models would ideally be able to conditionally generate designs that are customized for an individual user's aesthetic preferences, as communicated through a reference image or a text prompt. To enable this sort of conditioning, we train a predictive model to directly estimate CLIP embeddings of parametric bicycle designs using the dataset of 1.4M synthetically-generated bike designs. By calculating CLIP embeddings for parametric designs, we can quickly evaluate their similarity to text prompts or reference images.

5 Benchmarking Metrics and Procedure

The design objectives detailed above yield a set of 10 design objectives and 15 design constraints, which are explained in detail in the Appendix A.4. However, to compare the performance of different generative models, we desire a concise set of summary metrics to evaluate sample sets. This section describes Bike-Bench's metrics and general benchmarking procedure.

5.1 Metrics

Though there are countless possibilities for generative model evaluation metrics [24], we select three: validity, optimality, and similarity, which we describe below. Unlike other benchmarks with singular objectives, models benchmarked on Bike-Bench should be compared using all three scoring metrics.
Validity (Constraint Satisfaction Rate): To evaluate a model's constraint satisfaction ability, we measure the proportion of generated designs that simultaneously satisfy all constraints.
Optimality (Hypervolume Metric): We calculate the hypervolume metric over any designs that simultaneoulsy satisfy all constraints. The Hypervolume metric is a staple of multi-objective optimization literature which calculates the overall multi-objective optimality of a set of designs. Hypervolume generally benefits from larger and more diverse solution sets. More details on hypervolume metric and its calculation are included in the Appendix.
Similarity (Maximum Mean Discrepancy): To capture a model's ability to match the manifold of existing designs, we evaluate distributional similarity to a subset of the BIKED dataset withheld from the model during training. We select Maximum Mean Discrepancy, a common kernel-based statistical discrepancy measure used to compare sets of samples. Similarity remains an important facet of design synthesis since it helps enforce two goals: 'realism' and 'design space coverage' [24].

This medley of performance scores roughly capture a model's ability to match the manifold of existing designs, avoid constraint violation and optimize design objectives. The function of validity and optimality are self-explanatory for design – maximally useful designs must be both feasible and optimized. The role that manifold similarity plays in design synthesis is less obvious. Just as human designers guide their decision-making based on the space of existing designs, design-space modeling can also assist algorithmic design synthesis approaches, a key motivation for the use of generative models. Statistical similarity serves to enforce two desirable attributes in generative models for design. The first is 'design realism' – what does it mean to be a bike and not something entirely different, say, a wheelbarrow? The second desirable attribute is 'design space coverage, ' encouraging models to learn more than just a small niche subspace of the overall breadth of bicycle designs, expanding their utility.

5.2 Benchmarking Procedure

Bike-bench's principal benchmark evaluates conditional generative modeling, providing 10,000 conditional cases (text prompt, rider anthropometric dimensions, use case, and reference image) and collecting one sample per condition from the generative model. Distributional similarity, constraint satisfaction rate over all 15 constraints, and hypervolume metric over all 10 evaluation objectives are evaluated for the 10,000 resulting designs and their respective conditions.

We also provide an unconditional generative modeling benchmark, where generative models are trained to satisfy a singular fixed condition, then are queried for 1,000 samples. To make scores less sensitive to specific conditions, 10 unique conditional cases are given, models are trained 10 times, and mean scores over these 10 conditional cases are reported.

To avoid bias, models are trained only on the 3600 training samples from the original 4500 human-designed bicycles, and not on synthetically-generated designs (The synthetic dataset is used exclusively for evaluation of aesthetics scores). Models are allowed to train on either a mixed-datatype or continuous version of the data. Designs generated by models trained on continuous data are mapped back to a mixed-datatype representation prior to evaluation – booleans are decided using rounding, while one-hot categorical data is decided using argmax.

6 Models Benchmarked

Our principal goal in the benchmarking of baselines is to span a variety of types of design generation procedures ranging from LLMs to tabular generative models to optimization algorithms. Because we do not focus on extensively comparing methods within specific classes, we limit our benchmarking to just one or a few methods per class. The various features and attributes of the models tested are summarized in Appendix A.3.
Large Language Models To gauge the general performance of Large Language Models on bike-bench, we provide an LLM interface with prompts, text descriptions of design parameters and objectives, and a data loader, as described in Appendix A.3.5. We benchmarked OpenAI's o4-

mini-2025-04-16. At the time of benchmarking, May 10, 2025, this was the highest performing model on ArtificalAnalysis's LLM intelligence index, a blended evaluation suite of seven individual evaluations [55, 56, 57, 58, 59, 60]. To limit cost and benchmarking time, we used the standard inference effort setting, despite 'high effort' being used for leaderboard scores. To comply with the model's context window, we exposed it to 10 example conditions (text, rider geometry, and use case), 10 example designs, and the associated 10 sets of 25 evaluation scores resulting from the condition-design pairs. The model was then provided 10 new conditions (or one in the unconditional setting) and asked to generate a csv file with 10 new designs. During our initial experimentation, it was found that the model tended to duplicate designs. Thus, our prompts include an explicit instruction that each generated design must be unique. The model was queried repeatedly and independently to populate each of 10k + 10k designs used for conditional and unconditional evaluation.

Tabular Generative Models: Bike-bench, like many tabular data problems, features ordinal, boolean, and categorical features. Although categorical features can be one-hot encoded into booleans and booleans can be treated as continuous, such simplifications can have significant drawbacks. Training deep learning models directly on categorical data generally requires special modifications, particularly in categorical data generation settings where gradients must be propagated through this discretization. Several generative modeling formulations which learn to generate mixed continuous and categorical data have been proposed over the year. We benchmark two popular models, CTGAN and TVAE, proposed in [41]. These models are unconditional, do not consider design performance, and are benchmarked principally for their distribution-learning capabilities.

Optimization-Augmented Generative Models: In generative modeling problems related to engineering design, design optimality and diversity are commonly valued in addition to distributional similarity. One mechanism to encourage the generation of high-performing, diverse, and realistic designs it to augment a model's classic statistical similarity objective with an auxiliary loss factoring in the optimality and diversity of a generated data batch. This involves calling the evaluation functions during training, much like the loop of an optimization algorithm. Since this loss serves to steer the generative model toward more optimal and constraint-satisfying solutions, this procedure effectively constitutes an augmentation-based optimization process. This optimization-augmentation has been demonstrated alongside GAN models using a Determinantal Point Process (DPP)-based auxiliary loss [61]. We benchmark such a performance-augmented GAN and further extend this DPP-based training to a VAE and DDPM. For the DDPM, we test this optimization-augmentation both in training and during inference, as guidance (labeled with '-G'). All optimization-augmented benchmarks are labeled with the 'OA-' prefix in our benchmarks. More details on our implementation are included in the Appendix.

Optimization: We evaluate both gradient-based and heuristic optimization. We benchmark plain aggregation-based gradient descent [62] and gradient-based Exact Pareto Optimal (EPA) search [63], both implemented in the LibMoon optimization library [64]. All gradient-based algorithms are applied to one-hot encoded data with constraints enforced using a 1000x weighted penalization function. We also benchmark a mixed-variable adaptation of the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [65], a staple evolutionary optimization algorithm, as implemented in pymoo [66]. Whereas the gradient-based algorithms benefit from the efficiency of gradient information, the mixed-variable NSGA-II learns directly on the mixed-variable data without obfuscation caused by one-hot encoding or continuous relaxation of constraints. Optimization algorithms are only benchmarked on the unconditional test case. All optimization algorithms are run 10 times, one for each of the 10 test conditions. Rather than forcing a final design set of 1k per condition, the final population (featuring less than 1k designs in each algorithm) is taken as the final design set.

7 Benchmarking Results

In this section, we present and analyze the benchmarking results. Both the unconditional and conditional generation results are compiled in Table 1, with averages over model classes presented in Table 2. All models are evaluated on validity (constraint satisfaction rate), optimality (hypervolume metric), and similarity (maximum mean discrepancy).

7.1 Discussion

The relatively low validity rate of the dataset baseline $(2.7/2.75\%)$ may initially seem surprising. Structural safety factor is a key driver of this low validity: 96% and 84% of dataset designs fail the

Table 1: Full Benchmark Results for unconditional and conditional generation. Models are separated by class (LLM / Tabular-GM / OA-GM / Optimizater), with class-averaged results appearing in Table 2. To provide a grounded reference point for validity and optimality scores, we include a row, *dataset*, italicized, which benchmarks the trivial 'generative' procedure of random selection from the training dataset. High validity and optimality scores are best, while low similarity scores are best. The best non-dataset scores in every metric are bolded.

	Unconditional Generation			Conditional Generation		
	Validity (↑)	Optimality (↑)	Similarity (↓)	Validity (↑)	Optimality (↑)	Similarity (↓)
dataset	*2.70%*	*0.2428*	*0.0006*	*2.75%*	*0.4151*	*0.0007*
o4-mini	0.28%	0.0115	0.5026	0.63%	0.2615	0.3460
TVAE	0.05%	0.0334	0.0400	0.00%	0.0000	0.0365
CTGAN	0.58%	0.1496	**0.0252**	0.87%	0.4090	0.0324
OA-VAE	80.55%	0.2712	0.1081	**81.65%**	**0.4357**	0.0693
OA-GAN	24.16%	0.3228	0.1066	22.34%	0.4152	0.1791
OA-DDPM	54.24%	0.2750	0.0740	40.46%	0.3801	0.0590
OA-DDPM-G	40.57%	0.2957	0.0261	40.30%	0.4120	**0.0253**
Grad-Agg	98.55%	0.3648	0.3824	-	-	-
EPO	92.00%	0.2625	0.2286	-	-	-
NSGA-II	**100.00%**	**0.4293**	0.5051	-	-	-

Table 2: Benchmark results averaged by model class

	Unconditional Generation			Conditional Generation		
	Val. (↑)	Opt. (↑)	Sim. (↓)	Val. (↑)	Opt. (↑)	Sim. (↓)
Large Language Model	0.28%	0.0115	0.5026	0.63%	0.2615	0.3460
Tabular Generative Models	0.32%	0.0915	**0.0326**	0.44%	0.2045	**0.0345**
Optimization-Augmented Models	49.99%	0.2912	0.0787	**46.18%**	**0.4108**	0.0832
Optimization Algorithms	**96.85%**	**0.3522**	0.3720	-	-	-

planar and eccentric safety factor constraints, respectively. This systematic under-engineering is one of the main sources of bias in the BIKED dataset, arising from the fact that many designers never adjust tube thickness parameters, since they are not visually prominent in the BikeCAD user interface [40]. This presents an interesting test for generative models, which, to satisfy structural constraints, generally have to deviate from the norms of the dataset to systematically thicken certain tubes. This exemplifies the principal challenge of Bike-Bench: **To simultaneously succeed in validity, optimality, and similarity, models must subtly and tactfully deviate from the distribution of the dataset to achieve design goals.**

In general, performance across classes of models is much more variable than performance within classes. We discuss the performance of each class of models one by one. Before doing so, we note a few general comments. First, conditional and unconditional scores should not necessarily be compared. For example, because hypervolume increases with set size given sufficient diversity, we expect significantly higher optimality scores in conditional generation due to the 10x larger set size. We generally expect validity and similarity scores to be relatively similar in unconditional versus conditional generation. Second, for models with very low validity, hypervolume is calculated over very few designs, thus may be somewhat sensitive.

Large Language Model: The rich history of bicycle design has extensive textual documentation and undoubtedly appears in common LLM training datasets. Consequently, LLMs can theoretically leverage contextual knowledge of the design parameters and the evaluation criteria to gain an edge over other classes of models. However, they must overcome key disadvantages in their lack of specialization for tabular data, as well as limited context windows, which may prevent them from observing the full dataset. OpenAI's o4-mini, the lone LLM tested, was a clear underperformer across all metrics, scoring the worst in distributional similarity in all unconditional generation metrics and in similarity during conditional generation. Notably, it did outperform tabular generative models in both validity and optimality during conditional generation, suggesting that its contextual knowledge or exposure to evaluation criteria gave it a slight edge over pure distribution-learning models. o4-mini's

performance indicates, unsurprisingly, that LLMs have room for growth in constrained engineering design problems and tabular data domains.

Tabular Generative Models: The tabular generative models trained are unconditional and agnostic to design constraints and objectives. Consequently, they displayed some of the lowest validity and optimality scores, but achieved consistently strong similarity scores. Notably, they underperformed the random dataset sampling baseline in every metric, highlighting the importance of performance-aware training, compared to pure performance-agnostic distribution learning.

Optimization-Augmented Generative Models: Optimization-Augmented Generative models achieved the most balanced set of scores. They demonstrated strong validity and optimality, second only to direct optimization in unconditional tests. Their similarity scores consistently placed them ahead of optimization algorithms and the LLM, but typically behind tabular generative models. The Optimization-Augmented VAE (OA-VAE) achieved high validity scores at the expense of similarity, achieving worst- or second-worst-in-class similarity scores. OA-GAN was generally underwhelming across all metrics. Finally, the guided diffusion model (OA-DDPM-G) was particularly outstanding in its distributional similarity performance, while still maintaining validity and optimality scores commensurate with most other optimization-augmented models. Guided diffusion outperformed unguided conditional diffusion in nearly all metrics.

Optimization: Optimization algorithms cannot be easily applied in the conditional generation case, thus are limited in their application. This is because they assume evaluation functions are deterministic based solely on the optimization's design parameters. Since many of Bike-Bench's evaluation functions depend on conditional information, this conditioning must be static for optimization to succeed. As the only class of non-data-driven methods, optimization is naturally a weak performer in similarity. Noting these limitations, optimization is the clear winner in both validity and optimality for unconditional design generation. The performance of optimization algorithms indicate considerable room for growth for generative models in validity and optimality.

8 Conclusion and Future Directions

We proposed Bike-Bench, a constrained engineering design benchmark for generative models, comprised of 10 multiphysics and human-centered objectives and 15 geometric and safety constraints. We benchmarked a variety of design generation procedures including an LLM, tabular generative models, optimization algorithms, and optimization-augmented generative models. Our benchmarking results suggest that tabular generative models supersede other baselines in similarity performance, optimization outperforms in validity and optimality, optimization-augmented generative models reach a strong middle ground, and LLMs generally underperform.

We encourage further benchmarks of generative models, optimization algorithms, and design generation procedures that transcend boundaries. LLMs with larger context windows or tabular data adapters, Vision Language Models (VLMs), foundation models for tabular generation, and optimization-augmented mixed-variable models may be of particular interest. Lastly we advocate for the development of more engineering design benchmarks which, alongside Bike-Bench, may hopefully expand the frontier of generative AI to successful engineering design automation and beyond.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[4] S. Oh, Y. Jung, S. Kim, I. Lee, and N. Kang, "Deep generative design: integration of topology optimization and generative models," *Journal of Mechanical Design*, vol. 141, no. 11, p. 111405, 2019.

[5] L. Wang, Y.-C. Chan, F. Ahmed, Z. Liu, P. Zhu, and W. Chen, "Deep generative modeling for mechanistic-based learning and design of metamaterial systems," *Computer Methods in Applied Mechanics and Engineering*, vol. 372, p. 113377, 2020.

[6] R. Wu, C. Xiao, and C. Zheng, "Deepcad: A deep generative network for computer-aided design models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6772–6782.

[7] A. H. Nobari, A. Srivastava, D. Gutfreund, K. Xu, and F. Ahmed, "Link: Learning joint representations of design and performance spaces through contrastive learning for mechanism synthesis," *Transactions on Machine Learning Research*, 2024.

[8] L. Regenwetter, A. H. Nobari, and F. Ahmed, "Deep generative models in engineering design: A review," *Journal of Mechanical Design*, vol. 144, no. 7, p. 071704, 2022.

[9] American Council of Engineering Companies Research Institute, "Economic assessment of the engineering + design services industry," 2024, available at: https://www.acec.org/wp-content/uploads/2024/10/ACEC-Research-Institute-2024-Economic-Assessment-Forecast-Final-1.pdf.

[10] Royal Academy of Engineering, "Engineering economy & place," 2022, available at: https://raeng.org.uk/media/o0hkijvw/engineering-economy-and-place-report-june-2023.pdf.

[11] W. Beitz, G. Pahl, and K. Grote, "Engineering design: a systematic approach," *Mrs Bulletin*, vol. 71, p. 30, 1996.

[12] D. M. Buede and W. D. Miller, *The engineering design of systems: models and methods*. John Wiley & Sons, 2024.

[13] H. Saunders, "Mechanical engineering design," 1985.

[14] F. Faruqi, Y. Tian, V. Phadnis, V. Jampani, and S. Mueller, "Shaping realities: Enhancing 3d generative ai with fabrication constraints," *arXiv preprint arXiv:2404.10142*, 2024.

[15] A. Gaier, J. Stoddart, L. Villaggi, and S. Sudhakaran, "Generative design through quality-diversity data synthesis and language models," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2024, pp. 823–831.

[16] L. Regenwetter, G. Giannone, A. Srivastava, D. Gutfreund, and F. Ahmed, "Constraining generative models for engineering design with negative data," *Transactions on Machine Learning Research*, 2024.

[17] A. C. Doris, D. Grandi, R. Tomich, M. F. Alam, M. Ataei, H. Cheong, and F. Ahmed, "Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation," *Journal of Computing and Information Science in Engineering*, vol. 25, no. 2, p. 021009, 2025.

[18] R. V. Woldseth, N. Aage, J. A. Bærentzen, and O. Sigmund, "On the use of artificial neural networks in topology optimisation," *Structural and Multidisciplinary Optimization*, vol. 65, no. 10, p. 294, 2022.

[19] D. Ha and D. Eck, "A neural representation of sketch drawings," *CoRR*, vol. abs/1704.03477, 2017. [Online]. Available: http://arxiv.org/abs/1704.03477

[20] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[21] S. Kim, H.-g. Chi, X. Hu, Q. Huang, and K. Ramani, "A large-scale annotated mechanical components benchmark for classification and retrieval tasks with deep neural networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 175–191.

[22] S. Koch, A. Matveev, Z. Jiang, F. Williams, A. Artemov, E. Burnaev, M. Alexa, D. Zorin, and D. Panozzo, "Abc: A big cad model dataset for geometric deep learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9601–9611.

[23] K. D. Willis, Y. Pu, J. Luo, H. Chu, T. Du, J. G. Lambourne, A. Solar-Lezama, and W. Matusik, "Fusion 360 gallery: A dataset and environment for programmatic cad construction from human design sequences," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–24, 2021.

[24] L. Regenwetter, A. Srivastava, D. Gutfreund, and F. Ahmed, "Beyond statistical similarity: Rethinking metrics for deep generative models in engineering design," *Computer-Aided Design*, vol. 165, p. 103609, 2023.

[25] *Dated: Guidelines for Creating Synthetic Datasets for Engineering Design Applications*, ser. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, vol. Volume 3A: 49th Design Automation Conference (DAC), 08 2023. [Online]. Available: https://doi.org/10.1115/DETC2023-111609

[26] J. H. Panchal, M. Fuge, Y. Liu, S. Missoum, and C. Tucker, "Special issue: Machine learning for engineering design," *Journal of Mechanical Design*, vol. 141, no. 11, p. 110301, 10 2019. [Online]. Available: https://doi.org/10.1115/1.4044690

[27] P. Wollstadt, M. Bujny, S. Ramnath, J. J. Shah, D. Detwiler, and S. Menzel, "Carhoods10k: An industry-grade data set for representation learning and design optimization in engineering applications," *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 6, pp. 1221–1235, 2022.

[28] S. Yoo and N. Kang, "Deepwheel: Generating a 3d synthetic wheel dataset for design and performance evaluation," *arXiv preprint arXiv:2504.11347*, 2025.

[29] E. Whalen, A. Beyene, and C. Mueller, "Simjeb: simulated jet engine bracket dataset," in *Computer Graphics Forum*, vol. 40, no. 5. Wiley Online Library, 2021, pp. 9–17.

[30] S. Hong, Y. Kwon, D. Shin, J. Park, and N. Kang, "Deepjeb: 3d deep learning-based synthetic jet engine bracket dataset," *Journal of Mechanical Design*, vol. 147, no. 4, 2025.

[31] N. J. Bagazinski and F. Ahmed, "Ship-d: Ship hull dataset for design optimization using machine learning," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 87301. American Society of Mechanical Engineers, 2023, p. V03AT03A028.

[32] A. Cobb, A. Roy, D. Elenius, F. Heim, B. Swenson, S. Whittington, J. Walker, T. Bapty, J. Hite, K. Ramani *et al.*, "Aircraftverse: a large-scale multimodal dataset of aerial vehicle designs," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 524–44 543, 2023.

[33] S. Shin, D. Shin, and N. Kang, "Topology optimization via machine learning and deep learning: a review," *Journal of Computational Design and Engineering*, vol. 10, no. 4, pp. 1736–1766, 2023.

[34] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Advances in computer vision: proceedings of the 2019 computer vision conference (CVC), volume 1 1*. Springer, 2020, pp. 128–144.

[35] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–35, 2021.

[36] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[37] B. Trabucco, X. Geng, A. Kumar, and S. Levine, "Design-bench: Benchmarks for data-driven offline model-based optimization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 21 658–21 676.

[38] Z. Masood, M. Usama, S. Khan, K. Kostas, and P. D. Kaklis, "Generative vs. non-generative models in engineering shape optimization," *Journal of Marine Science and Engineering*, vol. 12, no. 4, p. 566, 2024.

[39] L. Regenwetter, B. Curry, and F. Ahmed, "Biked: A dataset for computational bicycle design with machine learning benchmarks," *Journal of Mechanical Design*, vol. 144, no. 3, p. 031706, 2022.

[40] L. Regenwetter, C. Weaver, and F. Ahmed, "Framed: An automl approach for structural performance prediction of bicycle frames," *Computer-Aided Design*, vol. 156, p. 103446, 2023.

[41] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in neural information processing systems*, vol. 32, 2019.

[42] Ipsos, "Global advisor: Cycling across the world 2022," 2022, accessed: 2024-12-04. [Online]. Available: https://www.ipsos.com/en/global-advisor-cycling-across-the-world-2022

[43] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!" *Frontiers in ICT*, vol. 2, 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fict.2015.00013

[44] C. Gordon, R. Walker, I. Tebbetts, J. McConville, B. Bradtmiller, C. Clauser, and T. Churchill, *1988 Anthropometric Survey of US Army Personnel-Methods and Summary Statistics. Final Report*, 1989.

[45] M. A. McDowell, C. D. Fryar, and C. L. Ogden, "Anthropometric reference data for children and adults: United states, 1988-1994," April 2009. [Online]. Available: https://stacks.cdc.gov/view/cdc/5301

[46] M. A. McDowell, C. D. Fryar, R. Hirsch, and C. L. Ogden, "Anthropometric reference data for children and adults: United states, 1999-2002," July 2005. [Online]. Available: https://www.cdc.gov/nchs/data/ad/ad361.pdf

[47] C. D. Fryar, Q. Gu, and K. M. Flegal, "Anthropometric reference data for children and adults: United states, 2011-2014," August 2016. [Online]. Available: https://stacks.cdc.gov/view/cdc/40572

[48] C. D. Fryar, Q. Gu, and C. L. Ogden, "Anthropometric reference data for children and adults: United states, 2007-2010," October 2012. [Online]. Available: https://www.cdc.gov/nchs/data/series/sr_11/sr11_252.pdf

[49] L. Regenwetter and F. Ahmed, "Design target achievement index: A differentiable metric to enhance deep generative models in multi-objective inverse design," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 86236. American Society of Mechanical Engineers, 2022, p. V03BT03A046.

[50] D. G. Wilson and T. Schmidt, *Bicycling science*. MIT press, 2020.

[51] P. Burt, *Bike Fit 2nd Edition: Optimise Your Bike Position for High Performance and Injury Avoidance*. Bloomsbury Publishing, 2022.

[52] P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, *Usability Evaluation in Industry*. Boca Raton: Chapman and Hall/CRC, 2014.

[53] C. Cheng, "New trends of ergonomics and its importance in modern industrial design," in *International Conference on Enterprise Information Systems*, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:43386669

[54] M. Hoor, "The bicycle as a symbol of lifestyle, status and distinction. a cultural studies analysis of urban cycling (sub) cultures in berlin," *Applied Mobilities*, vol. 7, no. 3, pp. 249–266, 2022.

[55] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang *et al.*, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[56] L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi *et al.*, "Humanity's last exam," *arXiv preprint arXiv:2501.14249*, 2025.

[57] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "Gpqa: A graduate-level google-proof q&a benchmark," in *First Conference on Language Modeling*, 2024.

[58] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv:2103.03874*, 2021.

[59] M. Tian, L. Gao, S. Zhang, X. Chen, C. Fan, X. Guo, R. Haas, P. Ji, K. Krongchon, Y. Li *et al.*, "Scicode: A research coding benchmark curated by scientists," *Advances in Neural Information Processing Systems*, vol. 37, pp. 30 624–30 650, 2024.

[60] N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica, "Livecodebench: Holistic and contamination free evaluation of large language models for code," *arXiv preprint arXiv:2403.07974*, 2024.

[61] W. Chen and F. Ahmed, "Padgan: Learning to generate high-quality novel designs," *Journal of Mechanical Design*, vol. 143, no. 3, p. 031703, 2021.

[62] K. Miettinen, *Nonlinear multiobjective optimization*. Springer Science & Business Media, 1999, vol. 12.

[63] D. Mahapatra and V. Rajan, "Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6597–6607.

[64] X. Zhang, L. Zhao, Y. Yu, X. Lin, Y. Chen, H. Zhao, and Q. Zhang, "Libmoon: A gradient-based multiobjective optimization library in pytorch," *Advances in Neural Information Processing Systems*, 2024.

[65] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[66] J. Blank and K. Deb, "Pymoo: Multi-objective optimization in python," *Ieee access*, vol. 8, pp. 89 497–89 509, 2020.

[67] A. Sharp, *Bicycles & tricycles: an elementary treatise on their design and construction, with examples and tables*. Longmans, Green, 1896.

[68] O. Oke, K. Bhalla, D. C. Love, and S. Siddiqui, "Tracking global bicycle ownership patterns," *Journal of Transport & Health*, vol. 2, no. 4, pp. 490–501, 2015.

[69] D. S. De Lorenzo and M. Hull, "Quantification of structural loading during off-road cycling," 1999.

[70] L. B. Lessard, J. A. Nemes, and P. L. Lizotte, "Utilization of fea in the design of composite bicycle frames," *Composites*, vol. 26, no. 1, pp. 72–74, 1995.

[71] M. Godo, D. Corson, and S. Legensky, "An aerodynamic study of bicycle wheel performance using cfd," in *47th AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition*, 2009, p. 322.

[72] D. Covill, P. Allard, J.-M. Drouet, and N. Emerson, "An assessment of bicycle frame behaviour under various load conditions using numerical simulations," *Procedia engineering*, vol. 147, pp. 665–670, 2016.

[73] P. Oja, S. Titze, A. Bauman, B. De Geus, P. Krenn, B. Reger-Nash, and T. Kohlberger, "Health benefits of cycling: a systematic review," *Scandinavian journal of medicine & science in sports*, vol. 21, no. 4, pp. 496–509, 2011.

[74] T. L. Hamilton and C. J. Wichman, "Bicycle infrastructure and traffic congestion: Evidence from dc's capital bikeshare," *Journal of Environmental Economics and Management*, vol. 87, pp. 72–93, 2018.

[75] O. Edenhofer, *Climate change 2014: mitigation of climate change*. Cambridge University Press, 2015, vol. 3.

[76] L. Regenwetter, Y. Abu Obaideh, and F. Ahmed, "Multi-objective counterfactuals for design: A model-agnostic counterfactual search method for multi-modal design modifications," *Journal of Mechanical Design*, vol. 147, no. 2, 2025.

A Appendix A: Extended Details on Background, Datasets, Evaluation Criteria, Metrics, and Models

A.1 Background: Data-Driven Bicycle Design

Bicycle design is a complex engineering problem with a rich history of optimization through scientific and engineering innovation [67]. Bicycle design catalyzed numerous significant advancements in mechanical engineering and remains a significant research field to this day. Bicycles themselves revolutionized transportation and remain ubiquitous in today's society. In 2015, at least 42% of households globally owned a bike [68] and 35% of adults surveyed across 28 countries in 2022 rode a bike at least once a week [42]. Thanks to this history, the widespread continued use, and the hugely varied use cases and subjective preferences among users, the bicycle design space is rich and vibrant – ideal for data-driven methods.

Data and computation has played an increasingly important role in bicycle design science [69], predominantly focusing on computational simulation and optimization [70, 71, 72]. However, little published research had explored bicycle design-space exploration or big-data applications before the BIKED [39] dataset was released in 2021. BIKED demonstrated some of the first applications of design-space modeling, deep learning, and generative AI in bicycle design.

Data-driven tools are well-positioned to continue the legacy of innovation in bicycle design science. Indeed, data-driven design tools stemming from BIKED have already been integrated into professional bicycle design software (see www.bikecad.ca/ai). Thanks to the continued ubiquity of the bicycle in modern society, design innovation may yield more optimal or better-customized bikes, potentially increasing ridership. Such an increase in ridership could subsequently impart further societal impact by improving public health [73], traffic congestion [74], and climate change [75].

A.2 Datasets and evaluation criteria

A.2.1 Details on Data Collection from Human Subjects

All participants in our human subject data collection were adults (aged 18 or order), paid an hourly rate of 9 GBP. All participants were informed that their data may be used for the training and evaluation of ML/AI models and consented to this. The study passed relevant institutional review procedures without issue.

A.2.2 Building a Classification Dataset for Usability

Bike-Bench uses the binary (yes/no) ratings collected from human raters to predict the proportion of raters that would consider a design 'easy to use.' Rather than averaging the scores for each bike, yielding the distribution of scores shown in 3, ratings can alternatively be aggregated to form a simple classification dataset. To assess the overall consensus, factoring in a margin of error resulting from potential sampling inaccuracies, we can use a two-sided binomial test. Taking the population size of the group with the smallest number of valid users, 46, we evaluate the 99% confidence interval to be approximately $\mathbf{20\%}$. Thus, for any design with at least 70% consensus among respondents, we can be 99% confident in the consensus value. In other words, an average score of at least 0.7 confidently indicates that more than half of raters would consider the bike usable, whereas an average of at most 0.3 indicates that less than half of raters would consider the bike usable. Following these thresholds, **49** bikes are identified as usable and **51** as unusable, resulting in a total of **100 confident classification labels**.

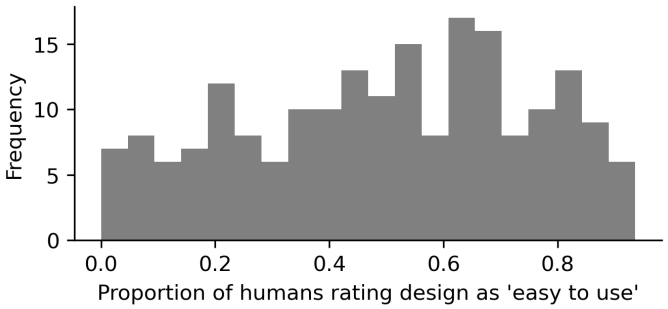

Figure 3: Distribution of average usability scores in Bike-Bench's human-sourced dataset

A.2.3 Limitations and Assumptions of Datasets and Evaluators

Bike-Bench makes numerous key assumptions that may cause significant inaccuracies in design evaluation. A few key assumptions and limitations are listed below, but this is not a comprehensive list. We encourage practitioners using models trained on Bike-Bench's data or using Bike-Bench's design evaluation tools for real-world bicycle design to undertake additional validation and verification steps.

- Design Representation: Bike-Bench's design representation is expressive enough to cover a wide variety of designs adhering to a 'conventional' diamond-frame bike. It does not span

entirely different topological layouts of bicycles, such as recumbent bikes, nor does it model complex suspension systems, such as mountain bike rear suspensions.

- Dataset bias: The original dataet of 4500 human-design bikes features certain biases, particularly in less visually-prominent features like tube thickess values. We refer readers to BIKED [39] and FRAMED [40] for more in-depth discussion.
- Geometric Feasibility: Our set of closed-form geometric checks is not comprehensive, and primarily covers commonly-seen 'mistakes' made by generative models. It is not a guarantee of feasibility.
- Structural Evaluation: Bike-Bench assumes isotropic material properties, and substitutes any inherently non-isotropic materials (bamboo, carbon, other) with steel for the purposes of evaluation. Real metal tubes nonetheless display some anisotropic properties, causing incurabilities. Simulations are based on a simplified 3D frame model which approximates tube joints and does not incorporate certain frame features, such as wheel cutouts. Detailed discussion and comparisons to experimental validation are included in FRAMED [40]
- Aerodynamics: The assessment of aerodynamic drag in a single speed and direct headwind and based on a basic cyclist model is a simplified assessment of bicycle aerodynamics. Real cycling undergoes a variety of different wind speeds at different angles, causing aerodynamic interactions in different flow regimes.
- Aesthetics: Although the trained parametric-to-CLIP embedding model suffices to capture subjective details (e.g. futuristic cyberpunk-style) [76], it struggles to capture technical details of bicycle components (e.g. number of chainrings or disk versus rim brakes).

A.2.4 Visualization of geometrically infeasible bikes

To give readers some intuition regarding the nature of geometric infeasibility, a few geometrically infeasible bikes are visualized in Figure 4.

Figure 4: Bike-bench features numerous closed-form constraint checks to identify common geometric infeasibilities such as disconnected or colliding components (left 3). Less common infeasibilities in frame geometry are flagged if the design causes errors during 3D reconstruction (right 3).

A.3 Models and metrics

A.3.1 Details of Hypervolume Calculation

In multi-objective design optimization, designs are often compared in the objective space – the space comprised of the set of all objective scores. One design "dominates" the other if it is superior in every single objective. To calculate the hypervolume metric of a sample set, we measure the hypervolume of the region in the objective space where a corresponding design would be dominated by any design in the sample set. To keep hypervolume bounded, it is typically calculated with respect to some reference point, which we select to be the combination of maximum (worst) objective scores in every unique objective, calculated by evaluating every design in BIKED with a random conditional configuration.

A.3.2 Standardization of computation budget

Although exact comparison of computation budget across classes of methods was deemed out of scope, we made reasonable efforts to standardize computational budget allocations. Excepting the closed-source o4-mini model, which was queried through API calls, all methods were benchmarked on the same workstation (5900x processor and 3090Ti GPU) and tuned to yield maximum performance with a maximum of 1 hour for optimization, model training and inference.

A.3.3 Summary of Model Features

A summary of select features and attributes of the various design generation methods tested is compiled in Table 3.

Table 3: Select features and attributes of various design generation methods

	Mixed Datatype	Calls Evaluators	Data-driven	Conditional Generation
o4-mini	✓	×	✓	✓
CTGAN	✓	×	✓	×
TVAE	✓	×	✓	×
OA-GAN	×	✓	✓	✓
OA-VAE	×	✓	✓	✓
OA-DDPM	×	✓	✓	✓
OA-DDPM-G	×	✓	✓	✓
Grad-Agg	×	✓	×	×
EPO	×	✓	×	×
NSGA-II	✓	✓	×	×

A.3.4 Details on performance-augmented model training:

The DPP-based auxiliary loss proposed in PaDGAN [61] uses a single aggregate design quality metric. To condense bike-bench's many objectives and constraints into a differentiable aggregate quality metric, we propose the following aggregation scheme:

$$s(x) = \sum_{i=1}^{n_o} \frac{o_i(x)}{w_{o_i}} + \sum_{i=1}^{n_c} g(\frac{c_i(x)}{w_{c_i}}) \tag{1}$$

All n_o objectives, o_i, and n_c constraints, c_i, are scaled by weighting parameters, w_{o_i} and w_{c_i}. Each parameter is set automatically based on the mean absolute value of the scores seen through a random pairing of each of the n_D points in the dataset, D, and randomly-sampled conditions, C_j:

$$w_{o_i} = \frac{\sum_{i=0}^{n_D}(|o_i(D_j, C_j)|)}{n_D}, \; w_{c_i} = \frac{\sum_{i=0}^{n_D}(|c_i(D_j, C_j)|)}{n_D} \tag{2}$$

Constraints are additionally fed through a nonlinear scaling function to push them across the constraint boundary and a safe margin away, without rewarding extreme constraint satisfaction:

$$g(x) = \begin{cases} \frac{\alpha e^{\beta x}}{\beta} & x \leq 0 \\ \alpha(x + \frac{1}{\beta}) & x \geq 0 \end{cases} \tag{3}$$

This continuous penalty function linearly increases for $x > 0$, and morphs into an increasingly gradual decay for $x < 0$, and also features a continuous first derivative. We use $\alpha = 10$ and $\beta = 10$ in our testing. Modulating these terms can encourage a model to focus more on objectives or constraints. For details on the DPP-based loss, given an aggregate quality function, we refer readers to [61].

A.3.5 LLM Prompts

A series of three prompts is given to the LLM model. Following the first prompt, the dataset descriptions (see Sec. A.5) and evaluation criterion descriptions (see Sec. A.4), along with a short introduction of each (see codebase) are given to the model. Following the second prompt, the condition, design, and scoring example data is given to the model. Following the third prompt, the test condition is given to the model. Conditional information is concatenated into a text string of the form: 'Rider Body Dimensions: Upper leg length - [ULL], Lower leg length - [LLL], Arm length - [AL], Torso length - [TL], Neck and head length - [HNL], Torso width - [TW]. Use Case: [Road Biking/Mountain Biking/Commuting]. Marketing Description: [Text Prompt]'. The prompts are listed as follows:

- I will ask you to create some bicycle designs. The bicycle designs are subject to a set of conditions: a text prompt, some rider dimensions, and a use case. Each design is defined by 70 variables, which I will describe. Some of these are categorical variables, and I will provide you with the possible values for these variables. Others are continuous. I will describe the design variables shortly. Designs are evaluated according to a set of 25 criteria. The first 10 are objectives, while the last 15 are constraints. Here are the design variable descriptions and the evaluation criteria:
- I will also provide a dataset of existing bicycle designs. These are useful as a reference point, because it may be difficult to satisfy constraints and objectives if you deviate too far from the space of existing designs. I will also provide a set of objective scores for these bicycle designs. Since these criteria are dependent on the conditional information, each bike design is paired with a randomized condition set for the purpose of evaluation. Thus, I will provide: 1) a text file with 10 rows of conditions, 2) a csv file with 10 rows of bikes and 70 columns of parameters, and 3) a csv file of 10 sets (rows) of 25 scores (columns). Here are these files. Please have a look to try to gain an understanding of the design space and the evaluation criteria.
- Having examined the existing designs and the evaluation of these designs alongside their associated conditions, I hope you have gained an understanding of the design space and design objectives. I will now ask you to create bicycle designs. Please deliberate on a strategy for creating high-performing designs that satisfy the constraints and objectives. I will provide you with 10 new conditions. Please create 10 unique bicycle design that satisfies the constraints and objectives. Important: You are not allowed to generate the same bike 10 times. Each design must be unique! Please provide the designs in a 10x70 csv file. No index or headers. Only the 10x70 values. Here are the 10 conditions. Remember! Every design must be unique.

A.4 Summary of Constraints and Objectives

Table 4: Summary of Bike-Bench's evaluation functions, classified as objectives or constraints. Inputs and evaluator type are also specified.

Category	Objectives	Constraints	Evaluator	Inputs
Geometric Feasibility	0	13	Closed-Form + Predictor	Bike
Structural Soundness	4	2	Predictor	Bike
Aerodynamics	1	0	Predictor	Bike + Rider
Ergonomics	3	0	Closed-Form	Bike + Rider + Use case
Human-Centered Usability	1	0	Predictor	Bike
Aesthetics	1	0	Predictor	Bike + Text/Image/Embedding

We describe Bike-Bench's constraints and objectives below. A summary of all evaluation functions is included in Table 4, and a text description of each is included as follows:

- Usability Score: [Objective] The predicted 'usability,' as rated by a human, with 0 being the most usable and 1 being the least usable. Predicted by a regression model trained on human-collected ratings.
- Drag Force: [Objective] The predicted drag force in N incurred by the cyclist in a 10 m/s headwind, as predicted by a regression model trained on computational fluid dynamics simulation data.
- Knee Angle Error: [Objective] The difference between the minimum knee angle of the cyclist and the optimal reference range. May include a penalty term if the rider's geometry is completely incompatible with the bike.
- Hip Angle Error: [Objective] The difference between the torso-to-upper-leg angle of the cyclist and the optimal reference range. May include a penalty term if the rider's geometry is completely incompatible with the bike.
- Arm Angle Error: [Objective] The difference between the torso-to-arm angle of the cyclist and the optimal reference range. May include a penalty term if the rider's geometry is completely incompatible with the bike.

- Cosine Distance to Embedding: [Objective] The cosine distance in the CLIP embedding space between the rendered bike image and the target text or image embedding.
- Mass: [Objective] The mass in kg of the bike frame, as predicted by a regression model trained on finite element analysis data.
- Planar Compliance: [Objective] A composite planar compliance score for the bike frame, as predicted by a regression model trained on finite element analysis data.
- Transverse Compliance: [Objective] A transverse compliance score for the bike frame, as predicted by a regression model trained on finite element analysis data.
- Eccentric Compliance: [Objective] A composite eccentric compliance score for the bike frame, as predicted by a regression model trained on finite element analysis data.
- Planar Safety Factor: [Constraint] Constraint quantified as 1.5 minus the safety factor under planar loading, as predicted by a regression model trained on finite element analysis data.
- Eccentric Safety Factor: [Constraint] Constraint quantified as 1.5 minus the safety factor under eccentric loading, as predicted by a regression model trained on finite element analysis data.
- Saddle height too small: [Constraint] Constraint indicating that the saddle height collides with the top of the seat tube.
- Seat post too short: [Constraint] Constraint indicating that the seat post doesn't reach the seat tube given the prescribed saddle height.
- Head tube lower extension too great: [Constraint] Constraint indicating that the down tube doesn't properly intersect with the head tube.
- Head tube length too great: [Constraint] Constraint indicating that the head tube is so short that the top tube and down tube intersect.
- Certain parameters must be positive: [Constraint] Constraint indicating that at least one parameter that should be strictly positive is negative.
- Chain stay should be greater than wheel radius: [Constraint] Constraint indicating that the chain stay length is smaller than the wheel radius, creating a collision.
- Chain stay should be greater than BB: [Constraint] Constraint indicating that the vertical drop from the rear axle to bottom bracket is greater than the chain stay length, creating an impossibility.
- Seat stay should be greater than wheel radius: [Constraint] Constraint indicating that the seat stay length is smaller than the wheel radius, creating a collision.
- Down tube must reach head tube: [Constraint] Constraint indicating that the down tube is too short to reach the head tube.
- The pedal shouldn't intersect the front wheel: [Constraint] Constraint indicating that the front wheel would intersect the pedal in its forward position, causing a collision when turning.
- The crank shouldn't hit the ground when it is in its lower position: [Constraint] Constraint indicating that the crank hits the ground during its rotation.
- RGB value should be less than 255: [Constraint] Constraint indicating that frame RGB values were set at higher than 255.
- Predicted Frame Validity: [Constraint] Constraint indicating some abstract issue with the frame, as predicted by a classification model trained to identify CAD models that failed to regenerate.

A.5 Summary of Design Variables

We include a text descriptions of all 70 parameters in the standard bike–bench design representation scheme. All lengths are measured in mm. All angles are measured in degrees. Parameter names are largely based on the BikeCAD internal XML labeling scheme.

The 70 variables are described as follows:

- ‘CS textfield’: [Continuous] The length of the chain stay tubes.
- ‘BB textfield’: [Continuous] Bottom bracket drop, measured as the vertical drop from the rear axle to the center of the bottom bracket. By convention, positive values imply the bottom bracket lies below the axle.
- ‘Stack’: [Continuous] The vertical distance from the top of the head tube relative to the bottom bracket.
- ‘Head angle’: [Continuous] The angle of the head tube with respect to horizontal, in degrees.
- ‘Head tube length textfield’ [Continuous] The length of the head tube.
- ‘Seat stay junction0’: [Continuous] The length along the seat tube from the top of the seat tube to the junction with the seat stays. By convention, this is measured to the center of the seat stays.
- ‘Seat tube length’: [Continuous] The length of the seat stay tubes.
- ‘Seat angle’: [Continuous] The angle of the seat tube with respect to horizontal.
- ‘DT Length’: [Continuous] The length of the down tube.
- ‘FORK0R’: [Continuous] Fork offset, measured as the perpendicular distance from the front axle to the head tube axis.
- ‘BB diameter’: [Continuous] The diameter of the bottom bracket
- ‘ttd’: [Continuous] Top tube outer diameter.
- ‘csd’: [Continuous] Chain stay outer diameter.
- ‘ssd’: [Continuous] Seat stay outer diameter.
- ‘dtd’: [Continuous] Down tube outer diameter.
- ‘Chain stay position on BB’: [Continuous] The distance along the length of the bottom bracket from its edge to the center of the chain stay tubes.
- ‘SSTopZOFFSET’: [Continuous] The offset from center plane of the bike of the joints connecting the seat stays to the the seat tube.
- ‘Head tube upper extension2’: [Continuous] The length from the top of the head tube to the junction with the top tube. By convention, this is measured to the center of the top tube.
- ‘Seat tube extension2’: [Continuous] The length from the top of the seat tube to the junction with the top tube. By convention, this is measured to the center of the top tube.
- ‘Head tube lower extension2’: [Continuous] The length from the bottom of the head tube to the junction with the down tube. By convention, this is measured to the center of the down tube.
- ‘SEATSTAYbrdgshift’: [Continuous] The distance along the center plane of the bike from the seat stay and seat tube junction to the seat stay bridge, if present on the bike.
- ‘CHAINSTAYbrdgshift’: [Continuous] The distance along the center plane of the bike from the outer rim of the bottom bracket to the chain stay bridge, if present on the bike.
- ‘SEATSTAYbrdgdia1’: [Continuous] The diameter of the seat stay bridge, if present on the bike.
- ‘CHAINSTAYbrdgdia1’: [Continuous] The diameter of the chain stay bridge, if present on the bike.
- ‘SEATSTAYbrdgCheck’: [Boolean] A boolean value indicating whether the seat stay bridge is present on the bike.
- ‘CHAINSTAYbrdgCheck’: [Boolean] A boolean value indicating whether the chain stay bridge is present on the bike.
- ‘Dropout spacing’: [Continuous] The distance between the rear dropouts.
- ‘Wall thickness Bottom Bracket’: [Continuous] The tube wall thickness of the bottom bracket.
- ‘Wall thickness Top tube’: [Continuous] The tube wall thickness of the top tube.
- ‘Wall thickness Head tube’: [Continuous] The tube wall thickness of the head tube.

- ‘Wall thickness Down tube’: [Continuous] The tube wall thickness of the down tube.
- ‘Wall thickness Chain stay’: [Continuous] The tube wall thickness of the chain stay.
- ‘Wall thickness Seat stay’: [Continuous] The tube wall thickness of the seat stay.
- ‘Wall thickness Seat tube’: [Continuous] The tube wall thickness of the seat tube.
- ‘Wheel diameter front’: [Continuous] The outer diameter of the front wheel.
- ‘RDBSD’: [Continuous] The difference between rear wheel outer diameter and bead seat diamater, roughly approximating the tire thickness.
- ‘Wheel diameter rear’: [Continuous] The outer diameter of the rear wheel.
- ‘FDBSD’: [Continuous] The difference between front wheel outer diameter and bead seat diamater, roughly approximating the tire thickness.
- ‘Display AEROBARS’: [Boolean] A boolean value indicating whether the bike has aerobars.
- ‘BB length’: [Continuous] The length of the bottom bracket.
- ‘Head tube diameter’: [Continuous] Head tube outer diameter.
- ‘Wheel cut’: [Continuous] The diameter of the cutout of seat tube for the rear wheel, if using an aerodynamic tube type.
- ‘Seat tube diameter’: [Continuous] Seat tube outer diameter.
- ‘bottle SEATTUBE0 show’: [Boolean] A boolean value indicating whether the bike has a bottle holder on the seat tube.
- ‘bottle DOWNTUBE0 show’: [Boolean] A boolean value indicating whether the bike has a bottle holder on the down tube.
- ‘Front Fender include’: [Boolean] A boolean value indicating whether the bike has a front fender.
- ‘Rear Fender include’: [Boolean] A boolean value indicating whether the bike has a rear fender.
- ‘BELTorCHAIN’: [Boolean] A boolean value indicating whether the bike has a chain (True) as opposed to a belt.
- ‘Number of cogs’ [Integer] The number of cogs on the rear wheel.
- ‘Number of chainrings’ [Integer] The number of chainrings attached to the crank.
- ‘Display RACK’: [Boolean] A boolean value indicating whether the bike has a rack.
- ‘FIRST color R_RGB’: [Continuous] The red component of the primary paint color of the bike.
- ‘FIRST color G_RGB’: [Continuous] The green component of the primary paint color of the bike.
- ‘FIRST color B_RGB’: [Continuous] The blue component of the primary paint color of the bike.
- ‘SPOKES composite front’: [Integer] If applicable, the number of composite spokes in the front wheel minus two (a value of 1 is a trispoke wheel).
- ‘SPOKES composite rear’: [Integer] If applicable, the number of composite spokes in the rear wheel minus two (a value of 1 is a trispoke wheel).
- ‘SBLADEW front’: [Continuous] If applicable, the width of the front wheel composite spokes.
- ‘SBLADEW rear’: [Continuous] If applicable, the width of the rear wheel composite spokes.
- ‘Saddle length’: [Continuous] The length of the saddle.
- ‘Saddle height’: [Continuous] The bertical distance from the saddle to the bottom bracket.
- ‘Down tube diameter’: [Continuous] The diameter of the down tube.
- ‘Seatpost LENGTH’: [Continuous] The length of the seat post.
- ‘MATERIAL’: [Categorical] The material of the bike frame. Possible values are: ‘ALUMINIUM’, ‘CARBON’, ‘STEEL’, ‘TITANIUM’, ‘BAMBOO’, ‘OTHER’.

- 'Head tube type': [Categorical] The style of head tube. Possible values are: '0', '1', '2', '3'. 0 is aerodynamic, while 1 and 2 are standard round tubes with no distinction in this representation scheme. 3 is a tapered head tube.
- 'RIM_STYLE front': [Categorical] The style of the front rim. Possible values are: 'DISC', 'SPOKED', 'TRISPOKE'. Despite the name, trispoke class implies composite spokes but does not necessarily imply three composite spokes.
- 'RIM_STYLE rear': [Categorical] The style of the rear rim. Possible values are: 'DISC', 'SPOKED', 'TRISPOKE'. Despite the name, trispoke class implies composite spokes but does not necessarily imply three composite spokes.
- 'Handlebar style': [Categorical] The style of the handlebars. Possible values are: '0', '1', '2'. 0 is a drop bar, 1 is a mountain bike bar, 2 is a bullhorn bar.
- 'Stem kind': [Categorical] The style of stem. Possible values are: '0', '1', '2'. 0 is a stem that features a sharp and immediate angle away from the head tube. 1 is a stem that features a sharp angle some distance away from the head tube. 2 is a stem that features a gradual angle away from the head tube after initially extending in line with the head tube.
- 'Fork type': [Categorical] The style of fork. Possible values are: '0', '1', '2'. 0 is a standard fork, 1 is a fork with mountain bike shocks, 2 is a time trial bike fork.
- 'Seat tube type': [Categorical] The style of seat tube. Possible values are: '0', '1', '2'. 0 is aerodynamic, while 1 and 2 are standard round tubes with no distinction in this representation scheme.

A.6 Ethics and Societal Impact

Bike-Bench aims to advance the capabilities of Generative AI models for engineering design. In general, AI has many positives and negatives, most of which are not particularly pertinent to this work. However, we would like to acknowledge some of the pros and cons of generative models for engineering design. A principal risk of AI in engineering, particularly generative models, is safety. Generative models are usually probabilistic and when human safety is in question, even small chances of failure are unacceptable. Designs and design decisions created by AI must be held to the same (or higher) scrutiny and engineering standards as human designs. Engineering design AI also stands poised to deliver notable societal benefits. By lowering the barrier of entry to design, it may democratize design, allowing more, and especially underprivileged individuals, to participate in the design process. By increasing design throughput, it may realize better products, more efficient design, and greater design customization for individuals. We advocate for the careful and ethical use of AI in engineering design and beyond.