

Exploring Human and Language Model Alignment in Perceived Design Similarity using Ordinal Embeddings

Matthew Keeler

Department of Mechanical and Process Engineering
ETH Zürich, Switzerland
Email: mkeeler@ethz.ch

Mark D. Fuge*

Department of Mechanical and Process Engineering
ETH Zürich, Switzerland
Email: mafuge@ethz.ch

Aoran Peng

School of Engineering Design and Innovation
Pennsylvania State University
University Park, PA

Scarlett Miller

School of Engineering Design and Innovation
Pennsylvania State University
University Park, PA

In recent years, Large Language Models (LLMs) and Vision Language Models (VLMs) have excelled at tasks requiring human-like reasoning, inspiring researchers in engineering design to use LMs as surrogate evaluators of design concepts. But do these models actually evaluate designs like humans? While recent work has shown LM evaluations sometimes fall within human variance on Likert-scale grading tasks, those tasks often obscure the reasoning and biases behind the scores. To address this limitation, we compare LM word embeddings (trained to capture semantic similarity) with human-rated similarity embeddings derived from triplet comparisons (“is A closer to B than C?”) on a dataset of design sketches and descriptions. We assess alignment via local tripletwise similarity and embedding distances, allowing for deeper insights than raw Likert-scale scores provide. We also explore whether describing the designs to LMs through text or images improves alignment with human judgments. Our findings suggest that text alone may not fully capture the nuances humans key into, yet text-based embeddings outperform their multi-modal counterparts on satisfying local triplets. Based on these insights, we offer recommendations for effectively integrating LMs into design evaluation tasks.

1 Introduction

The early stage of the engineering design process demands effective ways for evaluating similarities among design concepts. Maximizing diversity and encouraging novel solutions in concept generation have both been shown to lead to better performing final designs [1–3]. At their core, novelty and diversity are measures of similarity—novelty being how dissimilar a design is to all other designs and diversity being the coverage of designs in ‘similarity space’ [4,5]. Unfortunately, measuring similarity or its related metrics is challenging. In practice, design concepts are often given novelty or diversity scores by domain experts [6]. However, relying on domain experts to assign novelty or diversity scores is both time-consuming and costly, severely limiting how many concepts can be assessed.

In recent years, Large Language Models (LLMs) and Vision–Language Models (VLMs) have demonstrated widespread success at replicating human-like reasoning in benchmarks like taking math exams [7] and coding [8]. These successes have inspired researchers in engineering design to use LMs as surrogate design evaluators. Many studies in this field have reported strong alignment between model predictions and human judgments using scalar or Likert-

*Address all correspondence to this author.

scale evaluations of design quality and similarity [9, 10]. However, such uni-dimensional scores often mask the nuances of human perception and provide limited insight into why or how models agree with—or diverge from—human evaluators.

In contrast to Likert scores, triplet-based similarity embeddings offer a richer framework for capturing human perception, asking raters to answer “Is design A closer to B or to C?” rather than assigning an absolute score [11–13]. These ordinal judgments are cognitively simpler and yield more self-consistent responses [14, 15], while also encoding the relative features of similarity that are challenging to articulate in a single rating scale [16]. In deriving low-dimensional embeddings from triplet comparisons, we can spatially realize the multi-faceted nature of design similarity as perceived by humans.

Crucially, human triplet embeddings can be directly compared to pre-trained LM text embeddings—vectors learned to capture semantic similarity in text. This comparison enables analysis at two levels: local alignment, where the orderings of triplets match between human and model; and global alignment, where the distances between all pairs of designs are preserved across embedding spaces. Together, these measures offer a more powerful test of perceptual agreement than traditional Likert correlations—which remains the standard in benchmarking evaluation on engineering designs.

In non-design fields like computer vision, prior work has examined alignment differences between language models and human judgments using Likert scales and triplet embeddings—typically on arbitrary “random things” datasets [17, 18]. Engineering designs, however, offer a unique testbed: they are inherently multimodal and have decades of research devoted to understanding how to capture them semantically. Emphasizing these distinct modalities may yield novel insights for LM-based design evaluators. While earlier studies in language and vision have shown that strong LM–human agreement on Likert ratings does not necessarily translate to similar alignment in triplet-based comparisons, they have not systematically investigated how the mode of design presentation affects human-LM alignment. In particular, it remains unclear whether highlighting functional, behavioral, or structural aspects of a design can uncover the features human raters prioritize—and how those text-based modality descriptions compare to human alignment in VLMs, which can leverage visual cues that text-only encoders cannot access.

To address these gaps, this paper makes the following contributions:

1. Assembles a novel dataset of human-written descriptions for early-stage engineering designs, structured according to modality-isolating templates that emphasize functional, behavioral, and structural facets.
2. Introduces three complementary metrics—triplet error, novelty error, and Procrustes disparity—to quantify both local and global alignment between human triplet embeddings and model embeddings, revealing that no single metric suffices to capture overall agreement.

3. Benchmarks three static and context-sensitive text-embedding models (GloVe, MiniLM, and OpenAI text-embedding-3) on their ability to replicate human similarity perceptions across description templates.
4. Evaluates three state-of-the-art VLMs (GPT-4o, GPT-4.1, GPT-o1, GPT-o4-mini) on the same triplet tasks using design sketches, analyzing how visual reasoning impacts alignment with human judgments.
5. Discusses the implications of our initial study for integrating LLMs and VLMs into human-centered design workflows, highlighting both the potential and the limitations of AI-driven evaluation in complex, multi-dimensional design tasks.

The rest of the paper addresses these questions in §3-5. In this initial study, we provide empirical results from a variety of crowdsourced triplet-labeling and description-writing surveys. LM-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 and between text-embedding LMs and VLMs.

2 Related Work

Related work has been subdivided into the following sections: (1) work that tries to understand or metrically represent how humans perceive similarity (2) work that uses AI to align with human similarity judgments.

2.1 Modeling Human-Perceived Similarity

Similarity is a complex and multi-dimensional quality that does not lend itself to easy quantification. In the field of psychology, human similarity judgments are popularly thought to be a linear combination of shared and distinct features [19]. Similarity is commonly evaluated by either counting shared features [19] or assigning Likert values to pre-defined scoring criteria [20]. Similarity is then assessed with vector distances [20, 21], or with distributional divergence measures like Kullback-Liebler [22]

However, prior work has shown that the weighted importance humans associate with features is highly context-sensitive. Humans tend to focus on holistic qualities when judging items globally, but emphasize distinct or local features when comparing items side-by-side [23, 24]. This makes ratings like Likert scores especially prone to self-inconsistent responses. As raters gain broader exposure to a dataset over repeated questioning, their internal reference frame evolves, and the features they value when rating also evolve.

To mitigate these scoring inconsistencies, recent work has proposed collecting similarity judgments in the form of triplet comparisons (Is Design A closer to Design B or Design C?) [11, 12, 25]. Choosing the closest of two options or the “odd-man-out” has proven to result in greater self-consistency than Likert scores in human ratings of colored shapes [14], diverse faces [15], and large diverse image datasets [16]. Triplet queries avoid the need for global

1 scoring criteria and the embeddings that satisfy triplets can
2 express similarity-leveraged features that may be difficult
3 to articulate or generalize in a scoring rubric [11, 12, 25].
4 For these reasons, this study uses triplet embeddings as our
5 model for human-perceived similarity.

6 In engineering design, modeling perceived similarity
7 presents additional challenges in the form of multi-modality.
8 Designs are often described using a mixture of visual infor-
9 mation (e.g., geometry, form) and textual/functional knowl-
10 edge (e.g., what a system does or how it behaves). Prior
11 studies have shown that modality strongly influences how
12 humans compare items: people emphasize commonalities in
13 text-based comparisons, but emphasize differences in visual
14 comparisons [26]. Better understanding the hierarchy with
15 which humans prioritize multimodal information is critical
16 to understanding how surrogate evaluators should be used or
17 constructed.

18 To capture design complexity, several frameworks
19 have emerged to isolate modalities by decomposing de-
20 signs solution-neutral sub-functions [27, 28] or functional-
21 behavioral-structural (FBS) representations [29–31]. Our
22 study uses these decompositions to emphasize a particular
23 modality when designs are given to an LM—in an effort to
24 capture human-like abstraction.

25 2.2 Using AI to Replicate Human Similarity Judgments

26 Many recent efforts have sought to harness AI to mirror
27 the complexities of human perception. Outside of design dis-
28 ciplines—notably in clinical psychology—researchers have
29 begun aligning LLMs with brain activity measured via fMRI.
30 Affolter *et al.* proposed decoding fMRI signals to steer
31 LLMs toward brain-like word choices [32], while others
32 have proposed the inverse of predicting fMRI activity dur-
33 ing word-association tasks using LLMs [33, 34]. Kwon
34 *et al.* even propose that this LLM–fMRI synergy could un-
35 lock methods for translating engineers’ mental representa-
36 tions of design concepts during creative ideation [35]. Yet
37 Fegghi *et al.* caution that much of the apparent alignment
38 arises from trivial text features—sentence length, token posi-
39 tion, and the like—rather than genuine parallels in language
40 processing [36].

41 Parallel critiques have emerged from vision and NLP re-
42 search, where reliance on single-value scores often obscures
43 finer alignment discrepancies. Liu *et al.* show that pair-
44 wise or head-to-head grouping better reflects human judg-
45 ments of text quality than outright scores [37]. In visual
46 domains—from simple kernel experiments [14] to random-
47 image datasets [16]—triplet-style comparisons reveal axes
48 of human–model agreement that scalar metrics miss. Shen
49 *et al.* demonstrate the same in evaluating LMs in text-
50 summarization: head-to-head evaluations expose axes of
51 misalignment that would otherwise be obscured in a raw met-
52 ric [38]. Triplet-like analysis has also been used to help vi-
53 sualize how multi-modal information improves human-LM
54 alignment in text and image datasets—studying affects on
55 aligned dimensions and entropy scores that Likert scores
56 would not reflect [17, 39]. However, none of these studies ex-

plore datasets as complicated and multi-modal as engineer-
ing designs. They also study alignment measures like cosine
similarity—which is quite practical for text comparison but
less practical than global diversity or novelty disagreement
that engineers value in a good surrogate.

In engineering design, score-based quality or novelty
measures remain the norm, even as practitioners explore sim-
ilar questions regarding multi-modal impacts [40] and as-
sess whether surrogate evaluators can match human perfor-
mance [9, 10]. But focusing solely on raw scores—“good
enough” though they may seem—makes it nearly impossi-
ble to diagnose why a surrogate evaluator diverges from hu-
man judgment. Kwon *et al.* use human labeled triplets as
a baseline for comparing AI-similarity judgements, but only
on a dataset of simple chair designs and comparing against
a convolutional neural network [41, 42]. Because the CNN
can only take in the geometry of the 3D chair model, their
work only explores alignment in purely visual similarity. Our
work builds on Kwon *et al.* by applying triplet-based com-
parison methods—which spatially uncover alignment devi-
ations between humans and AI more effectively than Likert
scores—to state-of-the-art language models now used as sur-
rogate evaluators. We test this on a dataset of complex engi-
neering designs that combine textual and visual information,
and we link the resulting alignment metrics to novelty and
diversity which are metrics of value for surrogate evaluation
in the domain of concept evaluation.

3 Methods

To compare perceived similarity between LMs and hu-
mans, a baseline embedding must be established that rep-
resents human similarity groupings among designs. This
baseline embedding is constructed by minimizing the viola-
tions of human-provided triplet orderings. With this embed-
ding, disagreement in similarity assessment can be identified
by comparing the Euclidean distances between the human-
baseline and LLM-generated embedding spaces. Then mod-
ifications can be made to the descriptions provided to the
LM and the benefit in embedding/metric alignment can be
observed.

3.1 Milk-Frother Dataset

A dataset of ten hand-drawn designs from an undergrad-
uate Pennsylvania State University design course was used
in this work [25]. 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 de-
vice as well as a brief descriptive title. All ten sketches are
provided in Figure 1. This dataset has been used in many
works in the field of design evaluation [6, 9, 10, 25].

3.2 Collecting Human-Ordered Triplets

To capture human-perceived similarity, human raters
were tasked with ordering designs in the form of triplet
queries. Existing work has shown that humans can con-
sistently find the closer match between two candidate pairs

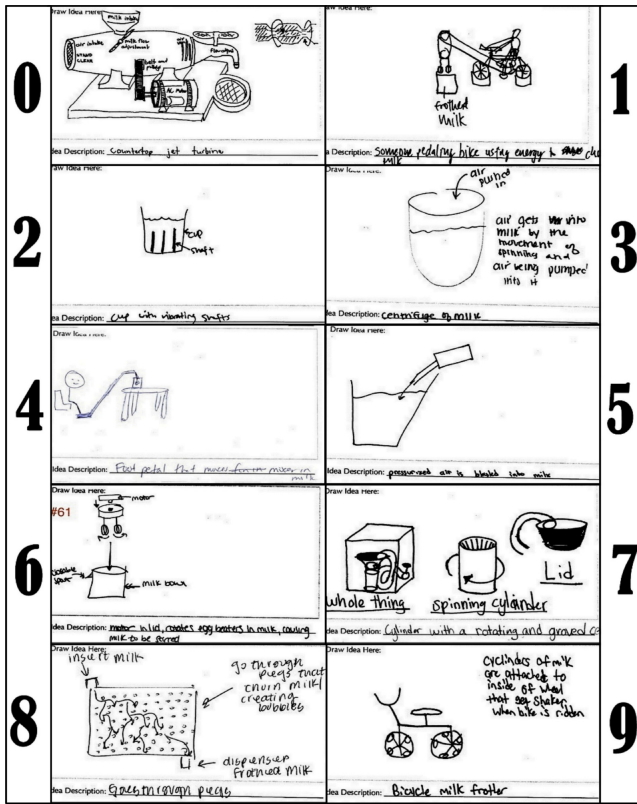


Fig. 1. Dataset of student-generated milk frother sketches. Numbers 0-9 label a corresponding milk frother design.

of designs [11, 25]. This is contrasted with the more difficult task of articulating or quantifying the closeness of pairwise comparisons—which requires development of domain-specific scoring criteria [43].

Using the Milk-Frother dataset, a survey conducted by Ahmed *et al.* [25] 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 University’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 Human Triplets

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, specifically using the `cblearn` Python library [44]. GNMDS deconstructs triplet orderings into two pairwise distances—one for each reference-candidate pair [13]. 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 [11, 12]. 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 similarity comparison is unknown [16, 30]. 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 [45], 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 Text Embedding Models

After creating the human-perceived similarity space, the next step was to select pretrained LMs to generate embeddings based on text descriptions of the Milk-Frother designs.

For this paper three text-embedding LMs were chosen, two open-sourced from the Huggingface sentence-transformers library: GloVe (Global Vectors) [46] and MiniLM-L12-v2 (MiniLM) [47], and OpenAI’s text-embedding-3 [48]. GloVe 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 [46]. GloVe was chosen as a baseline due to its widespread use and documentation. MiniLM-L12-

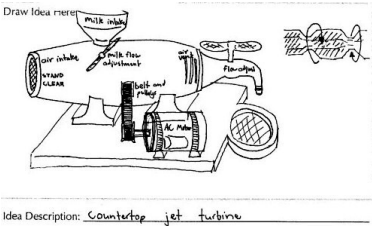
Image	Template 1			Template 2
	Function	Behavior	Structure	Limited Word
	Funnels an intake of poured milk. Aerates milk inside of the container. Mechanically mixes the air-milk solution. Dispenses the frothed milk product.	Intake milk flow is gravity-fed and corralled before flow stream is redirected. 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. An example of a single author's crowdsourced descriptions for each writing template

v2 is a newer Siamese Network built on a BERT-like pre-trained model [49, 50]. Whereas GloVe only models co-occurrence of whole words, MiniLM incorporates the context around words and can parse unrecognized words [49]. MiniLM-L12-v2 was chosen to represent a more state-of-the-art model. Finally to represent a state-of-the-art closed source model, OpenAI's text-embedding-3 was chosen [48]. This is OpenAI's best performing model on the standard Massive Text Embedding Benchmark (MTEB) for LLM-encoders [51].

For this work, only text-embedding language models were considered. These models directly encode textual similarity without stochasticity or the need for prompt-engineering like reasoning models. Well-tuned reasoning models may perform better on human-alignment tasks, but the stochasticity of these models makes it difficult to observe the effect of description writing changes.

3.5 Describing Milk-Frothers in Text

To use text-embedding models, we needed some way of taking the text and image data in the Milk Frother sketches and creating a purely text-based description. But how should the text be described for the LM? Does prioritizing certain distinguishing facets of a design lead to better human-LM alignment? To explore this question we had seven graduate students from Pennsylvania State University's engineering design labs write text description paragraphs for all ten Milk-Frother sketches. These seven students were given various writing templates that asked them to focus on different design facets.

3.5.1 Free-Form Template

To serve as baseline, the first writing prompt was instructionless (referred to as Free-Form or FF). Authors were given no additional information beyond explaining the design in sentence form. These text descriptions were given to the GloVe, MiniLM, and text-embedding-3 models to create

three LLM-generated embeddings for every author.

3.5.2 Function-Behavior-Structure Template

The next template was motivated by the work in Gero [29] 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 [31, 52–55].

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 subassemblies. 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 ?? . The exact instructions for this template can be found in the supplemental material document. This template will be referred to as 'FBS'.

3.5.3 Limited Word Template

For the final 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 ???. The exact instructions for this template can be found in the supplemental material document. This template will be referred to as ‘Lim’.

3.6 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.

3.7 Triplet Violations

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 embedding 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.

3.8 Pairwise Distances

The second metric compares the pairwise distances between the embeddings. Pairwise distances provide the best metric for understanding how similar two objects are perceived to be. 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 perceived similarity.

3.9 Novelty Difference

Novelty differences describe the deviation in 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 [25, 56]. Novel items are identified as designs that are far from the centroid of the embedding space [25, 57]. 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 LM embedding 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 [45]. 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.

3.9.1 Procrustes Disparity

The final metric measures the overall geometric alignment between embedding spaces using the Procrustes Disparity. To calculate the Procrustes Disparity, the 10-dimensional GNMDs embedding from the human consensus triplets is taken as baseline. Then the triplets from the high-dimensional LM embeddings are embedded using a new 10-dimensional GNMDs model. Procrustes alignment is done to scale and rotate the 10-dimensional LM-triplet embedding

1 to best match the 10-dimensional human-triplet embedding.
2 The disparity measure the Euclidean distance between the
3 aligned embeddings. Whereas the pairwise and triplet viola-
4 tions are effective at identifying localized disagreements, the
5 Procrustes Disparity is effective at capturing overall embed-
6 ding consistency.

7 3.10 Vision and Multimodality

8 While text-based LM embeddings are very systematic
9 at capturing similarities in written descriptions, these models
10 lack reasoning and their worldview is limited to their train-
11 ing corpus. To understand the limitations of these models,
12 we could have a reasoning LLM model perform triplet rat-
13 ings like the humans did. However, to account for variability,
14 this would require giving reasoning LLMs both a good writ-
15 ten description of the designs—which is unclear what that
16 would be—and good prompting information. A VLM, on
17 the other hand, could be given the images and instructions for
18 triplet rating exactly like the humans received. This would let
19 us observe the impact of a reasoning model without needing
20 to address the variability in written descriptions of designs.
21 Additionally, VLMs might be able to capture spatial infor-
22 mation that text embeddings cannot easily encode.

23 For this paper four state-of-the-art VLMs were tested, all
24 from OpenAI: GPT-4o, GPT-4.1, GPT-o1, and GPT-o4-mini.
25 These models were prompted to perform triplet comparisons
26 on the Milk Frother images three times for each model. The
27 exact prompting instructions can be found in the supporting
28 material document. After three trials, each models triplet rat-
29 ings were described by a majority vote. The triplets were
30 then given to GNMDs to make 10-dimensional embeddings
31 for comparison against the human embedding and the em-
32 beddings from the text-embedding models.

33 As a final test, GPT-4o was given the same writing tem-
34 plates as the humans—with some additional guidance to
35 ensure valid responses. These written descriptions to test
36 if VLM-written responses could pass as human written re-
37 sponses. This also allowed us roughly gauge how effectively
38 the writing templates were being followed by human authors,
39 as the VLM follows the template instructions very closely.

40 4 Results

41 4.1 Pre-Template Text-Model Embedding Heatmaps

42 Figure 2 displays the similarity diagnostics for the three
43 embeddings before human authors were given instructions
44 on crafting their design descriptions. Descriptions provided
45 without a template will be referred to as free-form (FF). The
46 four monochrome heatmaps represent the Euclidean pair-
47 wise distances found in the respective embedding. The red
48 heatmap denoted ‘Human10’ corresponds to the baseline 10-
49 dimensional GNMDs embedding given human triplets. The
50 color bar on the right of the heatmap shows the relative per-
51 centile rank of each pairwise distance as described in §3.6.
52 Light colors represent a small pairwise distance and thus a
53 high-degree of perceived similarity. For example, the human
54 raters believe designs one and four are very similar—which

from Figure 1 are both human-operated stirring devices.

Below the ‘Human10’ heatmap are the triplet violation
and novelty order metrics for the embedding. As mentioned
in §3.6, the ‘Human10’ baseline cannot satisfy all human-
provided triplets and instead violates 7.78% of the human la-
bels. About 4% of these unsatisfied triplets can be attributed
to transitive violations in the labeled triplet pool which are
impossible to satisfy. Below the reported triplet disagree-
ment is a sorted list based on increasing centroidal distance
or novelty. According to the ‘Human10’ embedding, human
raters perceive design eight as the most novel and design one
as the least novel.

The heatmaps in the middle column to the right of the
‘Human10’ heatmap are for embeddings generated by the
MiniLM, GloVe, and OpenAI text-embedding-3 models re-
spectively. These embeddings are found from averaging the
FF description embeddings for all seven authors. On the av-
erage FF description, the MiniLM model violates 32.5% of
the human-rated triplets, the GloVe model violates 38.89%,
and the OpenAI model violates 36.39%. Each LLM model
has about four to five times the violations of the human-
labeled GNMDs model.

Triplet disagreement between the models further mani-
fests in differences in the novelty ordering. Both the MiniLM
and GloVe embeddings agree that design nine is the most
novel, and the OpenAI model labels design nine as the sec-
ond most novel after design four. Based solely on the rank
order of novel items, none of the LLM models accurately re-
flect human-perceived novelty. This is particularly evident in
the placement of design eight—which humans believe is the
most novel design and language models believe ranks in the
bottom half of novelty scores.

On the far right are heatmaps which represent the differ-
ence between the ‘Human10’ and LLM heatmaps. Dark col-
ors 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 be-
lieves designs five and seven 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 similar-
ity. For the MiniLM model, design one consistently presents
blue squares—indicating that the MiniLM model tends to la-
bel 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 that may
reveal some of the rationale behind the LLMs embedding
placements. While the humans group design one as being
very similar to other stirring devices, the LLM models be-
lieve the design to be unique. We hypothesize that this is be-
cause the LLM is putting heavy emphasis on the word ‘bicy-

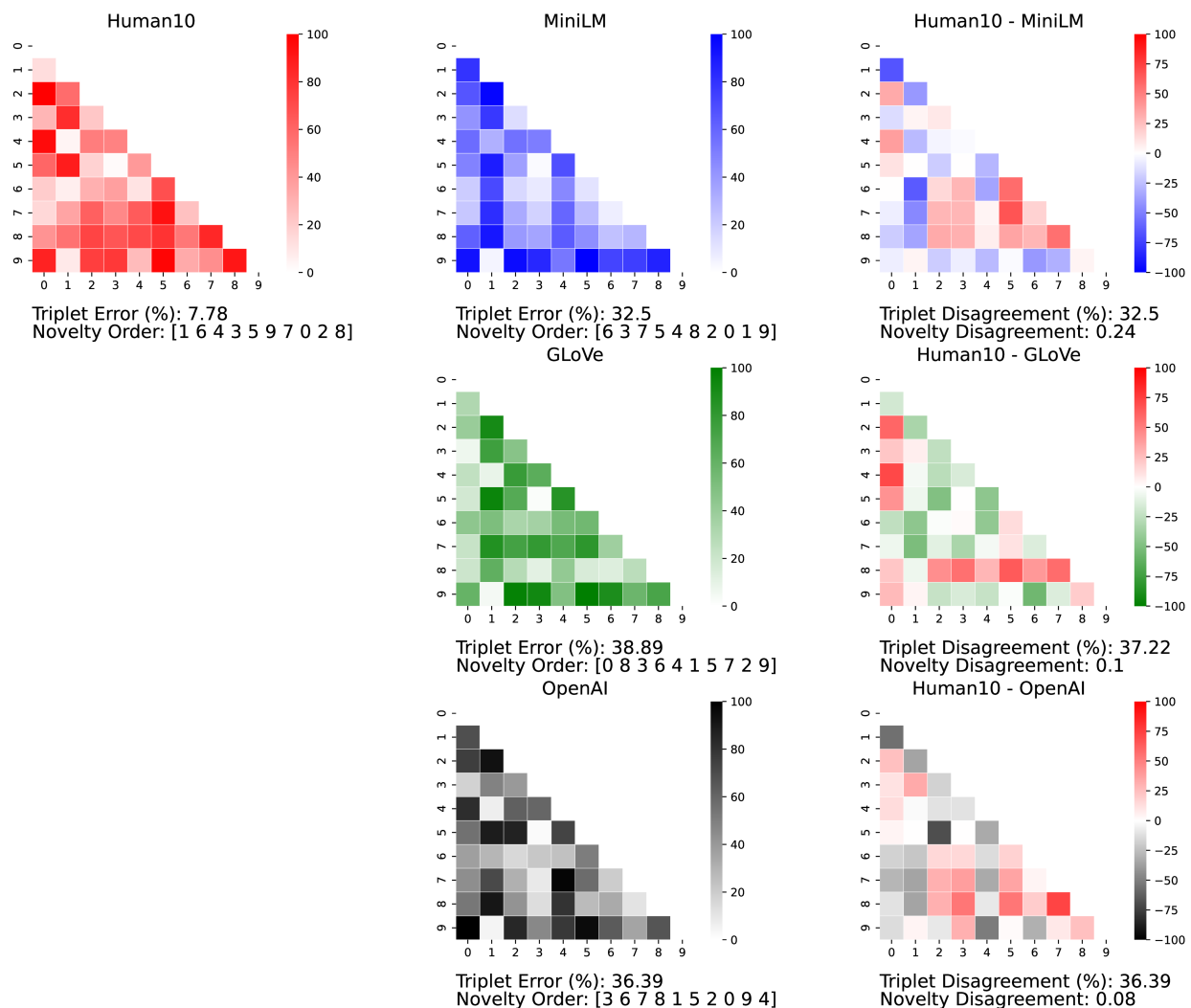


Fig. 2. Embedding pairwise distance heatmaps for pre-template written descriptions.

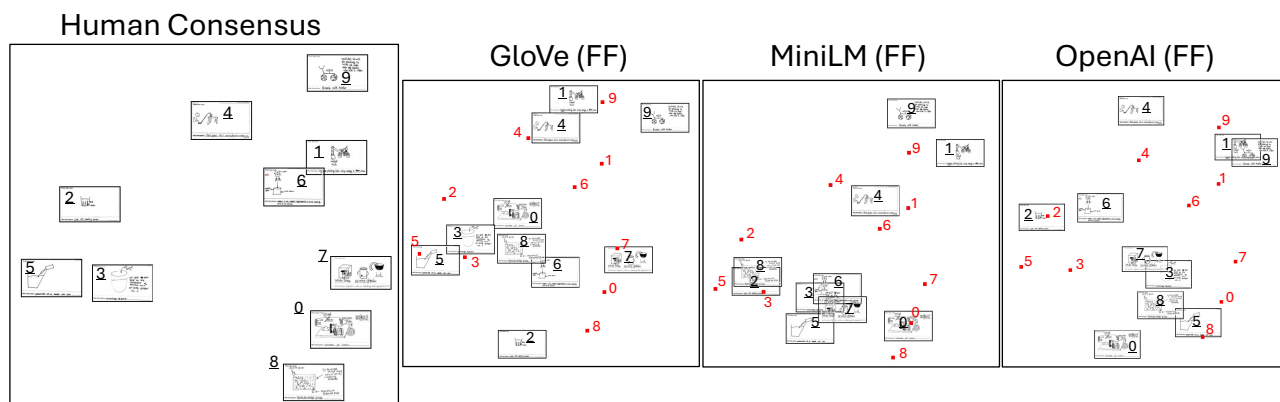


Fig. 3. Two-dimensional GNMDs reconstructions of human-triplet and LLM-description embeddings. Red dots represent the human-labeled placements for comparison.

cle’, which is an unusual word among the designs. Similarly, the language models consistently rate design nine—the other bicycle design—as more novel than the humans do. 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 transformer-LLM 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 transformer-based models, 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.

4.2 Pre-Template Text-Model 2D Visualizations

Figure 3 shows two-dimensional embeddings of the triplets which exist in higher-dimensional space in the human and LLM embeddings. Triplets are collected in the high-dimensional space and given to GNMDs to embed in the two-dimensional space. Because triplet embeddings are rotational and scale-invariant, Procrustes transformations are performed on all LLM-produced embeddings against the 2D human embedding. This is done to allow for easy visual comparisons to be drawn between embeddings. The red scatter points in the LLM embeddings represent placements of the human-labeled designs.

Figure 3 visualizes some general trends that we observed in the heatmaps. Ultimately, the LLMs have some noticeable clustering differences from the human embedding. The man-operated devices are universally grouped in the top right across all embeddings, but from there the models have significant disagreement on what clusters exist. The humans group three noticeable clusters—the man-operated mixers (1, 4, 9) plus the motorized mixer (6), the three cup designs (2, 3, 5), and the multi-component designs (0, 7, 8). The final two clusters appear to be largely governed by visual similarities. The language models—which only have textual descriptions of designs—do not identify the “cup” and “multi-component” clusters. Designs 0, 7, and 8 in particular are grouped very differently than in the human “multi-component” cluster. We believe this is because the descriptions do not capture the visual complexity of the design and humans are leveraging this feature in their similarity comparison ratings.

4.3 Post-Template Text-Model Results

Figure 4 shows the similarity metric results across all authors and all writing templates for the MiniLM model respectively. Complementary plots for the GloVe and OpenAi text-embedding models can be found in the supplemental material document. 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 §3.5.2; ‘Lim’ refers to the limited-word template; ‘3Fac’ refers to averaging the functional, behavioral, and structural descriptions; while ‘Stack’ refers to stacking the descriptions into one large composite description. Finally the ‘Raters’ label refers to the human raters who completed the triplet rating tasks which was then used to form the consensus human rating. This data is included to contextualize the variance of LM-alignment with the consensus against inter-human variance.

The top plot of Figure 4 shows box distributions for the seven author descriptions. Boxplot whiskers are found by adding 1.5 times the interquartile range. Red stars represent averaging the description embeddings across all authors. The bottom scatterplots show the full distribution of the LLM embedding results for a given author and writing template.

It is immediately clear that the text-only LM embeddings remain distinct from human embeddings regardless of the template used. For the MiniLM model across all three error metrics, the ‘Raters’ boxplot lies at a firmly lower alignment error than any LLM-produced embedding. Similar results are found for the other text-only LMs. We expect then, that a reasonable discriminator could easily identify embeddings produced by humans from embeddings produced by text-only LMs. It does not appear that writing descriptions for an LM can be a reasonable surrogate for human-triplet evaluation.

In Figure 4 red stars represent averaging all author embeddings for a given template. For the triplet violation error, averaging the embeddings consistently improves LLM-human alignment well beyond the median author error. However, averaging embeddings does not consistently improve the novelty and Procrustes error metrics from the median response. There are few reasons why this might be the case. Firstly, averaging embeddings can warp the scale of the embedding space as variances are smoothed. This is not problematic for triplet violations which are defined on local relationships, but global shape metrics like novelty and Procrustes errors are effected. It is also possible that averaging LLM embeddings are diluting the influence of nuanced vocabulary in favor of capturing general trends in similarity. This is useful for the triplet violations—as they are largely governed by proper clustering. However, for the LMs to understand the magnitude of distances required in novelty and Procrustes measures, emphasis on nuanced vocabulary may be needed. This is likely why the ‘Lim’ template—which prioritizes nuanced vocabulary—consistently shows improvement in averaging.

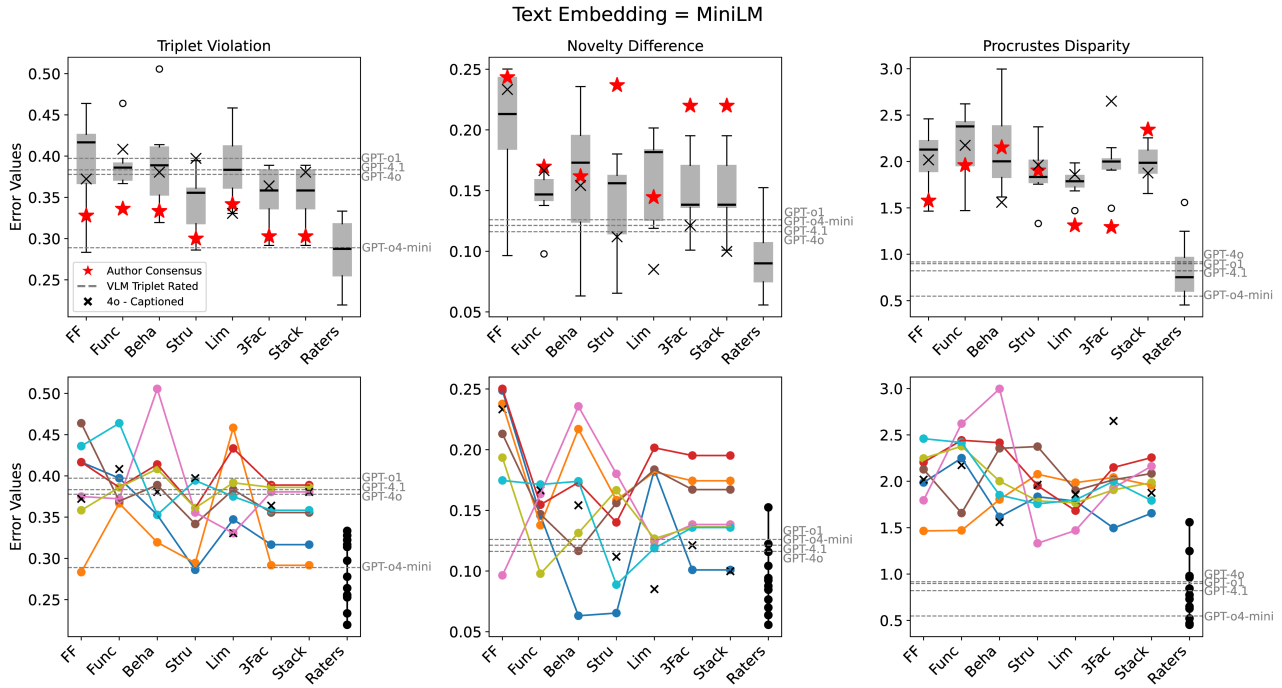


Fig. 4. Alignment error metrics for the MiniLM text embedding model across all writing templates: FreeForm, Functional, Behavioral, Structural, Limited Word, and combinations of them. Boxplots and full distributions are shown for all authored descriptions. Stars represent the embedding errors of the averaged author embedding. 'X' marks show the performance of descriptions written by the GPT-4o VLM. Dashed lines show the performance of VLMs given the triplet rating task. 'Raters' represents the individual human rater triplet responses to show human inter-variance.

4.4 VLM Captioning

The 'X' marks in Figure 4 are the results from descriptions written by GPT-4o. For the most part, the error values appear consistent with the median of author descriptions, if not better. It is hard to draw major conclusions from this test, but because the errors roughly follow the same trends as the human authors across the templates, we believe that human authors did a reasonable job of following the template instructions. Future study on VLM-captioning for text embedding is needed to see if this could be a viable surrogate for humans.

4.5 VLM Triplet Rating

The dashed lines from Figure 4 are the results from having various OpenAI VLMs perform the triplet rating tasks like the humans did. Across the board, the VLM models produce significantly more human-like Procrustes alignment—with all four models being well within human-rater variance. The VLMs also do better on average at modeling human-like novelty errors than the text-embedding models. However, the GPT-4o, GPT-4.1, and GPT-o1 produce significantly worse triplet violations than the averaged text embedding results. The results show a model can have relatively well-performing global embedding alignment while having relatively poor performing local triplet consistency. Multiple metrics are needed to fully understand human/LM alignment.

The GPT-o4-mini model, however, does achieve triplet violations well within human-rater variance. It is the only model—VLM or text-embedding—that achieves human-like performance on all tested metrics. GPT-o4-mini is OpenAI's newest reasoning model at the time of drafting this paper.

4.6 Structural Template 2D Visualizations

In Figure 5, the two-dimensional reconstructed embeddings are shown for language models on the average of the structural 'Stru' template. This is done in the same manor as the freeform descriptions in Figure 3. Now, given structural descriptions, the language models can capture both the 'man-operated' and 'cup design' clusters that the humans observed—where before only the man-operated cluster was preserved with freeform descriptions. Language models given the structural template still fail to capture the 'complex design' cluster (designs 0, 7, 8). We believe this case study supports the hypothesis that humans are first drawn to structural similarity between designs. This is especially true for simple designs which are comprised of few and distinguishing structures. However, when designs become more complicated (ie. 0, 7, 8) and cannot be described by a distinguishable structure, humans must prioritize additional information in their comparisons. These additional features of similarity are also needed to better define the cross-cluster comparisons and the distance between clusters.

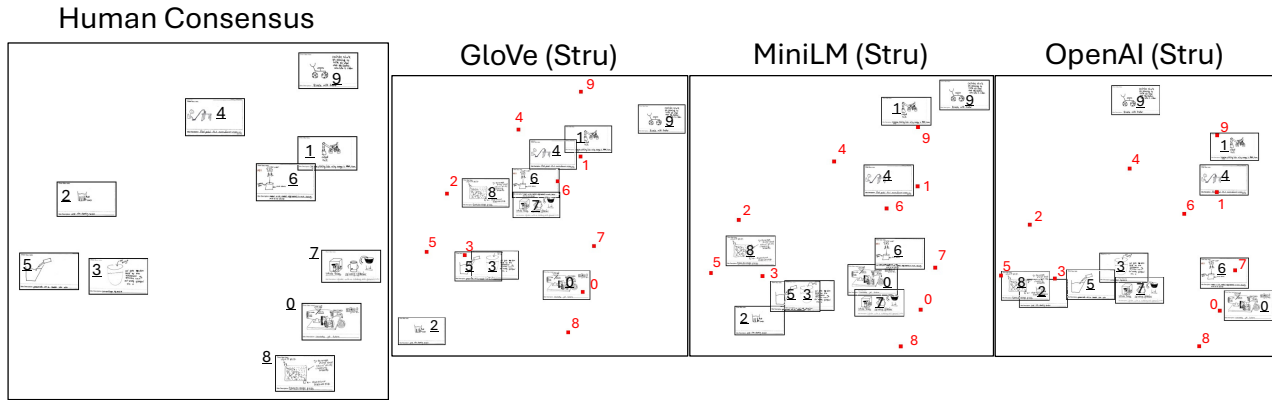


Fig. 5. Two-dimensional GNMDS reconstructions of human-triplet and LM-description embeddings from the structural template. Red dots represent the human-labeled placements for comparison.

5 Discussion

5.1 Individual Similarity Metrics Hide the Full Story

A common theme across the results of this paper lies in how heavily human-LM alignment measures can disagree. A specific model/template may result in the best triplet violation error, but lead to poor novelty error and Procrustes disparity. The distinction is the most clear between localized embedding alignment—like triplet violations, and global embedding—like novelty and Procrustes alignment. However, even among the more global measures there are templates/models with close to human-like novelty scores but poor Procrustes alignment. This is further evidence that human-LM alignment cannot be evaluated on single metrics, and similarity metrics unlock a larger understanding of alignment. Works which claim LM-human alignment on a single scoring rubric should be used cautiously.

5.2 Unique Vocabulary on Language Model Embeddings

The template-free heatmaps in Figure 2 demonstrate strong disagreement on the relative novelty orderings on designs with unique vocabulary—namely the designs one and nine which contain bicycles. Language models consistently rate these designs as more unique than human raters do. This case study provides evidence for the hypothesis that LMs are poor at contextualizing unique vocabulary around its greater role in the design description. While design one does contain a bicycle—a very unique structure among milk frothers—the bicycle is only a power source for a commonplace mixer architecture, and so the humans rate it as the least novel design. Design eight, on the other hand, does not contain unique components but includes a unique arrangement of components. Humans key into the unique arrangement and rate it as the most novel design, while LMs see commonplace vocabulary and rate it less novel.

If LMs are to be used to replicate human-perceived similarity of engineering designs, they require some way of better understanding the role of unique vocabulary in the design. Future work might involve the creation of graph structures that explain relations between components in the design. For

example, design one might be represented by a node that says “bicycle” connected by an energy flow relation to a node that says “egg-beater mixer”. These inputs might be useful for allowing language models to look beyond the component list to the greater composition as a whole.

5.3 The Missing Ingredient: Vision or Reasoning?

The post-template boxplots in Figure 4 show that text-encoder language models cannot achieve human-like embedding alignment, no matter what template or combination of templates is used—at least among the templating approaches we explored. While there may be other ways of describing designs in text that this paper did not attempt, a wide berth of description modalities was covered, and yet the results show that no authored description fell within human-variance across multiple metrics. This is evidence that text alone may not be sufficient for fully capturing the complexities of human similarity perception.

The most obvious missing component for the text models is the lack of vision. We can see from the freeform 2D visualizations that text-only models do reasonably well at recreating the human-perceived cluster of “man-operated stirring devices”. When giving structure-focused descriptions, the text-only models also do reasonably well at capturing the cluster of “cup-like” devices. However, the placement of the remaining—more visually complicated—designs and the spacing across clusters remains poor. These embedding differences manifest in large global alignment errors such as Procrustes disparity. The vision models, on the other hand, all perform well within human-variance on Procrustes disparity. Access to visual features appears necessary for good overall placement of groups in the similarity embedding space.

However, despite their human-like Procrustes disparity, the VLM models perform noticeably worse on triplet violation errors than the text models—except for the o4-mini, OpenAI’s newest reasoning model. It does not appear that o4-mini picks up more visual artifacts than the other VLMs based on similar embedding alignment measures. Rather, it seems that well-performing models need not only to en-

code relevant features but also to weigh those artifacts with human-like reasoning.

This may seem like an obvious finding, but it is interesting to see how much more effective the consensus 3Fac and Stack text embeddings are at modeling triplet violations compared to the VLMs—despite the VLMs having much better global embedding alignment. When an aggregate of author description embeddings is taken, the text models obtain a better weighing of human-biased language. This, in turn, gives the model a more human-like sense of how features are prioritized in human perception. We believe this makes the text embedding models better at fine-grained triplet responses—cases where all designs are similar. Correctly ordering fine-grained triplets requires capturing human feature importance hierarchies effectively. However, incorrectly flipping the order of a fine-grained triplet only leads to changes within in-cluster embedding coordinates. The text models may be better at fine-grained triplets, but the cross-cluster triplets they fail at—due to the lack of vision—ultimately make the VLMs better at global embedding alignment.

5.4 Insights and Future Work

Of course, the GPT-o4-mini model is a noticeable exception to trends seen in the other VLM models. GPT-o4-mini achieved in-human-variance alignment across all three tested metrics, greatly outperforming the other tested GPT models. Unfortunately, because this model is new and closed source, it is not immediately clear why the GPT-o4-mini outperforms the other models so drastically in triplet alignment. We would expect reasoning based models to be better at the non-reasoning GPT-4o and GPT-4.1, but o4-mini greatly outperforms its predecessor reasoning model in GPT-o1. Future work may involve doing a deep dive into why GPT-o4-mini excels in this similarity benchmark. It might be helpful to test o4-mini on images with the text deleted like was done in Picard *et al.* [10] to see if the improvement can be attributed to better image processing, or with only text descriptions to see if it is better reasoning. Finding what makes o4-mini perform well might help guide future use of LM-evaluators. It is promising that the newest LM evaluators achieve more human-like similarity scores.

While we do not know exactly what makes o4-mini effective, our study shows that replicating human-like similarity judgments requires a complex hierarchy of considerations. Human leveraged features change with the context of presented designs. Some model/template pairs do better at capturing and weighing a certain subset of features, while other pairs do better at another subset. This study shows that text-based models with human-provided descriptions might be better at encoding nuances needed to satisfy tight in-cluster triplets, while vision is needed to capture all features of similarity needed for global alignment. Future work should involve combining models into an ensemble for better performance. We believe this could lead to models which excel at both localized triplet alignment and global embedding alignment. GPT-o4-mini may already be a step

in this direction.

This paper presents an initial investigation into how multimodality affects alignment between human judgments and language-model embeddings. Using ten milk-frother designs as a case study, we isolated text-only, image-only, and hybrid representations, but the dataset’s small size limits our ability to generalize across engineering domains. Crucially, vision-language embeddings reproduce global similarity structures within human inter-rater variance, while text-based or reasoning-enhanced embeddings excel at fine-grained triplet judgments. These robust patterns motivate scaling up to larger, multi-domain concept sets—and exploring ensemble methods that leverage visual inputs to tighten overall embedding alignment, alongside textual reasoning to sharpen local agreement.

6 Conclusion

This study explored using ordinal triplet embeddings to model human similarity perception, and compared the alignment of that embedding to embeddings produced by text-only and vision language models. The results emphasize the need to look at multiple metrics in evaluating human-LM alignment, as localized triplet adherence and global embedding alignment can tell very different stories regarding LM performance. Based on large gaps between inter-rater alignment error and LM-alignment error, we conclude that text-only models are failing to capture some human-leveraged features of similarity, and modifying the way the design is presented in text cannot remove this gap. The gap is likely the result of lack of vision, as VLMs perform much better on global alignment measures. Text-only models still have a place though and achieve better localized triplet alignment on average. The choice of text-only evaluator or VLM depends on if localized triplet or global embedding consistency is desirable. We posit that achieving truly human-like similarity judgments across all metrics will require an ensemble approach that dynamically integrates textual and visual representations to reflect the context-dependent feature hierarchies inherent in human perception.

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] Dow, Steven P., Glassco, Alana, Kass, Jonathan, Schwarz, Melissa, Schwartz, Daniel L. and Klemmer, Scott R. “Parallel prototyping leads to better design results, more divergence, and increased self-efficacy.” *ACM Trans. Comput.-Hum. Interact.* Vol. 17 No. 4. DOI 10.1145/1879831.1879836.

- [2] Ward, Allen, Liker, Jeffrey K, Cristiano, John J and Sobek, Durward K. "The second Toyota paradox: How delaying decisions can make better cars faster." *Long Range Planning* Vol. 28 No. 4 (1995): p. 129. DOI 10.1016/0024-6301(95)94310-U.
- [3] Jansson, David G. and Smith, Steven M. "Design fixation." *Design Studies* Vol. 12 No. 1 (1991): pp. 3–11. DOI [https://doi.org/10.1016/0142-694X\(91\)90003-F](https://doi.org/10.1016/0142-694X(91)90003-F).
- [4] Shah, Jami J., Smith, Steve M. and Vargas-Hernandez, Noe. "Metrics for measuring ideation effectiveness." *Design Studies* Vol. 24 No. 2 (2003): pp. 111–134. DOI [https://doi.org/10.1016/S0142-694X\(02\)00034-0](https://doi.org/10.1016/S0142-694X(02)00034-0).
- [5] Johnson, Tyler A, Cheeley, Avery, Caldwell, Benjamin W and Green, Matthew G. "Comparison and extension of novelty metrics for problem-solving tasks." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 50190: p. V007T06A012. 2016. American Society of Mechanical Engineers.
- [6] Miller, Scarlett R., Hunter, Samuel T., Starkey, Elizabeth, Ramachandran, Sharath, Ahmed, Faez and Fuge, Mark. "How Should We Measure Creativity in Engineering Design? A Comparison Between Social Science and Engineering Approaches." *Journal of Mechanical Design* Vol. 143 No. 3 (2021): p. 031404. DOI 10.1115/1.4049061.
- [7] Lewkowycz, Aitor, Andreassen, Anders, Dohan, David, Dyer, Ethan, Michalewski, Henryk, Ramasesh, Vinay, Slone, Ambrose, Anil, Cem, Schlag, Imanol, Gutman-Solo, Theo, Wu, Yuhuai, Neyshabur, Behnam, Gur-Ari, Guy and Misra, Vedant. "Solving Quantitative Reasoning Problems with Language Models." (2022). URL 2206.14858.
- [8] Li, Yujia, Choi, David, Chung, Junyoung, Kushman, Nate, Schrittwieser, Julian, Leblond, Rémi, Eccles, Tom, Keeling, James, Gimeno, Felix, Dal Lago, Agustin, Hubert, Thomas, Choy, Peter, de Masson d'Autume, Cyprien, Babuschkin, Igor, Chen, Xinyun, Huang, Po-Sen, Welbl, Johannes, Goyal, Sven, Cherepanov, Alexey, Molloy, James, Mankowitz, Daniel J., Sutherland Robson, Esme, Kohli, Pushmeet, de Freitas, Nando, Kavukcuoglu, Koray and Vinyals, Oriol. "Competition-level code generation with AlphaCode." *Science* Vol. 378 No. 6624 (2022): p. 1092–1097. DOI 10.1126/science.abq1158.
- [9] Edwards, Kristen M., Tehranchi, Farnaz, Miller, Scarlett R. and Ahmed, Faez. "AI Judges in Design: Statistical Perspectives on Achieving Human Expert Equivalence With Vision-Language Models." (2025). DOI 10.48550/arXiv.2504.00938. URL <http://arxiv.org/abs/2504.00938>. ArXiv:2504.00938 [cs].
- [10] Picard, Cyril, Edwards, Kristen M., Doris, Anna C., Man, Brandon, Giannone, Giorgio, Alam, Md Ferdous and Ahmed, Faez. "From Concept to Manufacturing: Evaluating Vision-Language Models for Engineering Design." (2024).
- [11] 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>.
- [12] van der Maaten, Laurens and Weinberger, Kilian. "Stochastic triplet embedding.": pp. 1–6. 2012. DOI 10.1109/MLSP.2012.6349720.
- [13] Agarwal, Sameer, Wills, Josh, Cayton, Lawrence, Lanckriet, Gert, Kriegman, David and Belongie, Serge. "Generalized Non-metric Multidimensional Scaling." *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*: pp. 11–18. 2007. PMLR. ISSN: 1938-7228.
- [14] Demiralp, Cagatay Demiralp, Bernstein, Michael S. and Heer, Jeffrey. "Learning Perceptual Kernels for Visualization Design." *IEEE Transactions on Visualization and Computer Graphics* Vol. 20 No. 12 (2014): pp. 1933–1942. DOI 10.1109/TVCG.2014.2346978.
- [15] Li, Linjie, Song, Amanda, Malave, Vicente, Cottrell, Garrison and Yu, Angela. "Extracting Human Face Similarity Judgments: Pairs or Triplets?" *Journal of Vision* Vol. 16 No. 12 (2016): p. 719. DOI 10.1167/16.12.719.
- [16] 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.
- [17] Born, Frieda, Muttenthaler, Lukas, Greff, Klaus, Unterthiner, Thomas, Lampinen, Andrew K, Müller, Klaus-Robert and Mozer, Michael C. "Evaluating and supervising vision models with multi-level similarity judgments." *CCN 2024: the 7th annual conference on Cognitive Computational Neuroscience*. 2024.
- [18] Muttenthaler, Lukas, Greff, Klaus, Born, Frieda, Spitzer, Bernhard, Kornblith, Simon, Mozer, Michael C., Müller, Klaus-Robert, Unterthiner, Thomas and Lampinen, Andrew K. "Aligning Machine and Human Visual Representations across Abstraction Levels." (2024). DOI 10.48550/arXiv.2409.06509. URL <http://arxiv.org/abs/2409.06509>. ArXiv:2409.06509 [cs].
- [19] 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.
- [20] 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.
- [21] Nandy, Ananya, Dong, Andy and Goucher-Lambert, Kosa. "Evaluating Quantitative Measures for Assessing Functional Similarity in Engineering Design." *Journal of Mechanical Design* Vol. 144 No. 3 (2022): p. 031401. DOI 10.1115/1.4052302.
- [22] Chaudhari, Ashish, Bilionis, Ilias and Panchal, Jitesh. "Similarity in Engineering Design: A Knowledge-Based Approach." *Proceedings of the ASME 2019 In-*

ternational 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.

[23] Glazer, Rashi and Nakamoto, Kent. “Cognitive geometry: An analysis of structure underlying representations of similarity.” *Marketing Science* Vol. 10 No. 3 (1991): pp. 205–228.

[24] Krumhansl, Carol L. “Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density.” *Psychological Review* Vol. 85 No. 5 (1978): pp. 445–463.

[25] Ahmed, Faez, Ramachandran, Sharath Kumar, Fuge, Mark, Hunter, Samuel and Miller, Scarlett. “Interpreting Idea Maps: Pairwise Comparisons Reveal What Makes Ideas Novel.” *Journal of Mechanical Design* Vol. 141 No. 2 (2019): p. 021102. DOI 10.1115/1.4041856.

[26] Gati, Itamar and Tversky, Amos. “Weighting common and distinctive features in perceptual and conceptual judgments.” *Cognitive Psychology* Vol. 16 No. 3 (1984): pp. 341–370.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Gero, John S. and Kannengiesser, Udo. “A function–behavior–structure ontology of processes.” *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* Vol. 21 No. 4 (2007): pp. 379–391. DOI 10.1017/S0890060407000340.

[32] Affolter, Nicolas, Egressy, Beni, Pascual, Damian and Wattenhofer, Roger. “Brain2Word: Decoding Brain Activity for Language Generation.” (2020). URL <https://arxiv.org/abs/2009.04765>.

[33] Tang, Jerry, Du, Meng, Vo, Vy A., Lal, Vasudev and Huth, Alexander G. “Brain encoding models based on multimodal transformers can transfer across language and vision.” (2023). URL

<https://arxiv.org/abs/2305.12248>.

[34] Benara, Vinamra, Singh, Chandan, Morris, John X., Antonello, Richard, Stoica, Ion, Huth, Alexander G. and Gao, Jianfeng. “Crafting Interpretable Embeddings by Asking LLMs Questions.” (2024). URL <https://arxiv.org/abs/2405.16714>.

[35] Kwon, Elisa, Patterson, John D, Beaty, Roger E and Goucher-Lambert, Kosa. “Assessing the Alignment Between Word Representations in the Brain and Large Language Models.” *International Conference on Design Computing and Cognition* (2024): pp. 207–223.

[36] Feghhi, Ebrahim, Hadidi, Nima, Song, Bryan, Blank, Idan A. and Kao, Jonathan C. “What Are Large Language Models Mapping to in the Brain? A Case Against Over-Reliance on Brain Scores.” (2024). URL <https://arxiv.org/abs/2406.01538>.

[37] Liu, Yinhong, Zhou, Han, Guo, Zhijiang, Shareghi, Ehsan, Vulić, Ivan, Korhonen, Anna and Collier, Nigel. “Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators.” (2025). DOI 10.48550/arXiv.2403.16950. URL <http://arxiv.org/abs/2403.16950>. ArXiv:2403.16950 [cs].

[38] Shen, Chenhui, Cheng, Liying, Nguyen, Xuan-Phi, You, Yang and Bing, Lidong. “Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization.” *Findings of the Association for Computational Linguistics: EMNLP 2023*: pp. 4215–4233. 2023. Association for Computational Linguistics, Singapore. DOI 10.18653/v1/2023.findings-emnlp.278.

[39] Merx, Danny, Frank, Stefan and Ernestus, Mirjam. “Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge.” Chersoni, Emmanuele, Hollenstein, Nora, Jacobs, Cassandra, Oseki, Yohei, Prévot, Laurent and Santus, Enrico (eds.). *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*: pp. 1–11. 2022. Association for Computational Linguistics, Dublin, Ireland. DOI 10.18653/v1/2022.cmcl-1.1.

[40] Song, Binyang, Miller, Scarlett and Ahmed, Faez. “Attention-Enhanced Multimodal Learning for Conceptual Design Evaluations.” *Journal of Mechanical Design* Vol. 145 No. 4 (2023): p. 041410. DOI 10.1115/1.4056669.

[41] Kwon, Elisa and Goucher-Lambert, Kosa. “Comparing and Evaluating Human and Computationally Derived Representations of Non-Semantic Design Information.” *Journal of Mechanical Design* Vol. 146 No. 3 (2024): p. 031401. DOI 10.1115/1.4063567.

[42] Kwon, Elisa and Goucher-Lambert, Kosa. “EXAMINING THE BOUNDARY BETWEEN NEAR AND FAR DESIGN STIMULI.” *Proceedings of the Design Society* Vol. 3 (2023): pp. 1725–1734. DOI 10.1017/pds.2023.173.

[43] 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.

[44] Künstle, David-Elias and Luxburg, Ulrike von. “cblearn: Comparison-based Machine Learning in Python.” *Journal of Open Source Software* Vol. 9 No. 98 (2024): p. 6139. DOI 10.21105/joss.06139.

[45] 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.

[46] 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.

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

[48] OpenAI. *text-embedding-3-small Model Documentation* (2024). Available at <https://platform.openai.com/>.

[49] 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.

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

[51] Muennighoff, Niklas, Tazi, Nouamane, Magne, Loïc and Reimers, Nils. “MTEB: Massive Text Embedding Benchmark.” (2023). URL <https://arxiv.org/abs/2210.07316>.

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

[53] Gero, John S. and Kannengiesser, Udo. “The situated function–behaviour–structure framework.” *Design Studies* Vol. 25 No. 4 (2004): pp. 373–391. DOI 10.1016/j.destud.2003.10.010.

[54] Chen, Liuqing, Zuo, Haoyu, Cai, Zebin, Yin, Yuan, Zhang, Yuan, Sun, Lingyun, Childs, Peter and Wang, Boheng. “Toward Controllable Generative Design: A Conceptual Design Generation Approach Leveraging the Function–Behavior–Structure Ontology and Large Language Models.” *Journal of Mechanical Design* Vol. 146 No. 12 (2024): p. 121401. DOI 10.1115/1.4065562.

[55] Kan, Jeff and Gero, John. “Using the FBS ontology to capture semantic design information in design pro-

tol studies.” *Designing - Analysing Design Meetings* (2009): pp. 213–229.

[56] 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.

[57] Hwang, Dongwook and Wood, Kristin Lee. “Assessing the novelty of design outcomes: Using a perceptual kernel in a crowd-sourced setting.” *Design Computing and Cognition’20*. Springer (2022): pp. 411–422.