

## UTILIZING PRE-TRAINED LANGUAGE MODELS TO SUPPORT CIRCULAR DESIGN DECISION-MAKING

Ananya Nandy, Ashley Hartwell, KC Morris

National Institute of Standards and Technology, Gaithersburg, MD

### ABSTRACT

*Circular practices (e.g., extending product/component life and recovering end-of-life products) can divert products from waste streams, creating alternative material sources and resilient production systems. Recently developed design guidelines and principles can ensure that circularity is planned into a product design. However, when expertise on circularity is lacking, determining which of the numerous guidelines to prioritize and implement is challenging, especially for complex product assemblies. This study evaluates an approach to support the selection of relevant circular design principles for different product types, leveraging advances in natural language processing and semantic understanding. Several pre-trained language models are evaluated for their performance in ranking the circular design principles most relevant to two sets of consumer electronic products in comparison to human-determined ratings. The approach can reduce barriers to practically implementing circular product design principles, but results indicate that alignment with human decision-making is sensitive to model choices and product representation.*

Keywords: product design, circular economy, circular design, semantic search, large language models (LLMs)

### 1. INTRODUCTION

Circular product design (CPD) is an approach to the product development process that emphasizes planning for a product's end-of-life treatment at its inception while maximizing product value during each part of its life cycle [1]. It ensures that goods and their constituent materials remain in the economy and out of unwanted sinks such as landfills, retaining value and reducing the deleterious impacts of poor end-of-life management [2]. While many companies have established circular economy initiatives to meet corporate sustainability goals [3], more guidance is needed to determine the best strategies to apply during product design given a product's unique architecture, purpose, and features.

CPD principles, which serve as distilled sets of actions to increase product and material circularity [4], [5], have gained traction with designers and engineers as a way to enable

circularity through design. Past research has developed sets of general, sector, and product-specific CPD principles, but decision support methods to aid the practical operationalization of circular design are lacking [6], [7], [8].

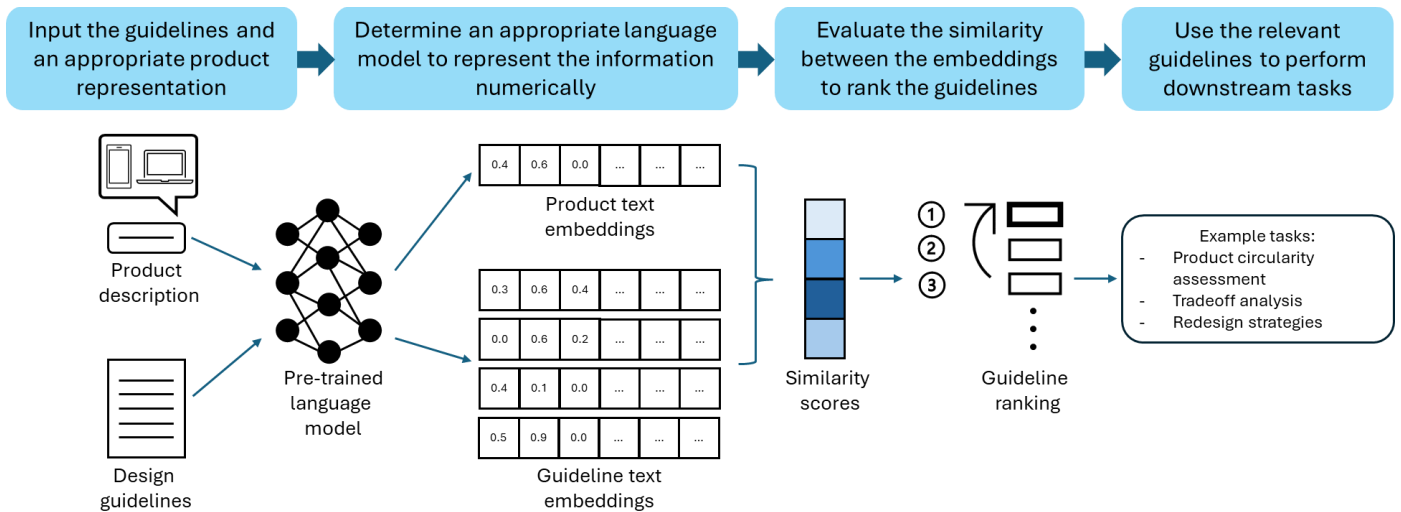
Previous work has leveraged natural language processing (NLP) to generate circularity themes for the purpose of educating designers and improving understanding of CPD principles [9]. This work builds upon efforts to facilitate the implementation of CPD principles within design practice, exploring the use of NLP to help designers choose from broad sets of circularity strategies in the absence of a full product definition (i.e., during early-stage design). Specifically, pre-trained language models (via semantic search or text generation) are used to rapidly surface the most relevant CPD principles for a diverse set of electronic products, using only commonly available information, like the product type or its list of components. Such an automated method can guide circular decision-making by highlighting the principles to be prioritized and targeted for increasing product circularity.

### 2. BACKGROUND

The development of design guidelines for circularity and the application of language models or semantic search to early-stage design are both relevant to the task addressed in this study. Prior efforts in these areas are outlined below.

#### 2.1 Circular Product Design Guidelines and Principles

Recent circular product design principles can be traced to established strategies from practices such as Eco-design, Design for Sustainability, and Cradle-to-Cradle design [7], [10]. In parallel, circularity practitioners and researchers are working with standards organizations such as ASTM International [5] and the European Committee for Standardization (CEN) [4] to develop and disseminate a common set of consensus-based circular product design guidelines for general-purpose use by a broad range of stakeholders. Additionally, emerging industry-specific circular design guidelines and research address the nuances of various sectors, spanning from electronics to textiles,



**FIGURE 1.** General envisioned approach to CPD guideline selection by automatically generating a ranking of the relevant guidelines to use for downstream tasks, such as identifying design feature changes or calculating circularity metrics

and their associated materials, production methods, and supply chains [11], [12], [13], [14], [15], [16].

Despite the proliferation of resources on design for circularity, operationalizing CPD guidelines and principles remains challenging. Additional scaffolding that accompanies these guidelines is needed to help designers prioritize actions that are most relevant to a specific product, make design changes, assess to what degree the guidelines have been applied, and determine the impact of changes on product circularity overall. This paper focuses on the prioritization and selection of design principles because all downstream actions are dependent on this step, yet it can be a significant challenge, particularly for those with limited expertise related to circularity.

## 2.2 Semantic Search in Design Methodology

Though design involves data of various modalities, including images and 3D representations, a large amount of design information is found in textual forms (e.g., descriptions, requirements, etc.). A desire to leverage this unstructured data for design has driven significant interest in NLP and information retrieval to support design tasks, such as design knowledge reuse [17] or design-by-analogy [18].

The development and release of pre-trained language models have further enhanced the ability to process textual data, through both text embedding and text generation. Transformer-based models, building from the BERT (Bidirectional Encoder Representations from Transformers) architecture, have improved the embedding of words, sentences, and documents (i.e., mathematical representations of textual data). These improvements enable greater semantic understanding, allowing search and information retrieval based on the meaning and context associated with queries rather than only keyword matches [19], [20]. Beyond text embeddings, prompt-based text generation models such as GPT-4 or the Llama model family enable chat-like interfaces that can be used for a variety of language-related tasks [21] [22].

Both types of models can increasingly be applied to design tasks. For example, work from NASA used semantic search (powered by text embedding models) to retrieve relevant lessons learned from prior projects, helping engineers identify risks for new design projects and facilitating knowledge transfer [23]. Research has also utilized text generation and semantic search for product life cycle assessment (LCA), augmenting limited data and automatically matching unstructured information about manufacturing processes to existing life cycle database processes [24]. Leveraging textual information and language models to support the application of design guidelines, therefore, is an opportunity to be explored for circular product design.

## 3. MATERIALS AND METHODS

The goal of the study was to facilitate CPD principle selection by systematically surfacing the principles most relevant for various product types. The task was formulated as follows: given different product types, rank the appropriate design guidelines in order of relevance to that product type. The general procedure is summarized in Figure 1 and implemented using Python, the *sentence-transformers* library [20], and an open-source large language model (LLM).

The success of the task was determined by relevance, which is assumed to vary based on product type. In this case, the product type refers to a categorization of similar products (e.g., laptops) and relevance is based on human-determined ratings. The task procedure was tested using a set of circular design guidelines and data from two sets of consumer electronic products from prior literature. Electronic products were chosen because the electronics industry is a prime sector for circular economy interventions due to its dependence on scarce or non-renewable materials, global supply chains, projected production growth, and high product value and durability. Furthermore, the electronics industry has taken an interest in product circularity and has initiated several efforts to apply circular design guidelines to their products [11], [12].

### 3.1 Data Preparation

The input data consisted of the product type and a representative product components list (including component variations for a single product type) as text. The results were evaluated against human-determined ratings, which rated the relevance of various circular design principles to different product types. Data from prior literature was augmented to obtain sets of product types, component lists, and relevance ratings, as outlined below.

**Data Source 1 (Products and Guidelines).** Bovea and Pérez-Belis introduced a set of circular design guidelines which are utilized in this work as the test set of guidelines to select from for product circularity improvement [25]. The set of circular design guidelines consists of 33 guidelines organized into 6 guideline groups (Extension of life span (ELS), Disassembly: Connectors (DC), Disassembly: Product structure (DPS), Product reuse (PR), Components reuse (CR), and Material recycling (MR)). While alternate guidelines may be used, this set was selected due to availability of relevance ratings.

In the prior work, the guideline groups were rated according to their relevance from 1 (low) to 3 (high) for improving the circularity of 10 product types (vacuum cleaners, hand blenders, heaters, kettles, iron, hair dryers, toasters, coffee makers, juicers and sandwich makers). The ratings were determined by the authors of the original study based on characteristics of each product type, such as its tasks, life span, durability, and performance [25]. Of the 10 product types, 3 (hair dryer, coffee maker, sandwich maker) were removed because each guideline group was rated equally relevant for those product types, leading to an uninformative ranking. The remaining product types and corresponding relevance ratings were used to evaluate the relevance of guidelines retrieved through semantic search.

**Data Augmentation for Source 1.** Product component information was unavailable in the dataset from Bovea and Pérez-Belis, although several products within each product type were disassembled to establish the relevance ratings in the initial study [25]. To include information about product components in the absence of a specific bill of materials (BOM), a list of general components for each product type was developed using the images provided in the original study and online research by the authors (shown in Table 1). Component alternatives (e.g., button or switch) were included when they were perceived to be common variants within the product type. The listed components were not exhaustive but representative of key components that may be found in the product type broadly.

**Data Source 2 (Products).** Babbitt et al. introduced a BOM dataset for consumer electronic products based on disassembly [26]. This dataset consisted of 95 unique products categorized into 25 product types, including the components and materials within each specific product. 5 of these product types and their specific products (4 drones, 3 gaming consoles, 14 laptops, 3 thermostats and 1 smartphone, which were differentiated within a product type by having a unique component list) were used as representative product types that spanned a variety of uses and

**TABLE 1.** Augmentation to data from Bovea and Pérez-Belis [25] to add product components

Product Type	Components
<b>hand blender</b>	Electric motor, cutting blades, immersion shaft housing, handle housing, power cord or battery, button or switch
<b>heater</b>	Heating element, power cord, thermostat, temperature control dial, exterior housing, safety shutoff
<b>iron</b>	Soleplate, thermostat, temperature control dial, insulated electrical cord, water reservoir, power cord, handle, body housing, switch
<b>juicer</b>	Liquid reservoir, power cord, cone attachment, electric motor, exterior housing, switch
<b>kettle</b>	Heating element, lid, water container, power cord, base housing, switch
<b>toaster</b>	Heating element, button or knob, timer or thermal sensor, lever, exterior housing, metal slot, power cord
<b>vacuum cleaner</b>	Motor, filter bag or canister, drive belt, brush roll or beater bar, handle, wheels, fan, hose, power cord reel, exterior housing, button or switch

structures. This subset was used as another source to evaluate the relevance of guidelines retrieved through semantic search.

**Data Augmentation for Source 2.** In the dataset from Babbitt et al., the products were disassembled and information about the product components and materials was available [26]. However, ratings establishing each guideline’s perceived relevance to each product type were unavailable. To maintain consistency with the first dataset and augment the dataset with relevance ratings, the procedure from Bovea and Pérez-Belis was applied to the second dataset of electronic products, as shown in Table 2. The authors rated the relevance of each guideline group (compared to each other) to each product type, considering the provided guiding questions [25]. Then, an internal collaborator with knowledge of the electronics sector reviewed the ratings and finally, modifications were made for consensus.

**TABLE 2.** Augmentation to data from Babbitt et al. [26] to add relevance ratings for each guideline group (ELS - Extension of life span, DC - Disassembly: Connectors, DPS: Disassembly: Product structure, PR: Product reuse, CR: Components reuse, MR: Material recycling)

	drone	gaming console	laptop	smart phone	thermostat
<b>ELS</b>	3	2	2	3	1
<b>DC</b>	2	2	3	3	2
<b>DPS</b>	3	1	2	2	1
<b>PR</b>	3	2	1	2	1
<b>CR</b>	2	1	3	2	1
<b>MR</b>	2	3	2	2	3

**Query Representation:** The queries consisted of the product type (not the specific brand) or the product type with its components list, taken from the datasets. The product type, captured in one or two words, provides the least possible amount of information about a product while still enabling category differentiation. Adding components increases the amount of specificity. When components were included, the query followed the template “{product type} made of {components}” (e.g., “hand blender made of electric motor, cutting blades, immersion shaft housing, handle housing, power cord or battery, button or switch”). The components list for specific products within the same product type could differ when BOM data was available.

**Retrieved Guidelines:** The language models were used to rank the set of design guidelines. The guidelines [25] were checked for spelling and grammar before data processing since subtle changes in wording may impact language models. One guideline was updated (“Use joints **than** can be disassembled rather **then** fixed joints” changed to “Use joints **that** can be disassembled rather **than** fixed joints”). The guidelines (e.g., “Promote monomaterial designs”) and their overarching group (e.g., “Material recycling”) were both included in the set to provide more granularity in retrieval, resulting in 39 total guidelines. Although relevance ratings were only available for the guideline groups, the rating assigned to a guideline group was assumed to apply to all guidelines within the group.

Note that the semantic content of the query (product information) vs. the retrieved text (guidelines) differ greatly in this task. Therefore, the current query is a starting point for product representation but may need refinement in the future.

### 3.2.3 Models

A subset of publicly available pre-trained models was tested, accounting for variability in training data or training objectives and subsequent performance on general or domain-specific tasks. The considered models were primarily limited to embedding or re-ranking models compatible with the *sentence-transformers* library. Additionally, one large text generation model was included for comparison (hosted through an on-site cluster). While a wide variety of pre-trained models are available for various tasks, none are trained for the specific task addressed here. Semantic search most closely describes the task since the query representation (i.e., the product) and the representation of the retrieved results (i.e., the design guideline or principle) are not the same in length or content type. Additionally, the models were utilized as-is with no domain-specific fine-tuning. The pre-trained models were not exhaustively tested but were chosen to reflect the diversity of available model types. Therefore, the models considered here do not necessarily indicate recommendations but evaluate the baseline potential for using language models to map product types to CPD guidelines without collecting significant amounts of data or training new models. Table 3 shows a summary of each model, its characteristics, and rationale for why it was chosen.

Two of the models required the additional input of a prompt

to achieve the task. The *SFR-Embedding-Mistral* model required one instruction sentence describing the task:

**Prompt:** “Given a product category, retrieve the circular design guidelines most relevant and applicable to the product category”

The *Llama-4-Maverick-17B-128E-Instruct-FP8* model required a detailed prompt to generate the outputs in a usable format:

**System Prompt:** “You are a product designer considering how to design products for a circular economy.”

**Prompt:** “Each line of the following list is an individual circular design guideline:

{list of guidelines}

Instructions: Score each individual circular design guideline on the list (from 0.00 to 1.00) in order of its relevance and applicability, compared to the other guidelines, to the following product category: {{product type}}/{product type} made of {components}. Return the score and an explanation of the score. Return the scores in the following format: provided guideline, score, explanation. Ensure every single one of the guidelines has a score (no skipping a guideline) and ensure that none of the guidelines have the same score (no tied scores). Provide no other formatting or content other than those indicated.”

These prompts were iteratively checked for erroneous outputs, though not exhaustively tested. Subtle changes to the prompt wording were not accounted for in the study but may impact the scores and rankings returned by prompted models, particularly for *Llama-4* which demonstrates stochasticity for even the same prompt. Only a single run of text generation was used to obtain the scores from *Llama-4* (although the prompt was re-run if results were not returned in the specified format).

### 3.3 Evaluation Measure

The measure used to evaluate the results against the human-determined relevance ratings is the Normalized Discounted Cumulative Gain (NDCG), which is commonly used for evaluating the ranking quality of information retrieval systems [33]. NDCG is bounded from 0 to 1, where 1 indicates perfect alignment of the retrieved ranking with known relevance ratings. NDCG relies on the assumption that items rated as highly relevant should be retrieved at higher ranks (i.e., earlier on the list) compared to moderately relevant items, which should in turn be retrieved at higher ranks than non-relevant items. The discounted cumulative gain (DCG) penalizes when highly relevant items appear at lower ranks and is calculated as follows up to the desired rank position  $k$ , where  $rel_i$  is the relevance rating at rank position  $i$ :

$$DCG_k = \sum_{i=1}^k \frac{2^{rel_i-1}}{\log_2(i+1)}. \quad (1)$$

DCG can be normalized by dividing by the ideal discounted cumulative gain (IDCG):

**TABLE 3.** Summary of models tested for the task of ranking relevant circular design guidelines

Model	Task trained for	Training data	Rationale for selection
<b>Text Embedding</b>			
<b>all-mpnet-base-v2 [27]</b>	STS (sentence textual similarity): short texts up to ~384 word pieces	1,170,060,424 sentence pairs	High performance for sentence similarity according to sentence-transformers library (on 4/14/2025) [28]
<b>multi-qa-mpnet-base-dot-v1 [27]</b>	Asymmetric search: questions and answers	214,988,242 question and answer pairs	Utilized in prior work and high performance in semantic retrieval according to sentence-transformers library (on 4/14/2025) [28]
<b>SFR-Embedding-Mistral [29]</b>	Multi-task: STS (sentence textual similarity), clustering, retrieval, reranking	Details not available	High performance on the MTEB v2 leaderboard considering the relevant tasks only (STS, Retrieval, Reranking) (on 4/14/2025) [30], [31]
<b>ms-marco-MiniLM-L12-v2 [20]</b>	Asymmetric search: passage retrieval and ranking/reranking	~500,000 query and passage pairs	Cross-encoder model type where pairs are embedded and scored directly, rather than producing embeddings separately and calculating similarity
<b>allenai-specter [32]</b>	Symmetric search: scientific paper (title & abstract) similarity	~ 684,000 training triples (query, positive, negative)	Trained on scientific data specifically Note: The Semantic Scholar dataset used for this model is also included in other models
<b>Text Generation</b>			
<b>Llama-4-Maverick-17B-128E-Instruct-FP8 [22]</b>	Chat/visual reasoning and natural language generation	~22 trillion tokens of public, licensed, and company data, up to August 2024	State-of-the-art large language model trained on a large quantity of data that can be used for a variety of tasks

$$NDCG_k = \frac{DCG_k}{IDCG_k}. \quad (2)$$

IDCG is calculated by ordering items with known relevance ratings in order of decreasing relevance and calculating DCG, until rank position  $k$ , for that order. By nature, NDCG<sub>k</sub> tends to increase as  $k$  increases. An empirical study of ranking systems reveals that individual queries can be sensitive to the cutoff  $k$ , therefore NDCG<sub>k</sub> is calculated both as an average across queries and at the query level [34].

### 3.5 Explanation of Retrieved Results

Because rankings are just a first step for decision-making, it is necessary to understand why the guidelines were ranked a certain way. While text embedding models are not inherently interpretable, explainability techniques can help demonstrate how specific words contribute to the model’s output score. An occlusion-based procedure, which iteratively masks each token of the retrieved result and scores the masked result with respect to the query, was used for importance attribution. Importance attribution was computed based on changes to the score when a word’s tokens were masked compared to the full query (e.g., if the score decreased when the tokens were masked, the word was considered important in retrieving the result). This procedure was implemented using Python and the *xtai* library [35] for the text embedding models. Since the text generation model (*Llama-4*) directly outputs scores for the task, it does not produce embeddings to assess using the occlusion procedure. However,

in comparison to text embedding models, this model can be instructed to produce natural language reasoning for its scores. Therefore, the generated explanations were examined instead.

## 4. RESULTS AND DISCUSSION

The results demonstrate the alignment of model-generated rankings with human-determined relevance ratings. We evaluate how different models perform and the impact of the information contained in the query. Additionally, we explore explainability and discuss the benefits and limitations of the overall approach.

### 4.1 Quantitative Evaluation of Models and Queries

The average NDCG score (across all queries) for each of the models, applied to each dataset, is shown in Table 4. A higher NDCG score indicates that the model more highly ranked guidelines with larger human-determined relevance ratings. The score is reported for  $k = 5$  (i.e., the top 5 results out of 39) because it is likely that users of a decision support tool would pay closer attention to guidelines returned in the top few ranks.

For the first dataset consisting of electrical appliances (vacuum cleaner, hand blender, etc.), the *Llama-4-Maverick-17B-128E-Instruct-FP8* model returned the guideline ranking most aligned with human ratings. For this dataset, using only the product type name was sufficient to achieve the highest alignment. Including product component information only improved alignment for specific text embedding models. For the second dataset consisting of electronic devices (smart phone,

**TABLE 4.** Average normalized discounted cumulative gain (NDCG) scores and standard deviation across all queries, considering each dataset and the different query representations (Product: “{product type}” or Product + Components: “{product type} made of {components}”)

NDCG <sub>k=5</sub>				
Model	Data Source 1 – Electrical appliances [25]		Data Source 2 – Electronic devices [26]	
	Product	Product + Components	Product	Product + Components
<b>Text Embedding</b>				
<b>all-mpnet-base-v2</b>	0.62 ± 0.28	0.70 ± 0.21	0.56 ± 0.32	0.50 ± 0.20
<b>multi-qa-mpnet-base-dot-v1</b>	0.64 ± 0.28	0.74 ± 0.23	0.50 ± 0.26	0.55 ± 0.13
<b>allenai-specter</b>	0.76 ± 0.18	0.74 ± 0.19	0.60 ± 0.15	0.66 ± 0.21
<b>ms-marco-MiniLM-L12-v2</b>	0.61 ± 0.17	0.70 ± 0.23	0.55 ± 0.023	0.45 ± 0.17
<b>SFR-Embedding-Mistral</b>	0.74 ± 0.14	0.71 ± 0.18	0.49 ± 0.17	0.57 ± 0.18
<b>Text Generation</b>				
<b>Llama-4-Maverick-17B-128E-Instruct-FP8</b>	0.91 ± 0.08	0.85 ± 0.11	0.58 ± 0.15	0.57 ± 0.20

laptop, etc.), the *allenai-specter* embedding model performed the best compared to human ratings. For this dataset, including the component information along with the product type led to the highest alignment. Again, including the component information did not universally improve alignment and was instead highly dependent on the specific model. While the general-purpose pre-trained models were able to perform the task reasonably well for the common household electrical appliances, NDCG scores were notably lower for the electronic device dataset. Perhaps unsurprisingly, the model (*allenai-specter*) trained specifically for a technical task (though not for circular economy or design) was better aligned with human ratings for this electronic device dataset compared to other models.

Another point to note was the high standard deviation in scores, indicating that results varied significantly by query (i.e., product type). The variation in guidelines retrieved was dependent on the product representation (i.e., the amount of information about the product included in the query) for each model type. However, the combination of those factors could also change based on the specific product type.

#### 4.2 Examples of Retrieved Results and Explanations

Laptops, laptop hard drives, and smart phones are examples examined in detail to better understand the types of guidelines retrieved using semantic search. The laptop category was chosen because there were several common components across the 16 different laptops, but each specific product also contained a unique set of components indicated by the BOM. The smart phone category was chosen because, in contrast, although there were 12 different smart phones in the dataset [26], their components were described in a consistent way in the BOM (all having a main body with a casing, circuit board, and battery, and a display). Finally, the laptop hard drive was included to investigate the application of guidelines at the component level.

**Laptop:** Figure 2 shows the laptop category, focusing on how subtle changes in the query information impact retrieved results

(using *allenai-specter*). The top 5 guidelines retrieved when the query contains the product type only vs. the product type and its components have no overlap. Including component information can subtly distinguish between the different laptops. When including components, there is complete overlap between the top 5 guidelines retrieved, but each rank order is slightly different.

**Laptop hard drive:** Product components reflect variation within a product type, while material variation is reflected in the components themselves. Demonstrating the impact of material variation, Figure 3 shows the retrieved guidelines for the hard drive component of a laptop and its materials, sourced from the provided bill of materials [26]. Retrieval using *multi-qa-mpnet-base-dot-v1* shows that adding material information to the query results in new guidelines from the Material Recycling guideline group appearing in the top 5. The positive importance attribution of the word “material” substantiates that they have likely been retrieved because of the additional information in the query.

**Smart phone:** Figure 4 shows the results returned by two different models (*allenai-specter*, which is specifically for scientific paper-related tasks, and *multi-qa-mpnet-base-dot-v1*, which is for general Q&A) when provided with both query types for the smart phone category. For these two models, there is only one overlapping guideline across the query types (“Use easily accessible joints” or “Adopt modular designs”), demonstrating the impact of including components in the query across different models. For *allenai-specter*, the guidelines retrieved with just the product type span multiple guideline groups (e.g., Extension of life span, Disassembly, Material recycling). When components are included, the top retrieved guidelines are primarily focused on disassembly. A focus on “Connectors” is further highlighted by its positive importance attribution (removing this word decreases the retrieval score of its containing guideline).

**Smart phone (text generation):** *Llama-4-Maverick-17B-128E-Instruct-FP8*, the text generation model, does not align best with

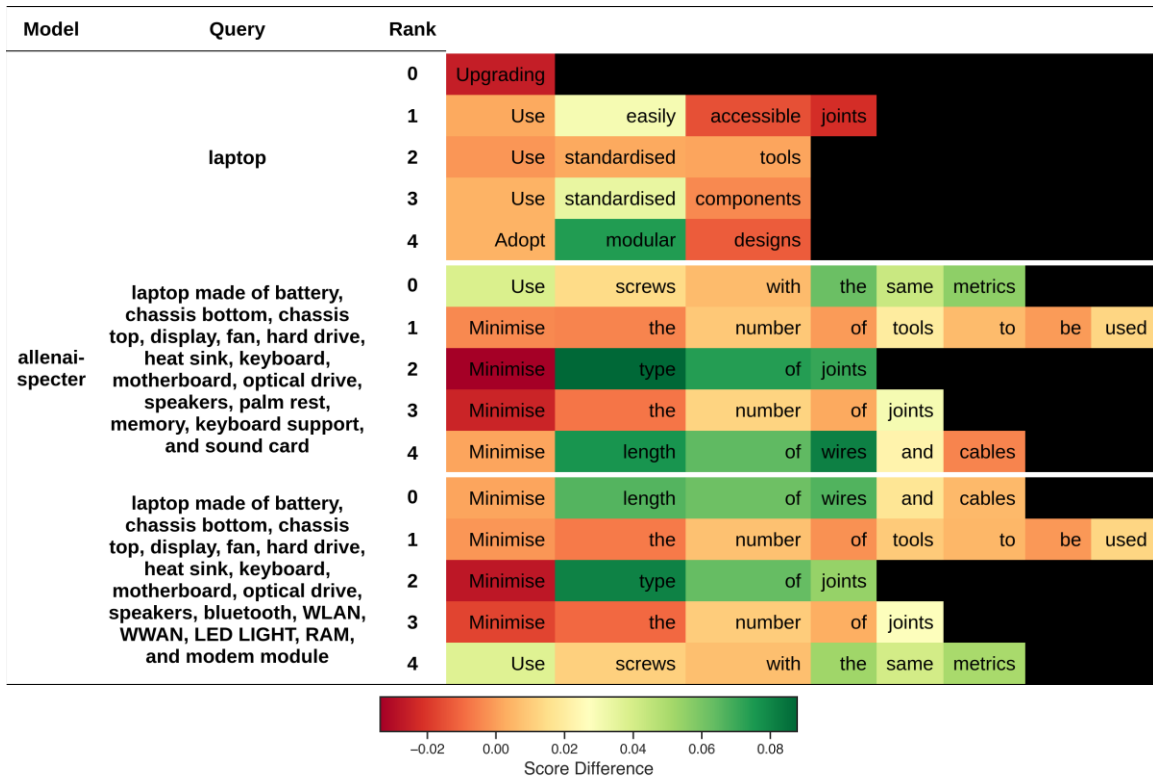


FIGURE 2. Top 5 guidelines retrieved from the *laptop* queries (comparing inputs of the same product type with varying components) using two of the text embedding models, along with word importance attributes (positive indicates more important)

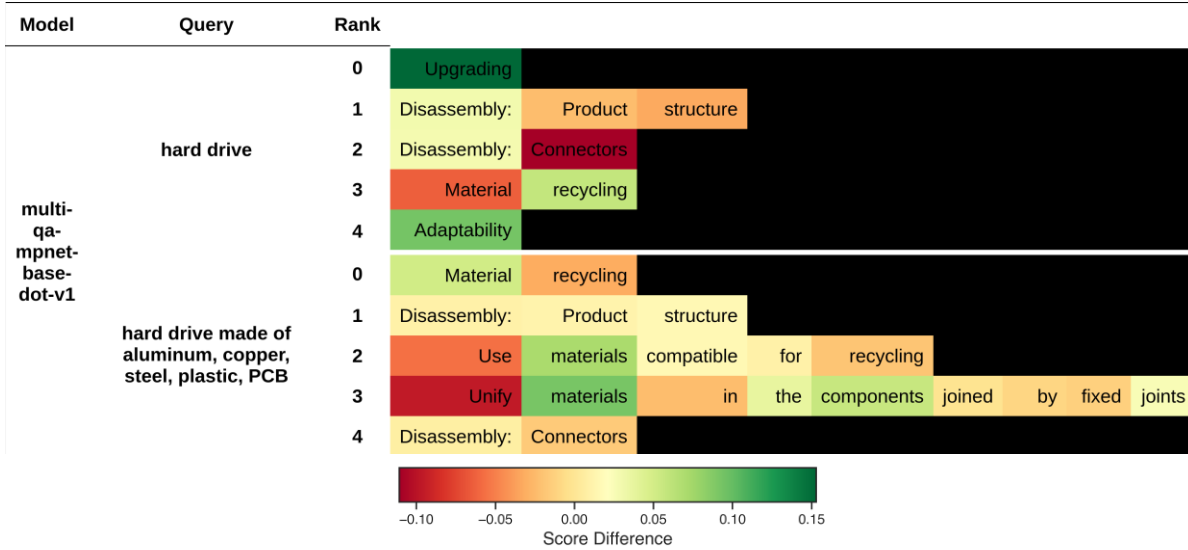
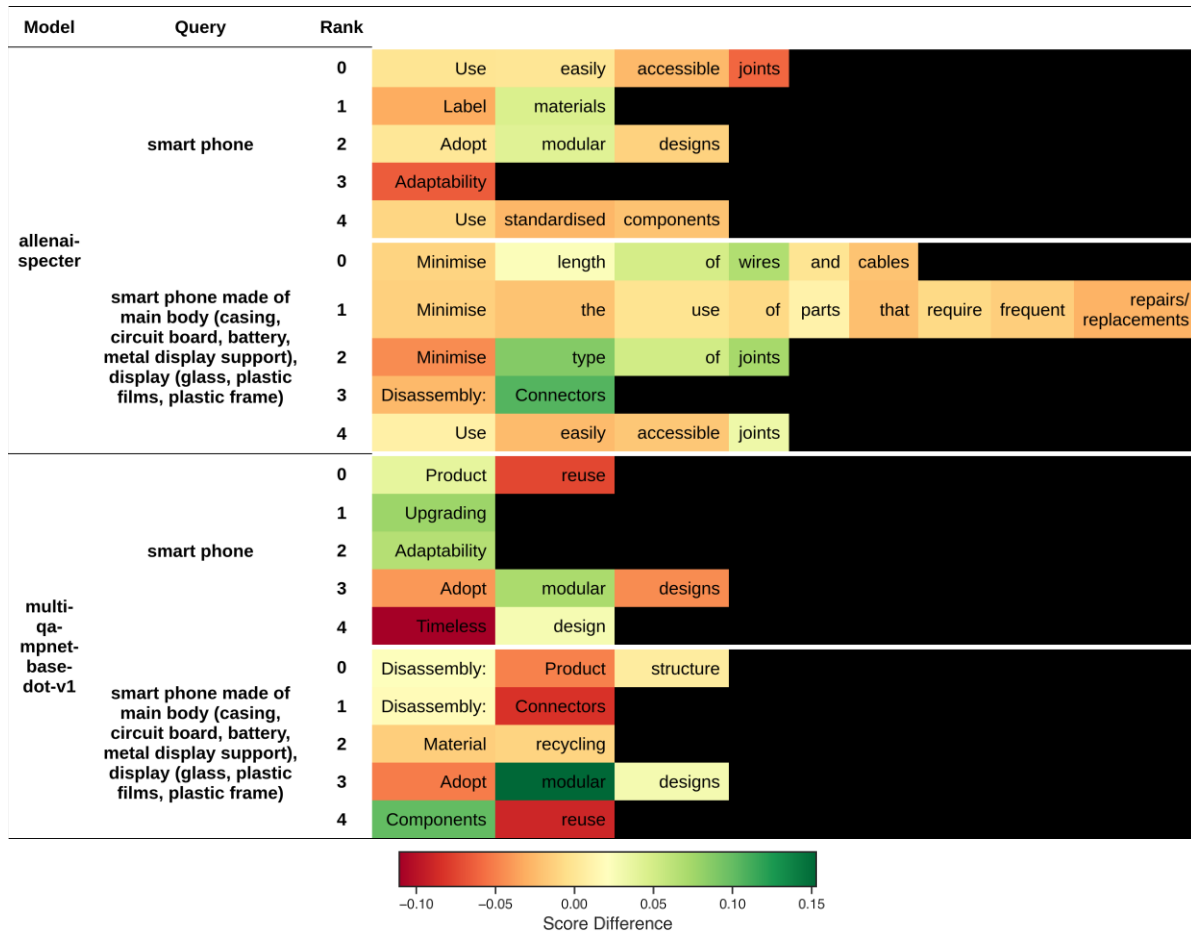


FIGURE 3. Top 5 guidelines retrieved from the *hard drive* queries (comparing inputs consisting of the component only or the component and its materials) using two of the text embedding models, along with word importance attributes (positive indicates more important)

human relevance ratings for the electronic device dataset, which includes the smart phone and laptop [26]. Furthermore, the model can pose problems due to stochasticity in its outputs. However, a benefit of text generation models is their ability to generate natural language explanations of their rankings. Table 5 shows the top 5 guidelines ranked by *Llama-4-Maverick-17B-*

*128E-Instruct-FP8* for a smart phone, along with a model-generated explanation. The current explanations are somewhat generic but demonstrate the potential for enhancing interpretability of automated decision support for circular design. Since guideline ranking is a first step to a longer process, interpretability may be valued over perfect alignment during



**FIGURE 4.** Top 5 guidelines retrieved from the *smart phone* queries (comparing inputs consisting of the product type only or the product type and its components) using two of the text embedding models, along with word importance attributes (positive indicates more important)

practical implementation.

### 4.3 Implications and Future Work

While organizations may desire to move towards circularity through design, identifying the right design strategies to implement for their specific products remains challenging. The goal of this study was to explore how computational approaches – specifically language models – can support the application of CPD guidelines. Approaches such as the one considered here have the potential to be integrated into design decision support tools, lowering the barriers to implementing circular design.

We used various pre-trained language models to rank CPD guidelines for various product types, finding that the ranking is influenced by the types of information used to describe the product. For example, it appears that including information about components can increase the ranks of guidelines related to disassembly, and including information about materials can increase the ranks of guidelines related to recycling. Therefore, identifying the right input query is critical for specifying system behavior when leveraging these systems for decision-making. Furthermore, in this task, guideline retrieval was sensitive to model choice, motivating the need to develop approaches

grounded in domain-specific data rather than solely relying on internal model knowledge (i.e., pre-trained model weights). Finally, the utility of the results – automatically prioritized guidelines – need to be evaluated in practice, considering the number and types of guidelines designers can feasibly implement. Such refinements would ensure improvements in alignment with designers’ intent and the overall utility of the investigated approach.

While the study indicates the potential to support CPD through LLM-based guideline ranking, several limitations are noted. Developing “ground-truth” relevance ratings requires expertise regarding all life cycle stages for each product type. This challenge is a motivation for automating guideline ranking in the first place but presents a limitation for feasibly validating the approach. Evaluating results against a single set of human-determined ratings (sourced from literature and through data augmentation here) is likely insufficient and alternative benchmarking mechanisms should be considered. Furthermore, it may be necessary to expand the information in the query (beyond product type or component names) and the set of design principles (beyond the short statements) to provide more context to the language models. For example, product-related technical

**TABLE 5.** Guidelines from Bovea and Pérez-Belis [25] retrieved from the *smart phone* queries (comparing inputs of the product type only or the product type and its components) using the *Llama-4-Maverick-17B-128E-Instruct-FP8* text generation model for scoring and explanation generation

Query	Model-Generated Rank	Model-Generated Explanation
<b>smart phone</b>	1. “Adopt modular designs” [DPS1]	“Highly beneficial for smartphones as it facilitates upgrading, repair, and component replacement.”
	2. “Use materials with a low environmental impact (recyclable/low energy content/etc.)” [MR3]	“Highly relevant for reducing the overall environmental footprint of smartphones.”
	3. “Upgrading” [ELS3]	“Very relevant for smartphones as it enables users to improve performance or add new features without replacing the entire device.”
	4. “Incorporate systems to monitor failing components” [PR5]	“Highly relevant for predicting and preventing component failure.”
	5. “Minimise using hazardous materials” [MR7]	“Critical for reducing environmental and health risks associated with smartphone production and disposal”
<b>smart phone made of main body (casing, circuit board, battery, metal display support), display (glass, plastic films, plastic frame)</b>	1. “Incorporate systems to monitor failing components” [PR5]	“Highly relevant as it can predict and prevent failures, enhancing the product's lifespan and reducing waste.”
	2. “Product reuse” [PR]	“Reusing smartphones or their components can significantly reduce electronic waste.”
	3. “Components reuse” [CR]	“Reusing components can significantly reduce the demand for new raw materials and decrease waste.”
	4. “Adopt modular designs” [DPS1]	“Modular designs can significantly enhance the upgradability, repairability, and recyclability of smartphones.”
	5. “Extension of life span” [ELS]	“A crucial guideline for smartphones as extending their lifespan can significantly reduce electronic waste.”

technical documents can be leveraged for more informative queries. Additionally, expansions to ASTM E3461 [5], a new standard with a rigorously-reviewed set of CPD principles, will contain more detailed guideline definitions. These sources could improve semantic matching by providing better context. In the future, computational representations of designs in modalities other than text (e.g., 3D models) could enable the surfacing of relevant circular guidelines as products are being designed, fitting more seamlessly into existing design workflows and tools.

Finally, guideline prioritization is only a first step towards product circularity. After a designer has gained familiarity with circular design principles, they may no longer need assistance surfacing relevant guidelines for their domain. Instead, they may need support on downstream steps, such as calculating performance indicators to assess the quantitative impacts of design decisions on circularity. Future work can link this quantitative assessment directly with guideline prioritization.

## 5. CONCLUSION

While circular product design (CPD) guidelines are becoming available for designers striving to create circular products, methods are needed to operationalize these guidelines so they can be applied to different design contexts, sectors, and product types. This work explores the application of pre-trained language models to rank CPD guidelines for their relevance to various product types within the electronics sector, demonstrating the potential for decision support tools to facilitate circular design. Factors such as information provided about the product (e.g., its

components) and the model type influence the alignment of automated guideline rankings with guidelines determined to be relevant by humans. While further investigation is necessary to comprehensively evaluate the integration of such models into circular decision-making, the findings provide a baseline upon which future work can build, highlighting avenues for further exploration.

*Software Disclaimer:* Certain software is identified in this paper to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the software identified is necessarily the best available for the purpose.

## REFERENCES

- [1] V. Ferrero, K. C. Morris, and B. Hapuwatte, “Adapting Consumer Product Design to the Circular Economy,” in *Volume 3B: 48th Design Automation Conference (DAC)*, St. Louis, Missouri, USA: ASME, Aug. 2022, p. V03BT03A049. doi: 10.1115/DETC2022-89542.
- [2] N. Mathur, N. Last, and K. C. Morris, “A process model representation of the end-of-life phase of a product in a circular economy to identify standards needs,” *Front. Manuf. Technol.*, vol. 3, p. 988073, Apr. 2023, doi: 10.3389/fmtec.2023.988073.
- [3] Ellen MacArthur Foundation, “We need to radically rethink how we design.” Accessed: Apr. 27, 2025. [Online]. Available: <https://www.ellenmacarthurfoundation.org/introduction-to->

circular-design/we-need-to-radically-rethink-how-we-design

[4] *Method to achieve circular designs of products*, EN45560:2024, Nov. 22, 2024.

[5] ASTM, *Standard Guide for Principles of Circular Product Design*, E3461-25.

[6] A. Mestre and T. Cooper, “Circular Product Design. A Multiple Loops Life Cycle Design Approach for the Circular Economy,” *Des. J.*, vol. 20, no. suppl, pp. S1620–S1635, Jul. 2017, doi: 10.1080/14606925.2017.1352686.

[7] B. M. Hapuwatte and I. S. Jawahir, “Closed-loop sustainable product design for circular economy,” *J. Ind. Ecol.*, vol. 25, no. 6, pp. 1430–1446, Dec. 2021, doi: 10.1111/jiec.13154.

[8] M. F. Aguiar and D. Jugend, “Circular product design maturity matrix: A guideline to evaluate new product development in light of the circular economy transition,” *J. Clean. Prod.*, vol. 365, p. 132732, Sep. 2022, doi: 10.1016/j.jclepro.2022.132732.

[9] B. Hapuwatte, N. Last, S. Karsli, G. Aher, K. C. Morris, and V. Ferrero, “From Principles to Practice: Product Design Foundations for a Circular Economy,” in *Volume 1: Acoustics, Vibration, and Phononics; Advanced Design and Information Technologies*, Portland, Oregon, USA: ASME, Nov. 2024, p. V001T02A044. doi: 10.1115/IMECE2024-146537.

[10] IDEO and Ellen MacArthur Foundation, “Circular Design Guide,” Sep. 2016.

[11] W. Ng and A. R. Vempati, “Cisco’s Circular Economy Journey and Embedding Circularity into Product Design,” Cisco, Jun. 2024.

[12] Circular Electronics Partnership, “Circular Electronics Design Guide,” Oct. 2024.

[13] J. Ko, G. B. Guedes, F. Badurdeen, I. S. Jawahir, K. C. Morris, and V. Ferrero, “Transitioning Towards Circular Consumer Electronics Products,” in *Volume 5: 29th Design for Manufacturing and the Life Cycle Conference (DFMLC)*, Washington, DC, USA: ASME, Aug. 2024, p. V005T05A003. doi: 10.1115/DETC2024-147949.

[14] Nike, “Circularity Workbook: Guiding the Future of Design.”

[15] ASOS and UAL Centre for Sustainable Fashion, “ASOS Circular Design Guidebook,” Nov. 2021.

[16] Association of Plastic Recyclers, “APR Design® Guide Overview,” APR Design® Guide Overview. Accessed: May 05, 2025. [Online]. Available: <https://plasticsrecycling.org/apr-design-hub/apr-design-guide-overview/>

[17] L. Siddharth, L. Blessing, and J. Luo, “Natural language processing in-and-for design research,” *Des. Sci.*, vol. 8, p. e21, 2022, doi: 10.1017/dsj.2022.16.

[18] S. Jiang, J. Hu, K. L. Wood, and J. Luo, “Data-Driven Design-By-Analogy: State-of-the-Art and Future Directions,” *J. Mech. Des.*, vol. 144, no. 2, p. 020801, Feb. 2022, doi: 10.1115/1.4051681.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 24, 2019, *arXiv: arXiv:1810.04805*. doi: 10.48550/arXiv.1810.04805.

[20] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence

Embeddings using Siamese BERT-Networks,” Aug. 27, 2019, *arXiv: arXiv:1908.10084*. doi: 10.48550/arXiv.1908.10084.

[21] OpenAI *et al.*, “GPT-4 Technical Report,” Mar. 04, 2024, *arXiv: arXiv:2303.08774*. doi: 10.48550/arXiv.2303.08774.

[22] H. Touvron *et al.*, “LLaMA: Open and Efficient Foundation Language Models,” Feb. 27, 2023, *arXiv: arXiv:2302.13971*. doi: 10.48550/arXiv.2302.13971.

[23] H. S. Walsh and S. R. Andrade, “Semantic Search With Sentence-BERT for Design Information Retrieval,” in *Volume 2: 42nd Computers and Information in Engineering Conference (CIE)*, St. Louis, Missouri, USA: ASME, Aug. 2022, p. V002T02A066. doi: 10.1115/DETC2022-89557.

[24] N. Preuss, A. S. Alshehri, and F. You, “Large language models for life cycle assessments: Opportunities, challenges, and risks,” *J. Clean. Prod.*, vol. 466, p. 142824, Aug. 2024, doi: 10.1016/j.jclepro.2024.142824.

[25] M. D. Bovea and V. Pérez-Belis, “Identifying design guidelines to meet the circular economy principles: A case study on electric and electronic equipment,” *J. Environ. Manage.*, vol. 228, pp. 483–494, Dec. 2018, doi: 10.1016/j.jenvman.2018.08.014.

[26] C. W. Babbitt, H. Madaka, S. Althaf, B. Kasulaitis, and E. G. Ryen, “Disassembly-based bill of materials data for consumer electronic products,” *Sci. Data*, vol. 7, no. 1, p. 251, Jul. 2020, doi: 10.1038/s41597-020-0573-9.

[27] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “Mpnnet: Masked and permuted pre-training for language understanding,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 16857–16867, 2020.

[28] *SentenceTransformers Documentation*. Accessed: May 05, 2025. [Online]. Available: <https://www.sbert.net/index.html>

[29] S. Y. Rui Meng Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, “SFR-Embedding-Mistral: Enhance Text Retrieval with Transfer Learning.” 2024. [Online]. Available: <https://www.salesforce.com/blog/sfr-embedding/>

[30] K. Enevoldsen *et al.*, “MMTEB: Massive Multilingual Text Embedding Benchmark,” *ArXiv Prepr. ArXiv250213595*, 2025, doi: 10.48550/arXiv.2502.13595.

[31] Hugging Face, “Embedding Leaderboard.” [Online]. Available: <https://huggingface.co/spaces/mteb/leaderboard>

[32] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, “SPECTER: Document-level Representation Learning using Citation-informed Transformers,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: ACL, 2020, pp. 2270–2282. doi: 10.18653/v1/2020.acl-main.207.

[33] C. Burges *et al.*, “Learning to rank using gradient descent,” in *Proceedings of the 22nd international conference on Machine learning - ICML '05*, Bonn, Germany: ACM Press, 2005, pp. 89–96. doi: 10.1145/1102351.1102363.

[34] S. K. Karmaker, P. Sondhi, and C. Zhai, “Empirical Analysis of Impact of Query-Specific Customization of nDCG: A Case-Study with Learning-to-Rank Methods,” in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Virtual Event Ireland: ACM, Oct. 2020, pp. 3281–3284. doi: 10.1145/3340531.3417454.

[35] *txtai*. [Online]. Available: <https://neurol.github.io/txtai/>