International Design Engineering Technical Conferences 2024
IDETC2024
August 25–28, 2024, Washington, DC

IDETC2024-143634

# A PICTURE OR A THOUSAND WORDS: DESIGN DESCRIPTION CRAFTING TO REPLICATE HUMAN SIMILARITY JUDGMENTS IN LARGE LANGUAGE MODELS

**Matthew Keeler[1], Mark Fuge[1,*], Aoran Peng[2], Scarlett Miller[2]**

[1] Department of Mechanical Engineering, University of Maryland, College Park, MD
[2] School of Engineering Design and Innovation, Pennsylvania State University, University Park, PA

## ABSTRACT

*Well-studied techniques that enhance diversity in early design concept generation require effective metrics for evaluating human-perceived similarity between ideas. Recent work suggests collecting triplet comparisons between designs directly from human raters and using those triplets to form an embedding where similarity is expressed as a Euclidean distance. While effective at modeling human-perceived similarity judgments, these methods are expensive and require a large number of triplets to be hand-labeled. However, what if there were a way to use AI to replicate the human similarity judgments captured in triplet embedding methods? In this paper, we explore the potential for pretrained Large Language Models (LLMs) to be used in this context.*

*Using a dataset of crowdsourced text descriptions written about engineering design sketches, we generate LLM embeddings and compare them to an embedding created from human-provided triplets of those same sketches. From these embeddings, we can use Euclidean distances to describe areas where human perception and LLM perception disagree regarding design similarity. We then implement this same procedure but with descriptions written from a template that attempts to isolate a particular modality of a design (i.e., functions, behaviors, structures). By comparing the templated description embeddings to both the triplet-generated and pre-template LLM embeddings, we explore ways of describing designs such that LLM and human similarity perception might better agree. We use these results to better understand how humans and LLMs interpret similarity in engineering designs and assess the implications for how LLMs should be used for design evaluation in the future.*

## 1. INTRODUCTION

In the field of engineering design, the initial phase of the design process demands the generation of a diverse set of candidate concepts. Techniques like Design-by-Analogy (DbA) promote concept diversity and creativity by having engineers draw inspiration from both closely related and seemingly unrelated fields [1–4]. Exposing designers to a variety of inspirational stimuli encourages them to make design considerations that they otherwise would not have, thus widening the scope of the candidate design space. To be used effectively, DbA techniques require methods for evaluating the similarity between designs [3]. Having a metric for describing similarity allows the diversity of a set of designs to be measured and unique or novel designs to be identified [5**?** ]. However, similarity is a complex and multi-dimensional quality that does not easily lend itself to quantification.

Developing a metric for evaluating design similarity has been a heavily researched subject for decades. Traditional methods involve identifying shared features between concepts and scoring those features with domain expert knowledge [6]. However, these methods are non-generalizable because they assume that all relevant features have been identified and lend themselves to metric quantification. More recent work has proposed crowdsourcing humans to perform triplet comparison tasks between designs (Is Design A closer to Design B or Design C?) [7**?** ]. Unlike traditional methods, these triplet queries do not require identification or metric evaluation of shared features. With these labeled triplet queries serving as constraints on the acceptable placement of designs, a low-dimensional embedding can be constructed and similarity can be measured as a Euclidean distance between designs.

Triplet embedding methods have demonstrated success in grouping designs by similarity as perceived by humans. Unfortunately, labeling triplet queries is expensive and the number of triplets required grows combinatorially with the number of designs in the embedding [8]. The expensive nature of triplet collection has motivated exploration into ways of automating this process. Ultimately, if a model could be developed to generate an embedding with human-like similarity considerations, it would render triplet collection methods obsolete.

Presently, there do exist pretrained models which generate similarity embedding spaces—notably, Large Language Models (LLMs). LLMs take an input of sentences or paragraphs and perform word vectorizations to embed that text into a pretrained

[*] Corresponding author: fuge@umd.edu

latent space [9–11]. In this latent space, similar text is plotted closer together than dissimilar text, so the same methods used for triplet embeddings can be applied here.

While using LLMs to generate a similarity embedding space for engineering designs is an interesting notion, it is not without some major complications of its own. Firstly, for this method to work, the LLM would have to group design similarity like a human would. Secondly, engineering design concepts contain both visual and textual information; LLMs would need these multi-modal design concepts to be translated into a purely textual description.

In this paper, we are interested in the question of whether LLM-based embedding methods could be used to approximate triplet-based embedding methods. Specifically, we want to know if there is a way to better write a textual description of an engineering design such that the LLM perceives design similarity like a human would. To tackle these broad questions, this paper proposes the use of crowdsourced triplet comparisons as a tool for generating a baseline embedding of human-perceived similarity. With this baseline, we can measure the success of an LLM-embedding at replicating human-perceived similarity judgments. Then the inputted textual description of the designs can be changed and the improvement measured. Using this framework, our paper works to answer the following questions:

1. How do different LLM embeddings differ with each other with regard to the reported similarity between design descriptions?
2. How do different LLM embeddings differ with those constructed by human triplet labeling with regard to the reported similarity between designs?
3. In what ways should design descriptions provided to the LLM be modified such that the LLM-generated embeddings more closely match the triplet-generated embeddings?
4. In what ways does focusing descriptions of Function, Behavior, Structure, or Visual elements of a design impact how similar LLM-generated embeddings become to human embeddings?

The rest of the paper addresses these questions in §3-5. We provide empirical results from a variety of crowdsourced triplet-labeling and description-writing surveys. LLM-generated embeddings are compared with triplet-generated embeddings using a variety of similarity metric considerations. Furthermore, we analyze the differences in results produced with and without the use of specific writing templates.

## 2. RELATED WORK

Related work has been subdivided into the following sections: (1) work that explains or questions how humans perceive similarity between items, (2) work that proposes methods for evaluating similarity, and (3) work that uses AI to characterize or mimic human similarity judgments.

### 2.1 Human-Perceived Similarity

Researchers have studied what humans value when evaluating the similarity between items. In the field of psychology, similarity has been described as a linear combination of shared and distinct features [6]. Once features are identified, matching functions can be defined which measure correlations between similar and unique features [6]. However, defining these features requires large pools of crowdsourced survey responses and reported features from participants vary widely depending on the relative context of the items [6, 12]. This makes it difficult to generalize what features humans value.

In the field of engineering design, attempts to better understand human similarity judgments have been largely pursued to enable Design-by-Analogy (DbA) strategies [1–4]. DbA has proven to be a powerful tool for encouraging novel designs in concept generation [2]. DbA requires that designs be decomposed into defining characteristics so that similarity evaluation and analogy retrieval can occur. The decomposition is most commonly done by breaking a design into solution-neutral subfunctions [1, 2]. Other methods include functional-behavioral-structural decompositions, which capture a wider net of design modalities [3, 13, 14]. However, past research suggests that humans value function-based similarity over behavioral and structural forms when performing similarity comparisons [3]. It is noteworthy that comparisons in this existing work have been performed on diverse design datasets. In our paper, we are interested in the modalities of human similarity judgment on a dataset of designs with the same core function, as well as how those modalities can be explained in text to improve LLM embedding disagreement.

### 2.2 Evaluating Similarity

To use design similarity in strategies like DbA, there must be a system for quantifying similarity. Traditionally, similarity is measured by defining a set of relevant features and then either counting shared features [6] or assigning metric values using domain knowledge [15]. With a strategy for associating metric values to features, similarity can be measured through vector distances [15**?** ], or by distributional divergence measures like Kullback-Liebler [16].

Each of these methods require identification of relevant features and a system for assigning metric values to those features. In practice, it may be impractical to limit a dataset of multi-faceted designs to a pool of defining features which may or may not lend themselves to easy quantification. Recent work has proposed avoiding feature identification by collecting similarity judgments in the form of triplet comparisons (Is Design A closer to Design B or Design C?) [7, 17**?** ]. In this implementation, similarity judgments are all relative to two reference designs. With a labeled list of triplet queries, a low-dimensional embedding can be created and similarity can be identified by Euclidean distances within the embedding [7, 17**?** ].

Our paper will use the methods triplet collection and embedding to explain and quantify human similarity judgments. We are interested in what design aspects humans key into when comparing similarity without prompting information and how and if pretrained LLMs can replicate those comparisons through engineered descriptions. Triplet methods are particularly well-suited for modeling human-perceived similarity, in that human raters can order triplet queries without explicitly identifying a list of scoring criteria.

### 2.3 Using AI to Replicate Human Similarity

Understanding human perceived-similarity is a powerful tool for novelty estimation and DbA, however creating a similarity embedding from crowdsourced data is expensive and time-consuming. Recent advances in AI have motivated work toward developing automated protocols for understanding and reproducing human similarity judgments. In work similar to this paper, researchers compared crowdsourced triplet embeddings with computational similarity scores [18] and with visual 3D image comparison software [**?** ]. Researchers highlighted the areas of disagreement between human and computational or AI similarity assessment, but no protocols were proposed to replicate human perception.

In terms of developing models to mimic human similarity scores, researchers have trained visual ImageNet software on crowdsourced similarity data [19]. In the field of psychology, researchers have equipped LLMs with algorithms designed to promote human-like inductive reasoning [20]. Related work has been done to validate similarity judgments on large documents from Latent Dirichlet Allocation (LDA) against human-perceived similarity [**?** ]. However, validating similarity scores from text-based algorithms has not been widely explored in the field of engineering design. Our paper seeks to better understand wheter LLMs can be used for the purposes of replicating human-perceived similarity as it relates to interpreting engineering designs.

## 3. METHODS

In order to compare perceived similarity between LLMs and humans, a baseline embedding must be established that represents human similarity groupings among designs. This baseline embedding is constructed by minimizing the violations of human-provided triplet orderings. With this embedding, disagreement in similarity assessment can be identified by comparing the Euclidean distances within the human-baseline and LLM-generated embedding spaces. With a method in place for comparing similarity assessment, modifications can be made to the descriptions provided to the LLM and the benefit can be observed.

### 3.1 Milk-Frother Dataset

A dataset of ten hand-drawn designs from an undergraduate Pennsylvania State University design course was used in this work [**?** ]. To form this dataset, students were asked to develop a device with the goal of frothing a basin of milk. Students were asked to provide a rough drawing of their device as well as a brief descriptive title. All ten sketches are provided in Figure 1.

### 3.2 Triplet Collection

To capture human-perceived similarity, human raters were tasked with ordering designs in the form of triplet queries (Is Design A closer to Design B or Design C?). Existing work has shown that humans can consistently find the closer match between two candidate pairs of designs [7**?** ]. This is contrasted with the

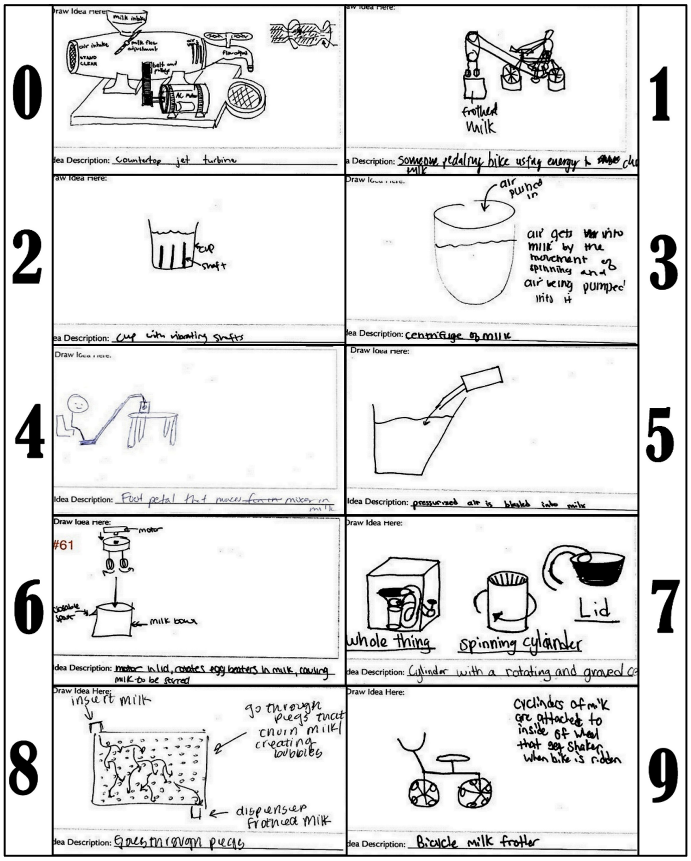

0
Draw Idea Here:
milk intake
air intake
STAND CLEAR
milk flow adjustment
air
Flow adjust
belt and pulleys
AC Motor
Idea Description: Countertop jet turbine
1
frothed milk
Idea Description: Someone pedaling bike using energy to ~~[illegible]~~ churn milk
2
Draw Idea Here:
cup
shaft
Idea Description: cup with vibrating shafts
3
air pushed in
air gets ~~[illegible]~~ into milk by the movement of spinning and air being pumped into it
Idea Description: centrifuge of milk
4
Draw Idea Here:
Idea Description: Foot petal that moves ~~forth~~ mixer in milk
5
Draw Idea Here:
Idea Description: pressurized air is blasted into milk
6
Draw Idea Here:
motor
#61
closable spout
milk bowl
Idea Description: motor in lid, rotates egg beaters in milk, causing milk to be stirred
7
Draw Idea Here:
whole thing
spinning cylinder
Lid
Idea Description: Cylinder with a rotating and graved ce
8
Draw Idea Here:
insert milk
go through pegs that churn milk creating bubbles
dispenser frothed milk
Idea Description: Goes through pegs
9
Draw Idea Here:
cyclinders of milk are attached to inside of wheel that get shaken when bike is ridden
Idea Description: Bicycle milk frotler

FIGURE 1: EXAMPLE OF STUDENT-GENERATED MILK FROTHER SKETCH

more difficult task of articulating or quantifying the closeness of pairwise comparisons—which requires development of domain-specific scoring criteria [21].

Using the Milk-Frother dataset, a survey conducted by Ahmed *et al.* [**?** ] was created to generate all 360 combinations of choosing a single reference design and two candidate designs from the ten available drawings. Raters were asked which of the two candidate designs was more similar to the reference design, and a triplet label was formed to express their chosen ordering. Raters were given no instructions for how they should evaluate design similarity. After every 18 triplets were labeled, raters were asked to provide a brief description of the thought process behind their triplet orderings.

This survey was completed by 15 members of Pennsylvania State Univesity's engineering design labs. Members consisted of 11 undergraduate students, 3 graduate students, and 1 faculty member. From this pool of 15 ratings per triplet query, a majority-vote was taken to represent a single consensus ordering for every triplet. This majority-voted list of all 360 possible triplet combinations was used to develop the human baseline embedding.

### 3.3 Embedding the Triplet Information

To make use of the human-labeled queries, a low dimensional embedding was constructed with the triplet orderings appearing as constraints on the acceptable placement of designs.

This paper used the Generalized Non-Metric Multidimensional Scaling (GNMDS) algorithm to generate an embedding from the majority-voted triplet pool. GNMDS deconstructs triplet orderings into two pairwise distances—one for each reference-candidate pair [**?** ]. The triplet label indicates which of these pairwise distances should be larger, and an inequality is established between the two pairwise distances. GNMDS associates a slack variable for each of these inequality constraints. If the constraint is satisfied, no slack is needed, but if the constraint is violated, the magnitude of violation is recorded as a Euclidean distance. The algorithm finds the embedding that best minimizes the sum of all triplet slack variables.

There are other popular algorithms that can also generate an embedding from triplet labels, including Stochastic Triplet Embedding and Crowd Kernel Learning [7, 17]. Both of these methods employ a probabilistic model for describing how well a triplet is modeled in the current embedding. While these methods are valid, GNMDS was chosen due to its intuitive and straightforward approach for evaluating embedding accuracy. GNMDS not only aims to reduce triplet violations, but considers the severity of the violation and encourages large margins of label satisfaction.

Optimally choosing the embedding dimension for an algorithm like GNMDS is non-trivial. When humans perform triplet rating tasks, the dimension of the feature-space that they use to perform similiarity comparison is unknown [3, 22]. If the embedding dimension is too small, some triplet orderings may be impossible to satisfy that would have been possible in a larger space. If the embedding dimension is too large, the algorithm gains freedom in design placement at the risk that triplet constraints no longer tightly bound the embedding space. This could lead to pairwise distances that misrepresent the human ratings, despite low triplet violation counts. For this paper, a 10-dimensional space was chosen following observations from Keeler and Fuge [8], which displayed larger reconstruction errors for under-approximation of the 'true' dimension-space. Ten dimensions were chosen to encourage a larger dimensional embedding, with the limit being the number of Milk-Frother designs to be plotted.

### 3.4 Large-Language Models

After creating the human-perceived similarity space, the next step was to select pretrained Large Language Models (LLMs) to generate embeddings based on text descriptions of the Milk-Frother designs.

For this paper, two LLMs were chosen from the Huggingface sentence-transformers library: GloVe and MiniLM-L12-v2 (MiniLM) [23]. GloVe (Global Vectors) is an unsupervised learning algorithm for obtaining vector representations for words. GloVe is an older model which is trained on word-word co-occurrence and develops a probabilistic model for those co-occurrences [9]. GloVe was chosen as a baseline due to its widespread use and documentation. MiniLM-L12-v2 is a newer Siamese Network built on a BERT-like pretrained model [10, 11]. Whereas GloVe only models co-occurrence of whole words, MiniLM incorporates the context around words and can parse unrecognized words [11]. MiniLM-L12-v2 was chosen to represent a more state-of-the-art model.

### 3.5 Similarity Metrics

To compare the embeddings produced by GNMDS on the human-labeled triplets and the LLMs on text-descriptions of the Milk Frother designs, three similarity metrics were considered: triplet violations, pairwise distances, and centroidal distances. Each of these metrics is used to describe how well an embedding agrees with the human-perceived similarity data.

In the first metric, all 360 possible triplet queries were collected and ordered based on the Euclidean distances in the embedding. These triplet orderings were then compared with the majority-voted orderings from the human raters. The number of triplets in the embeddding that violate their corresponding ordering in the human pool were recorded. This metric is the most powerful in that it directly measures the embedding's adherence to the human-provided responses. An embedding with no violations should directly reflect human relative similarity judgments. However, such an embedding is impossible, as the majority-voted pool has some triplets which violate the transitive property, and cannot be satisfied at the same time as another triplet with the same three designs.

The second metric compares the pairwise distances between the embeddings. Pairwise distances provide the best metric for quantifying perceived similarity between two designs. However, unlike the triplet violations, there is no easy way to obtain pairwise similarity scores from human raters directly. Instead, the GNMDS embedding must be used to describe the pairwise distances perceived by humans. In reality, however, the GNMDS embedding is not a perfect representation of the human triplet responses. Even if it were possible to satisfy all triplets, there can exist multiple embeddings that satisfy the same triplets but exhibit different pairwise distance matrices. However, it is the objective of GNMDS to minimize triplet violations with a large margin of certainty. For this paper we assume the GNMDS embedding is reasonably accurate in modeling the pairwise distances perceived by human raters.

To compare the pairwise distances across different embeddings, there must also be some way of normalizing them. GNMDS is a 10-dimensional embedding, while GLoVe and MiniLM are 300- and 384-dimensional embeddings, respectively. In reality, the absolute pairwise distances are irrelevant for describing perceived similarity. Only the relative similarity—where a pairwise distance ranks with respect to all other distances in the embedding—is meaningful for evaluating whether two designs are similar. For this reason, pairwise distances in each embedding are sorted and given 0-100 percentile scores for their rank among other distances in all subsequent uses in this paper. With this system, pairwise distances from the LLM-generated embeddings can be compared with the human-motivated GNMDS embedding. If a distance has a percentile of 0 (closest distance) in the LLM embedding and a percentile of 100 (largest distance) in the GNMDS embedding, this would indicate large disagreement in percieved similarity.

The final metric describes the centroidal distances in the embedding compared to the human baseline. This metric is motivated by existing work that uses triplet embeddings to rank design novelty [5**?** ]. Novel items are identified as designs that are far from the centroid of the embedding space [**?** **?** ]. This metric is used to identify how well two embeddings agree on perceived novelty. As a consequence, this metric can be used to determine if an LLM embedding technique can be a substitute to the more expensive job of triplet embedding for the purposes of novelty estimation.

Like the pairwise distance metric, the centroidal distance metric assumes that the GNMDS embedding reflects human-perceived novelty with reasonable accuracy. Also like the pairwise distance metric, there must be some system for comparing the centroidal distances across different embedding mediums. This could be done by comparing a list of designs ordered by their novelty—however, sorted lists do not account for the magnitudes of the centroidal distances. Two embeddings might similarly identify the same design as the most novel with disagreement in *how much* more novel it is. Instead, we describe novelty disagreement by the novelty error metric described in Keeler and Fuge [8]. This metric normalizes each centroidal distance by the largest distance in an embedding to create a relative novelty. It then takes the absolute difference of these normalized centroidal distances between embeddings and reports the average disagreement between relative novelty.

### 3.6 Description Crafting

Having developed a strategy for evaluating the similarity between embeddings, the final step is to develop protocols for writing text descriptions of the Milk-Frother sketches. Both the GloVe and MiniLM models work with paragraphs of text as inputs. To serve as baseline, seven graduate students from Pennsylvania State University's engineering design labs were asked to write text description paragraphs for all ten Milk-Frother sketches. Authors were given no additional instructions beyond explaining the design in sentence form. These text descriptions were given to the GloVe and MiniLM models to create two LLM-generated embeddings for every author.

After the instructionless data was collected, the same authors were given two templates for crafting design descriptions. The first template was motivated by the work in Gero [13] which categorizes design descriptions into three facets: functional, behavioral, and structural. Functions are the high-level goals of the design and detail *what* the design and its sub-assemblies are designed to accomplish. Behaviors are the physics and actions that describe *how* the functions are performed. Finally, structures describe the actual network of components and spatial relationships which work to perform the functions. This breakdown of design description schemas is widely used to capture the many facets of engineering designs in text [14, 24].

In the template, authors were asked to develop a bulleted list of high-level functions found in a particular Milk-Frother sketch. Authors were told to limit their functions to desired outcomes only, and to avoid mention of how the functions were accomplished. With this bulleted list of functional descriptions, authors were then asked to create a paired list of behavioral descriptions that described the actions taken to perform each function. In this section, authors were encouraged to use technical physics vocabulary and to avoid mention of the components of the sub-assemblies. Finally, authors were asked to create another paired list of structures which described the functional components at play. Authors were encouraged to include descriptions of spatial relationships but to avoid mention of actions or behaviors. These steps were taken in an effort to isolate each of the design description facets. Example results from one of the authors for the Milk-Frother sketch can be found in Table 1. The exact instructions for this template can be found in Appendix A.

For the second template, authors were asked to limit their description to three words or short phrases. This template was designed to capture the design components that authors immediately fixate on. Authors were told that two of these words should describe immediately eye-catching physical components of the design and that one of these should describe an action that the design is performing. Descriptive physics jargon was encouraged. Example results from one of the authors for the Milk-Frother sketch can be found in Table 1. The exact instructions for this template can be found in Appendix A.

## 4. RESULTS

### 4.1 Pre-Template Model Comparison

Figure 2 displays the similarity diagnostics for the three embeddings before human authors were given instructions on crafting their design descriptions. Descriptions provided without a template will be referred to as free-form (FF). The three monochrome heatmaps represent the Euclidean pairwise distances found in the respective embedding. The red heatmap denoted 'Human10' corresponds to the baseline 10-dimensional GNMDS embedding. The color bar on the right of the heatmap shows the relative percentile rank of each pairwise distance as described in §3.5. Light colors represent a small pairwise distance and thus a high-degree of perceived similarity. For example, the human raters believe designs one and four are very

Figures/Pre_Temp_Avgs.pdf

**FIGURE 2: PRE-TEMPLATE AVERAGED EMBEDDING RESULTS**

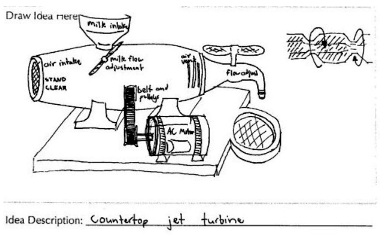
| Image | Template 1 | | | Template 2 |
|---|---|---|---|---|
| | Function | Behavior | Structure | Limited Word |
|  Draw Idea Here<br>milk intake<br>air intake<br>STAND CLEAR<br>milk flow adjustment<br>air vent<br>flow adjust<br>belt and pulleys<br>AC Motor<br>Idea Description: Countertop jet turbine | Funnels an intake of poured milk. Aerates milk inside of the container. Mechanically mixes the air-milk solution. Dispenses the frothed milk product. | Intake milk flow is gravity-fed and corralled before flow stream is redirected. Air is sucked inside, introducing turbulence. Air-milk solution is stirred through mechanical rotation. Outflow is guide by rear pressure and pushed through exit. | A conical intake at the top of the device with a lever-valve at the base. An air intake at the rear of the device that is covered by mesh-grating with air vents cut into the sides. A rotor attached to a belt and pulley system and an AC motor. A nozzle-faucet with an adjustable shut-off valve. | Turbine, motor, aeration |

TABLE 1: EXAMPLE DESCRIPTIONS FOR EACH TEMPLATE

similar—which from Figure 1 are both human-operated stirring devices.

Below the red heatmap are the triplet violation and novelty order metrics for the embedding. As mentioned in §3.5, the GNMDS baseline cannot satisfy all human-provided triplets and instead violates 17.78% of the human labels. About 6% of these unsatisfied triplets can be attributed to transitive violations in the labeled triplet pool which are impossible to satisfy. Below the reported triplet disagreement is a sorted list based on increasing centroidal distance or novelty. According to the GNMDS embedding, human raters perceive design nine as the most novel and design seven as the least novel.

The monochrome heatmaps to the right of the 'Human10' heatmap are for embeddings generated by the MiniLM and GloVe models, respectively. These embeddings are found from averaging the FF description embeddings for all seven authors. On the average FF description, the MiniLM model violates 33.06% of the human-rated triplets and the GloVe model violates 36.39%. Each LLM model has about double the violations of the GNMDS model.

Triplet disagreement between the three models manifests in differences in the novelty ordering. Both the MiniLM and GNMDS embeddings agree that design nine is the most novel. From there, the MiniLM model is fairly close to GNMDS in its reported novelty order. There are some notable points of contention, however, including the placement of design one. The MiniLM model believes design one is the second-most novel, whereas the human-based embedding places it in the bottom half. Despite only having 3% more violations than the MiniLM model, the GloVe model has a large amount of disagreement with the human and MiniLM models.

On the far right are heatmaps which represent the difference between the 'Human10' and LLM heatmaps. Dark colors on these heatmaps represent pairwise distances with large disagreement between the two models. In the top plot, red squares indicate a positive percentile difference—meaning the human model believes the designs are less similar than the LLM model suggests. For example, the human model believes designs 5 and 8 are much more dissimilar than the MiniLM model suggests.

By following an individual design in the heatmap, the model disagreement can be observed as it pertains to a single design. If a design's pairwise distant squares are consistently one color, this indicates a design that is very controversial between the two models with regard to its relative similarity. For the MiniLM model, design one consistently presents blue squares—indicating that the MiniLM model tends to label this design as more dissimilar to its neighbors than the human model does. This manifests in design one having a much higher perceived novelty in the MiniLM model.

The heavy disagreement over design one between the LLMs and human model is an interesting case which might reveal some of the rationale behind the LLMs embedding placements. The humans and MiniLM models agree that design nine is the most novel item. From Figure 1 we see that design nine has milk containers affixed to the spokes of a bicycle. The controversial design one also contains a bicycle, but this bicycle only acts as a power source to a more commonly exhibited stirring mechanism. While the humans group design one as being very similar to other stirring devices, the LLM models believe the design to be unique. We hypothesize that this is because the LLM is putting heavy emphasis on the word 'bicycle', which is an unusual word among the designs. Human raters are able to contextualize unusual structures by their role in the greater design, while LLM models are more heavily influenced by unique vocabulary.

Using a similar analysis, it can be seen the GloVe model consistently underrates the novelty of design zero compared to the human and MiniLM models. Design zero is a counter-top turbine design with many moving parts in a relatively complex orientation. However, the components themselves and underlying physics are not unique among the designs, and thus the description vocabulary is non-unique. The GloVe model is not able to leverage the context surrounding the commonplace vocabulary like the MiniLM model, and so it places its embedding solely on the individual word choices. This is especially problematic for the placement of design zero which needs many words to fully describe, diluting the perceived importance of any unique words. In this converse example to the design one scenario, humans may perceive a machine of non-unique components as unique, whereas LLMs are unable to capture the *compositional* novelty as effectively. Because the MiniLM model is better at leveraging contextual information and produces closer similarity scores to humans across all metrics, the MiniLM model will be used in all remaining experiments within the paper.

### 4.2 Template Effects on Triplet Violations

Figure 3 shows the triplet violation percentages for each author's descriptions. There are seven description templates displayed: 'FF' refers to free-form descriptions where authors were given no crafting instructions; 'Func', 'Beha', and 'Stru' denote the functional, behavioral, and structural template descriptions detailed in §3.6; 'Lim' refers to the limited-word template; 'FBS-Av' refers to averaging the functional, behavioral, and structural descriptions; while 'FBS-St' refers to stacking the descriptions into one large composite description.

The left side of Figure 3 displays the full distribution of embedding results for each author. In this plot, the template trends can be observed as they relate to a given author. On the right side of Figure 3 are boxplots that display the quartile values of the seven authors' responses across each template. Boxplot whiskers are found by adding 1.5 times the interquartile range.

From the boxplots, it can be seen that the median triplet violation scores for the individual FBS descriptions are about the same as for the template-free FF descriptions (∼ 38%). However, the individual FBS descriptions have much larger variances than the FF description results. The boxplot for the limited-word template is fairly close the FF template plot, with each quartile metric presenting a slightly higher violation percentage. From this data, it does not appear that it is advantageous to limit descriptions to a single design facet. However, it is interesting to see that the lowest observed triplet violations came from an author's structural and limited-word descriptions. While for the whole distribution, these templates do not present better results than the 'FF' description, it is possible for comparatively brief descriptions to outshine long multifaceted write-ups.

Both the 'FBS-Avg' and 'FBS-St' boxplots present lower median triplet violation percentages than the FF plot. In particular, the FBS-Avg has the best performing median (≈ 36%). More importantly, the FBS-Avg presents a much smaller variance than any other template. While there is not enough data to perform any significant hypothesis testing, it can also be seen from the line-plot that every author's violation score improves or remains the same from their 'FF' to 'FBS-Avg' descriptions with the exception of author two. These results suggest that the FBS-Avg template is at worst a minor improvement to the average FF description. A controlled template with a slightly lower median violation score is preferable to an uncontrolled and unpredictable free-form prompt.

The violation scores from templates that include multiple design facets appear lower and less variable than those from single-facet templates. There could be several reasons for this. The first would be that humans consider multiple design facets when performing similarity comparisons. From the survey results, raters reported factoring in actions, structures, and functions to justify their triplet orderings [**?** ]. It would make sense that if humans considered multiple modalities, that the LLM would need descriptions in each of these modalities to better replicate the similarity results. This hypothesis is corroborated by the fact that the FBS-Avg embedding has a much lower median than any of the individual FBS descriptions which comprise it. The FBS descriptions as individual pieces are unlikely to replicate human similarity groupings, but together they expose the LLM to multiple design modalities to promote more informed similarity comparison.

The second reason for the improvements might be explained by how the LLM models place the designs in an embedding. The placement is mostly determined by a balancing of word vectors that exist in the description. If a description contains a unique word that is not similar to a word in any other description, this could heavily influence its placement. The single-facet FBS descriptions tend to be shorter in length and are thus more influenced by unique words than the FBS-Av combination. Therefore, the decrease in variance could be attributed to longer descriptions being more robust to violations resulting from an author's idiosyncratic language.

### 4.3 Template Effects on Novelty Error

Figure 4 shows the full-distribution line plot and boxplot data but for the novelty error deviation between the MiniLM and GNMDS models. Like for the triplet violations, the median novelty error values for the FF and single-facet FBS templates are about the same. However, in the full-distribution line plot there are some key differences. One would expect that authors who had the highest triplet violation scores for a given template would also have the highest novelty error scores. This is the case for the most part with some key exceptions. Author two—who had the worst violation score for the 'Func' template—notably had the median novelty error for this same template.

Although the exact reason for the disagreement between the similarity metrics is unclear, this phenomenon is possible due to the varying degree of impact that triplets have on an embedding. While the triplet violation score is still the most powerful for describing how often humans and LLMs agree, the triplet violation score does not factor in the severity of a particular violation. If the LLM incorrectly orders a triplet containing three very similar designs, this will not have a large impact on the overall embedding layout. On the other hand, if the LLM incorrectly orders a triplet with two similar designs and one widely dissimilar design, this will greatly warp the embedding placements. The proportion of satisfied triplets should not be used blindly when describing the embedding's success at capturing human perception.

Moving on to the multi-faceted templates in Figure 4, there is more evidence of disagreement between the two similarity metrics. For the triplet violations, the 'FBS-Avg' template has a noticeably smaller variance and the lowest median of any template. For the novelty error, now the 'Lim' template has the lowest median and variance. While there is not enough evidence to suggest that the 'Lim' template is more suitable for novelty estimation, there is noticeable disagreement between the two metrics within this template in particular.

The 'Lim' template by nature has the shortest descriptions and thus is the most heavily influenced by idiosyncratic language. In the Figure 3 triplet violation plot, this can be seen where authors two and three have much worse triplet violation scores compared to the other authors. However, in the corresponding Figure 4 novelty error plot, these authors are both below the 3rd quartile error value. These results suggest that while limited gut-instinct vocabulary may not be effective for replicating all similarity comparisons, it might still be effective at capturing the general novelty. This might be because the LLM does not have enough modalities to correctly order hard-to-label but low-impact triplets. Instead, limited descriptions boil down designs to their most unique and defining features, effectively prioritizing distinct and space-defining triplets.

## 5. DISCUSSION, LIMITATIONS, AND FUTURE WORK

Overall, the LLM embeddings mostly agreed with the human-labeled triplet orderings. Before any templates were used, the averaged description embedded with the MiniLM model disagreed with the human pool 33% of the time—which is much more palatable after remembering that 6% of these triplets cannot be satisfied due to transitive violations. The GNMDS embedding itself still produced a triplet disagreement of 17.78%, and its objective is to minimize triplet violations. Future work is needed to better contextualize how effective a 33% violation rate is.

Case studies including the gross overrating of novelty of design one in Figure 2 suggest that the MiniLM model is heavily influenced by unique language like the word ‘bicycle’. Human raters are able to separate unique components of a design from unique roles in the design. If a unique component is performing a commonplace role, humans are unlikely to label this component as unique. The limited ability for LLMs to contextualize atypical language is a noticeable limitation in their ability to replicate human similarity judgments. Future work is needed to identify models which might be better at contextualizing biasing vocabulary.

After using the FBS template architecture, none of the single-facet FBS descriptions appeared to be more effective than the template-free descriptions at satisfying triplets. Averaging all of the FBS descriptions for an individual author, however, produced a lower median violation and a much smaller variance. We believe the FBS-Av is showing improvement for two reasons. First, based off of their survey responses, humans are considering multiple modalities when performing triplet comparisons. If the goal is to have the LLM best replicate these similarity comparisons, it requires a description which captures any and all relevant modalities. The second reason involves mitigating the effects of idiosyncratic language when averaging across multiple descriptions. If an author has a unique way of describing a particular function, the effects will be less impactful after the structure and behavior embeddings are factored in.

The post-template responses for the novelty error showed different trends than the corresponding triplet violation plots. Although the two metrics are closely related, it is possible for an embedding with many violations to have a worse novelty error than an embedding with fewer violations. This is because triplets comparing dissimilar designs have a greater impact on embedding placement than triplets comparing similar designs. The limited word template in particular had among the worst boxplots for the triplet violation percentage but the best and least variable boxplot for the novelty error. We hypothesize that this is because the limited word template does not include enough information to inform hard-to-label triplets which compare similar items. Instead, this template heavily prioritizes more impactful triplets by categorizing designs only by their most defining features. In the future it would be interesting to see if descriptions could be crafted which put the most weight on the limited word template but also contain FBS modal descriptions with less weight on the LLM placement.

One major limitation of our work is the inability to quantify how much of the disagreement between LLM and GNMDS similarity assessments can be attributed to humans having access to visual information that the LLM does not. From the survey responses on the triplet labeling tasks, some raters cited ‘visual complexity’ other form-based criteria to justify their similarity comparisons. These visual identifiers are not easily described in text to enable LLM judgment on this modality. We are interested in future work where human raters would perform triplet comparisons on the same textual descriptions as the LLM. It is possible that without visual information humans would have a much closer agreement with the LLM models.

Only descriptions from seven authors were used for LLM embedding generation, and so the results of this paper lack strong statistical confidence. In the future, we would like to crowdsource additional pre- and post-template descriptions. It would also be beneficial to have some method of measuring how effectively an author followed a particular set of template guidelines. While for the most part our seven authors followed the template instructions, we noticed for the ‘Functional’ template in particular that authors had difficult time keeping their functions solution-neutral. Some authors included behavioral information for *how* functions were being accomplished in the design, which was against the template’s intentions. We believe that writing solution-neutral functions for the sketches was difficult for authors because all designs share the same core function: frothing milk. Existing research by Gill *et al.* suggests that humans most strongly value solution-neutral functions when comparing designs, but theses results were between designs with *different* core functions [3]. Nonetheless, the ‘Functional’ template might need to be modified for future experimentation to better facilitate solution-neutral descriptions.

In our analysis of the provided descriptions, we also noticed one particular inconsistency that we believe is strongly influencing the post-template LLM embedding results. When providing structural descriptions in particular, some responses would purely contain design elements and some responses would include the word ‘milk’ when authors described how those structures interacted. To a human reader, inconsistent use of the word ‘milk’ should not be a major factor in similarity judgments, as every design contains milk to be frothed. However, the LLM is unaware that all designs share this commonality, and so when one description contains the word ‘milk’ and another does not, the LLM sees this as major evidence of textual dissimilarity. Out of curiosity, we omitted the word ‘milk’ from all descriptions and replotted Figures 2-4. The results showed a substantial decrease in the variability of the ‘Structure’ template—where the word choice was the most inconsistent. These results were omitted from this manuscript, as the broader consequences of editing the human-provided responses are unknown, however, the findings do lead us to believe that inconsistent mention of implied or universal vocabulary may have large consequences in LLM-reported similarity. For future template guidelines, discouraging language that applies to every design might be practical.

| Problem / Model | gan_cnn_2d | | cgan_cnn_2d | | diffusion_ | |
|---|---|---|---|---|---|---|
| beams2d | green!25 3.40e+08 | green!25 3.20e+08 | 2.16e+08 | 2.10e+08 | 2.55e+08 | |
| | 1.82e-01 | green!25 1.99e-09 | green!25 4.84e-01 | 4.23e-27 | 1.88e-01 | |
| heatconduction2d | 2.33e-03 | 8.67e-05 | 3.14e-03 | 9.23e-05 | green!25 6.46e-03 | |
| | 7.50e-01 | 6.21e-04 | 8.12e-01 | 7.15e-04 | 1.88e-01 | |
| photonics2d | -1.62e+04 | -4.87e-01 | green!25 -1.09e+04 | green!25 -4.53e-01 | -1.95e+04 | |
| | green!25 6.45e-01 | 2.07e-37 | 3.16e-01 | 8.91e-49 | 5.06e-02 | gre |

## ACKNOWLEDGMENTS

We acknowledge the funding and support provided by the National Science Foundation via award #2155072. We would also like to thank Roger Beaty and Dan Johnson for helpful discussions on which LLM models would be appropriate to address in the paper.

## REFERENCES

[1] Hirtz, Julie, Stone, Robert B., McAdams, Daniel A., Szykman, Simon and Wood, Kristin L. “A functional basis for engineering design: Reconciling and evolving previous efforts.” *Research in Engineering Design* Vol. 13 No. 2 (2002): pp. 65–82. DOI 10.1007/s00163-001-0008-3.

[2] Fu, Katherine, Murphy, Jeremy, Yang, Maria, Otto, Kevin, Jensen, Daniel and Wood, Kristin. “Design-by-Analogy: Experimental Evaluation of a Functional Analogy Search Methodology for Concept Generation Improvement.” *Research in Engineering Design* Vol. 26. DOI 10.1007/s00163-014-0186-4.

[3] Gill, Amaninder Singh, Tsoka, Arnold N. and Sen, Chiradeep. “Dimensions of Product Similarity in Design by Analogy: An Exploratory Study.” Vol. Volume 7: 31st International Conference on Design Theory and Methodology. 2019. DOI 10.1115/DETC2019-98252.

[4] Bhatta, Sambasiva R. and Goel, Ashok K. “From design experiences to generic mechanisms: Model-based learning in analogical design.” *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* Vol. 10 No. 2 (1996): p. 131–136. DOI 10.1017/S0890060400001372.

[5] Siangliulue, Pao, Arnold, Kenneth C., Gajos, Krzysztof Z. and Dow, Steven P. “Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas.” *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*: pp. 937–945. 2015. ACM, Vancouver BC Canada. DOI 10.1145/2675133.2675239.

[6] Tversky, Amos. “Features of similarity.” *Psychological Review* Vol. 84 No. 4 (1977): pp. 327–352. DOI 10.1037/0033-295X.84.4.327. Place: US Publisher: American Psychological Association.

[7] Tamuz, Omer, Liu, Ce, Belongie, Serge, Shamir, Ohad and Kalai, Adam Tauman. “Adaptively Learning the Crowd Kernel.” (2011). URL https://arxiv.org/abs/1105.1033.

[8] Keeler, Matthew and Fuge, Mark. “Fewer Triplets Than You Think: Novelty Error Converges Faster Than Triplet Violations in Ordinal Embeddings.” *Volume 6: 35th International Conference on Design Theory and Methodology (DTM)*: p. V006T06A020. 2023. American Society of Mechanical Engineers, Boston, Massachusetts, USA. DOI 10.1115/DETC2023-116696.

[9] Pennington, Jeffrey, Socher, Richard and Manning, Christopher. “GloVe: Global Vectors for Word Representation.” Moschitti, Alessandro, Pang, Bo and Daelemans, Walter (eds.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*: pp. 1532–1543. 2014. Association for Computational Linguistics, Doha, Qatar. DOI 10.3115/v1/D14-1162.

[10] Reimers, Nils and Gurevych, Iryna. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” (2019). URL https://arxiv.org/abs/1908.10084.

[11] Wang, Wenhui, Wei, Furu, Dong, Li, Bao, Hangbo, Yang, Nan and Zhou, Ming. “MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers.” *Proceedings of the 34th International Conference on Neural Information Processing Systems*: pp. 5776–5788. 2020. Curran Associates Inc., Red Hook, NY, USA.

[12] Goldstone, Robert L., Medin, Douglas L. and Halberstadt, Jamin. “Similarity in context.” *Memory & Cognition* Vol. 25 No. 2 (1997): pp. 237–255. DOI 10.3758/BF03201115.

[13] Gero, John S. “Design Prototypes: A Knowledge Representation Schema for Design.” *AI Magazine* Vol. 11 No. 4 (1990): pp. 26–36. DOI 10.1609/aimag.v11i4.854.

[14] Kan, Jeff and Gero, John. “Using the FBS ontology to capture semantic design information in design protocol studies.” *Designing - Analysing Design Meetings* (2009): pp. 213–229.

[15] McAdams, Daniel A. and Wood, Kristin L. “A Quantitative Similarity Metric for Design-by-Analogy.” *Journal of Mechanical Design* Vol. 124 No. 2 (2002): pp. 173–182. DOI 10.1115/1.1475317.

[16] Chaudhari, Ashish, Bilionis, Ilias and Panchal, Jitesh. “Similarity in Engineering Design: A Knowledge-Based Approach.” *Proceedings of the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 7: 31st International Conference on Design Theory and Methodology*. DETC2019-98272. Anaheim, CA, August 18–21, 2019, 2019. DOI 10.1115/DETC2019-98272.

[17] van der Maaten, Laurens and Weinberger, Kilian. “Stochastic triplet embedding.”: pp. 1–6. 2012. DOI 10.1109/MLSP.2012.6349720.

[18] Nandy, Ananya and Goucher-Lambert, Kosa. “Do Human and Computational Evaluations of Similarity Align? An Empirical Study of Product Function.” *Journal of Mechanical Design* Vol. 144 No. 4 (2022): p. 041404. DOI 10.1115/1.4053858.

[19] Roads, Brett D. and Love, Bradley C. “Enriching ImageNet with Human Similarity Judgments and Psychological Embeddings.” (2020). URL https://arxiv.org/abs/2011.11015.

[20] Bhatia, Sudeep. “Inductive reasoning in minds and machines.” *Psychological Review* DOI 10.1037/rev0000446.

[21] McTeague, Chris Patrick, Duffy, Alexander, Hay, Laura, Vuletic, Tijana, Campbell, Gerard, Choo, Pei Ling and Grealy, Madeleine. “Insights into design concept similarity judgements.” *15th International Design Conference (Design 2018)*. 2018.

[22] Hebart, Martin N., Zheng, Charles Y., Pereira, Francisco and Baker, Chris I. “Revealing the multidimensional mental representations of natural objects underlying human similarity judgments.” *Nature human behaviour* Vol. 4 No. 11 (2020): pp. 1173–1185. DOI 10.1038/s41562-020-00951-3.

[23] “Hugging Face – The AI community building the future.” (2024). URL https://huggingface.co/.

[24] Ralph, Paul and Wand, YA. “Proposal for a Formal Definition of the Design Concept. Design Requirements Engineering: A Ten-Year Perspective.” (2009).

## APPENDIX A. FBS TEMPLATE INSTRUCTIONS

Figures/Temp_Compare_Viol.pdf

FIGURE 3: POST-TEMPLATE TRIPLET VIOLATIONS

Figures/Temp_Compare_Nov.pdf

**FIGURE 4: POST-TEMPLATE NOVELTY ERROR**

Figures/FBS_Tem_Desc.png

Figures/FBS_Tem_Ex.png

Figures/Ex_Temp_Sketch.png

Figures/Lim_Tem.png