DRS BOSTON 20 DRS 24 JUNE 24 — 28

Exploring human-centered design method selection strategies with large language models

Vivek Rao[a,b,*], Yuanrui Jerry Zhu[c], Timothy Yang[c], Euiyoung Kim[d], Alice Agogino[c], Kosa Goucher-Lambert[c]

[a]Duke University, Pratt School of Engineering, USA
[b]University of California, Berkeley, Haas School of Business, USA
[c]University of California, Berkeley, Dept. of Mechanical Engineering, USA
[d]Delft University of Technology, Dept. of Design, Organization and Strategy, the Netherlands

*Corresponding author e-mail: vivek.rao@duke.edu

doi.org/10.21606/drs.2024.956

Abstract: In human-centered design (HCD) projects, designers select and use a variety of design methods in pursuit of a desired outcome. Given the prominence of method selection in designer behavior, what distinguishes a design team's method selections from design method selection based on frequency or probability? To explore this question, we compare HCD methods suggested by the publicly-available large-language model, GPT-3.5, to 402 novice design team method selections over five offerings of a design projectbased learning course at a large public university. We observe that GPT-3.5 appears to represent design method knowledge held in method repositories like theDesignExchange well. We also observe that GPT-3.5's method selection recommendations appear to poorly distinguish between HCD phases, and appear limited to highly specific aspects of HCD phases. These findings highlight the unique contribution of human design cognition in design decision-making relative to LLM's, and herald the promise of human-AI teaming in design method selection.

Keywords: design theory and methodology; human-AI collaboration; human-centered design; design methods.

1. Introduction

Design plays a role in problem-solving in domains ranging from engineering to public health, to law, and beyond (Bazzano et al., 2017; Brown, 2008; Kim et al., 2018; Verganti, 2009). Central to this diversity of applications are design methods (Avle et al., 2017; Chasanidou et al., 2014; Geis et al., 2008; Roschuni et al., 2015), which allow practitioners to adapt design

approaches to a specific context. Gericke et al. describe a design method as "a specification of how a specified result is to be achieved" in design (Gerrike et al., 2017). Design methods are a hallmark of human-centered design (HCD), an approach that has been adopted widely in both education and practice (Kramer et al., 2016; Tomiyama et al., 2009). An understanding of the prevalence of design methods (Fuge & Agogino, 2015; Kramer et al., 2017) has informed work seeking to understand *why* design methods are chosen, particularly the patterns (Fuge & Agogino, 2014; Poreh et al., 2018) and decision-making strategies (Rao et al., 2021) underlying these selections. Given their widespread usage, design methods provide an intriguing area to explore what Lloyd et al. have described as dialogues between humans and artificial intelligence (AI) systems – two way, and often multimodal, collaborations in pursuit of design outcomes (Lloyd et al., 2022).

In this preliminary work, we explore how design method selections among novice design teams differ from design method recommendations from probabilistic synthesis of general information, here represented by large language models (LLMs), neural networks trained on large quantities of textual data (Brants et al., 2007; Manning, 2022; Radford et al., 2019) to explore the following research questions:

> **R1.** Are methods selected by design teams similar or different to methods recommended by an LLM?
>
> **R2.** What similarities and differences exist between human- and LLM-sourced rationales for method selections?

We employ the following methodology:

- **Data Collection.** We source novice human design teams' method selections and selection justifications from five separate offerings of a project-based engineering design course. Separately, we use the OpenAI GPT-3.5 API to generate design method recommendations in response to project- and design phase-specific prompts reflective of the novice teams. GPT-3.5 has no constraints on the methods available to it, in contrast to the novice design teams, who must select from a curated set of 12-18 methods that vary across design phase.
- **Summary Data Analysis of Method Selections.** Methods selected by novices and recommended by GPT-3.5 are examined for prevalence and differences across phase and project type.
- **Quantitative Data Analysis of Method Selection Justifications.** Latent Dirichlet Analysis (LDA) is used to identify prevalent topics among justifications for method selections.

The key contributions of this work are a comparison of human and LLM-advised design method selections, quantification of differences in method prevalence, and a preliminary comparison of design team- and LLM-sourced method selection justifications. We note that while this study focuses on an engineering design course with students of many disciplines,

the practice of human-centered design methods spans design disciplines, and we believe this investigation holds transferable insight to these other, adjacent disciplines.

2. Related work

2.1 Design methods in the design process

As Daalhuizen et al. define it, 'design methods' are "formalised [representations] of a design activity" that can "support designers in achieving a goal" (Daalhuizen et al., 2019; Daalhuizen & Cash, 2021). Design methods have long been central to design research (Jones, 1992), with design methods acknowledged as a key contribution of design research (Cash et al., 2023; Gericke et al., 2022). Method repositories such as TheDesignExchange (tDX) and IDEO's Method Cards have found success in making design methods accessible to practitioners (Kramer et al., 2017; *Method Cards*, n.d.). Despite their prominence in research, design methods have been critiqued for their lack of 'transfer' to practice (Jagtap et al., 2014). To address this, active research explores what constitutes a 'good' method (Cash et al., 2023; Gericke et al., 2022)).

How and why methods are *selected* is a key dimension of methods' transfer and ultimate utility. Gericke et al. found that a reason professionals selected methods was their prior experience with the method (Gericke et al., 2016). Novice designers, in contrast, appear to have a range of motivations for choosing methods, many of which evolve over the course of a design project (Kim et al., 2022; Poreh et al., 2018; Rao et al., 2021).

In this work, we extend on prior research to consider the unique role that LLMs may play in supporting or automating method selection. Preliminarily, we examine both the *what* of methods – which methods are selected differentially by human and LLM's – and the *why* – the justification for the selection.

2.2 Artificial intelligence in engineering design

How designers can leverage ML and AI, and how ML and AI are reshaping designers are a central area of design research (B. Song et al., 2022). The implications of ML and AI for design have been described elsewhere (Allison et al., 2021; Panchal et al., 2019); applications have been demonstrated in accelerating numerous types and aspects of design, from automating UX design workflows (Yang, 2018) to accelerating concept development (Camburn et al., 2020). Recent work has illustrated how AI in 3D CAD systems can inform experts' and novices' modalities of inspiration (Kwon et al., 2021; Rao et al., 2022).

The recent widespread adoption of LLMs (Min et al., 2023) has opened new possibilities for AI in the design process through natural language processing and understanding. As of this writing, a predominant approach of many LLMs is to employ *generative pre-trained transformers* (GPT's); a comprehensive overview of transformer architecture and their generative applications in pre-trained contexts is provided by Radford et al. (Radford et al., 2018, 2019). Briefly, transformers rely on attention-based models, which provide more

context to a model by focusing on various parts of the input data sequence (Vaswani et al., 2017); the transformer architecture consists of parallel encoder and decoder layers that allow for efficient processing. Recent work by Ma et al. showed that LLMs generated more feasible concepts in an ideation exercise than human crowd workers on an online platform, but human workers' concepts had more novelty (Ma et al., 2023).

Two challenges of leveraging LLMs are salient. First, evaluating LLM results remains an area of exploration; a common method is to compare results with a ground truth or using a mature statistical test within the domain (Le Mens et al., 2023; Picard et al., 2023; Wan et al., 2023), while other approaches interpret the result more qualitatively (Goel et al., 2023). Ma et al. leveraged existing constructs adapted to domain-specific applications, such as the *convex hull volume metric* for representing the novelty of a set of design concepts, to evaluate LLM outputs (Ma et al., 2023). Second, biases are a significant risk in working with LLMs given their training methodology. Research has shown cultural, societal, and historical data may introduce bias during the training process of the model and be reflected on the model response (Kolisko & Anderson, 2023).

In this work, we seek to examine how LLMs can contribute to design method selection. One prior study, by Fuge et al., examined the application of ML to method recommendation (Fuge & Agogino, 2014); here, we extend on Fuge's work by applying OpenAI's GPT-3.5 LLM as opposed to recommender systems. We also extend on Ma et al.'s work by contrasting LLM-sourced design recommendations with human-sourced selections, applied here to method selections rather than concepts (Ma et al., 2023). To evaluate the outputs of LLMs, we examine prevalence of recommended methods, and use topic modeling to identify patterns across these recommendations as a whole (see Section 2.3).

2.3 Topic modeling in engineering design research

Topic modeling is an approach that seeks to provide deeper understanding of a given set of qualitative data, by identifying categories of the given data in an unsupervised manner. Among various topic modeling methods, Latent Dirichlet Allocation (LDA) has gained popularity because of its high efficiency and structural framework and its applicability to many fields of study (Feuerriegel & Pröllochs, 2021; Tran et al., 2019). LDA and similar approaches have limitations: they require large amounts of data; resulting topics require high levels of subjective evaluation to interpret; and a researcher has little control over the focus of resulting topics (Ornstein et al., 2022).

Recent applications of LDA in engineering design have included identifying different levels of design requirements (Chen et al., 2021), extracting insights from smartphone product and homeshare service reviews (Joung & Kim, 2021; Kiatkawsin et al., 2020), and improving the design of nursing beds (Yuan et al., 2023). These applications of topic modeling seek to enhance *design outcomes.* Automated semantic analyses, like LDA, have a long history of use in studying *design methodology* as well. In the specific context of analyzing human design data, Chan and Schunn showed that LDA could reveal insights about design conversations in

cross-cultural innovation teams, and argued that such an approach "[provides] formal patterns" to augment researchers' understanding of a given design activity (Chan & Schunn, 2017). Previously, Song et al. described how Latent Semantic Analysis (LSA) could be used to describe tacit knowledge in engineering design (S. Song et al., 2003), and Dong established LSA as a benchmark method for conceptualizing design team communication (Dong, 2005). Our previous work developed a qualitative codebook to describe a design team's method selection intentionality (Rao et al., 2021).

In this work, we leverage LDA to interpret the differences between novice design team- and LLM-sourced method selections. Our previous framework for describing design team intention in method selections is inappropriate given the questionable meaning of 'intention' of LLM-sourced justifications; accordingly, we use topic modeling to identify content-based themes within human- and machine-generated method selection justifications.

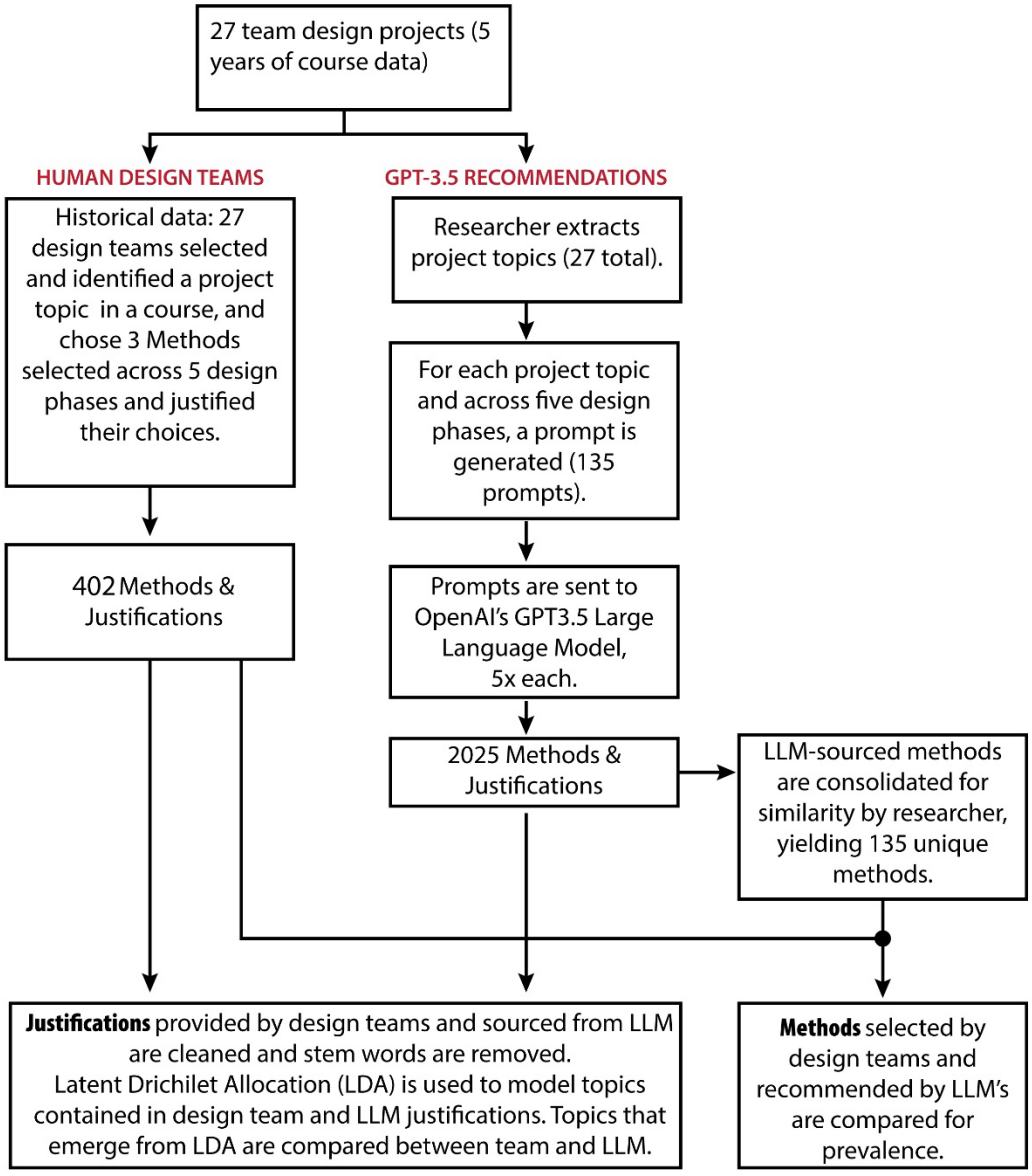

Figure 1 An overview of the research methodology and data sourcing.

3. Materials and methods

Our work contrasts method selections by students enrolled in an HCD project-based learning course at a major public university with methods advised by GPT-3.5, an LLM (Fig. 1).

3.1 Course & project details

Over five separate offerings of the course from 2017-2021, student teams identified a project topic (Table 1) to pursue, and were asked to select three HCD methods from a subset of 12-18 methods sourced from tDX for each of five phases underpinning the HCD process: Research, Analyze, Ideate, Build and Communicate (Kramer et al., 2016). Students chose these methods in conjunction with their progress through the HCD process, which they pursued over the course project. This resulted in a total of 402 design method selections and

justifications. Two researchers reviewed the student projects and gave them a theme - that is, the essence of the goal of the project as established by the student team. The researchers then classified projects by Ceschin and Gaziulusoy's innovation classification framework (Ceschin & Gaziulusoy, 2016), assigning an innovation level of *product, product-service,* or *spatio-social* to the project topic. This framework was chosen for its generalizability to a variety of innovation projects – which characterized the diversity of projects in the class – and for its thorough and highly-cited description of the distinction between the four levels of innovation. We note that the *socio-technical system* innovation level did not describe any projects in the data examined. Classification was based on both the scope of the student's design objective, and the design objective reached in the course (Table 1).

Table 1 Example Project Themes (5 of 27), as LLM Prompt Strings, and innovation type classifications

Project Theme, as LLM Prompt String	**Innovation Classification**
"reminding patients to take medicines and connect them to family members",	Product-Service
"finding available seating areas in public workspaces",	Spatio-Social
"helping manage electronic devices' wiring"	Product

3.2. LLM & prompt details

OpenAI's GPT-3.5 uses a neural network with a claimed 175 billion parameters, representing textual data scraped from the internet (*OpenAI Platform*, n.d.), presumably including tDX among many other sources. We used the GPT-3.5 API (mode choice: text-davinci-003) to generate five responses to a prompt that alternated both the project theme and the design phase (each underlined in the example below, and alternated).

> "In a project about helping organizations identify and respond to social engineering attacks, what are the best three human-centered design methods you recommend during the analyze phase? Format your answer with the name of each method followed by a colon, followed by a justification for why you recommend the method, followed by a | character."

As attention-based models rely on context for superior output, few-shot prompts (Brown et al., 2020) – that is, prompts with examples of desired output – will often yield more desirable responses from LLM queries. However, for this particular application – modeling novice designers' method selections – we believe zero-shot prompts, that is, prompts with no examples, are more representative of an actual use case, as novice designers presumably would not be able to give validated or meaningful examples to construct a few-shot prompt.

A total of 135 possible prompt combinations were queried five times each; multiple queries were used as LLM's can generate varying output given their probabilistic nature. GPT-3.5

generated a total of 2,025 method selections and justifications. These represented 254 different methods, consolidated into 135 separate methods upon review by one researcher. For example, the LLM recommended the following methods: “low-fidelity prototyping”, “prototyping,” “hands-on prototyping,” and “iterative prototyping,” all of which were consolidated to “prototyping” as these were similar to tDX’s definition of prototyping – and distinct from other prototyping methods specified in tDX, such as ‘experience prototyping’ and ‘tangible prototyping.’

3.3. Analysis & topic modeling model selection

Given the size of our data corpus of method selection justifications – approximately 100,000 total words across 2427 separate documents (Table 2) – a topic modeling approach was chosen. To do so, we used Latent Dirichlet Allocation (LDA), introduced in Section 2.3. LDA was proposed (Blei et al., 2003) as a generative probabilistic topic modeling approach and has been applied in a number of application areas (Griffiths & Steyvers, 2004; Jelodar et al., 2019). LDA is an appropriate method to (1) model topics existing among human and machine method selection justifications, and (2) comparatively interpret these topics based on these topics’ word probabilities. One important element is coherence score (Mimno et al., 2011), which measures the semantic similarity among high-scoring words within each topic, and higher coherence score means greater semantic coherence and interpretability of the topic; the score ultimately informs the number of topics for our LDA algorithm.

We first employ the Python package **gensim** to identify key topics in the method selection justifications provided by design teams and the LLM. Before fitting the model, we preprocessed our text (Table 2) by tokening the sentences into lists of words (**NLTK**) and removing stop words and punctuations referenced in NLTK’s stopwords corpus and Python’s string.isalnum() function. We then extracted stem words using nltk’s PowerStemmer. We obtained a bag-of-words format for our text. As the number of topics is well understood to be the critical driver of LDA performance, we have chosen to focus exclusively on tuning related hyperparameters in our study.

We chose the optimal number of topics based on the coherence score from **gensim’s** coherence model, which is calculated as a measure of the average similarity between all pairs of words within the topic, to measure the performance of the model. By choosing the number of topics corresponding to the highest coherence score among 3 to 10 topics, we determined our final number of topics as 5 for design team justifications, and 6 for LLM-sourced justifications. For the Prior for Document-Topic Distribution (alpha) and Topic-Word Distribution (beta), we utilized the default parameters of ‘Symmetric’ and ‘None’, respectively. We acknowledge the inherent subjectivity in interpreting LDA’s suggested topics; given our research team’s experience with the design methodology domain, we believe our interpretations are robust in capturing the essence of topics as a starting point.

Table 2 Preprocessing text details for method selection justifications.

	LLM	**Team**
Number of documents (method selection justifications)	2025	402
Number of words	66496	30328
Number of words after removing stop words	40515	16855
Number of nouns	18533	6619

4. Results and discussion

Four emerging takeaways resulted from our work. We present preliminary results exploring the diversity of methods proposed (4.1) and the phase- and project-sensitivity of methods recommended (4.2 & 4.3). Next, we present a preliminary result exploring emerging differences in topic models of design team- and LLM-sourced justifications (4.4). Before examining these results, we show method selections and justifications for one project team in one phase as an example (Table 3).

4.1 The LLM generates a similarly diverse range of methods as human-curated design repositories

While versions of all of GPT-3.5-recommended methods, after consolidation, are represented in the tDX repository, GPT-3.5-recommende methods do not account for all of the 300+ methods cataloged in tDX. We do note that in some areas, such as prototyping, the LLM does not recommend methods with the specificity of tDX. Nonetheless, we interpret this finding to suggest that LLM's provide a repository of diverse design methods knowledge comparable to human-curated sources.

4.2. Phase-specific design method curation appears distinguished from LLM recommendations, especially in the Analyze and Communicate phases

The LLM appears to distinguish the Research, Ideate, and Build phases by more frequently selecting methods that are primarily associated with each of those phases – matching human team selections. Examining the highest-proportion methods selected by phase (Fig. 3), 'interviews' represent the highest proportion (26.7%) of methods recommended by the LLM during the Research Phase; this is also the case for human design team method selections (18.4% of selections were the '1:1 Interview'). Results are similar for the Ideate phase – with the LLM recommending 'brainstorming' most frequently at 16% of selections, while 'brainstorming' represented 15.5% of selections by human design teams – and the build phase, with the LLM recommending 'prototyping' 15.1% of the time. However, for the Analyze and Communicate phases, the LLM appeared to diverge from the design teams. For the Analyze phase, the top method recommended by the LLM was "interviews" (18.8%), whereas design teams selected 'empathy maps" (16.7%) most frequently. For the

Communicate phase, the top method recommended by the LLM was “interviews” (16.5%), whereas human design teams most frequently selected “envisionment videos” (20.8%).

Table 3 Example design team method chosen and LLM method recommended. Only one method and corresponding justification is shown for both team and LLM.

Phase	Design Team Method Chosen	Design Team Justification	LLM Method Recommended	LLM Recommendation Justification
Analyze	Reframing	“Before any major group discussion, our team already had a general opinion that reframing is appropriate for us. So unlike with the other methods, we readily agreed upon this method.”	Focus Groups	“To get feedback from Users about their expectations and perceptions about the wide variety of features provided by Smart Home devices.”

We consider both the Analyze and Communicate phases more closely, looking at the top-5 most-recommended methods in each phase (Fig. 4a). The LLM predominantly advises methods that largely do not distinguish between design phases. This is demonstrated in the top five methods selected in the Communicate phase (Fig. 4b), a phase in which designers translate their solutions into stories to persuade and inspire stakeholders. While human designers selected methods like ‘storyboards,’ ‘envisionment videos’ and ‘roadmaps’ most frequently, GPT-3.5 advised ‘interviews,’ ‘personas,’ and ‘prototyping.’ These methods are specific to other phases in a traditional HCD process, and indeed have little specificity to the Communicate phase of HCD.

These findings suggest three insights. First, GPT-3.5 advisory in design method selection may struggle to distinguish between design phases: based on the above examination, it appears that GPT-3.5’s associations with the Analyze and Communicate phases do not reflect those phases’ actual core activities. Second, given that GPT-3.5 suggests particular methods frequently – e.g., ‘interviews’ – across all design phases, LLM’s may have a particular set of terms that are strongly associated with HCD that are heavily used in responses to queries regardless of any other nuance. From the method results, these appear to be terms like “interviews,” “brainstorming,” and “prototyping.”

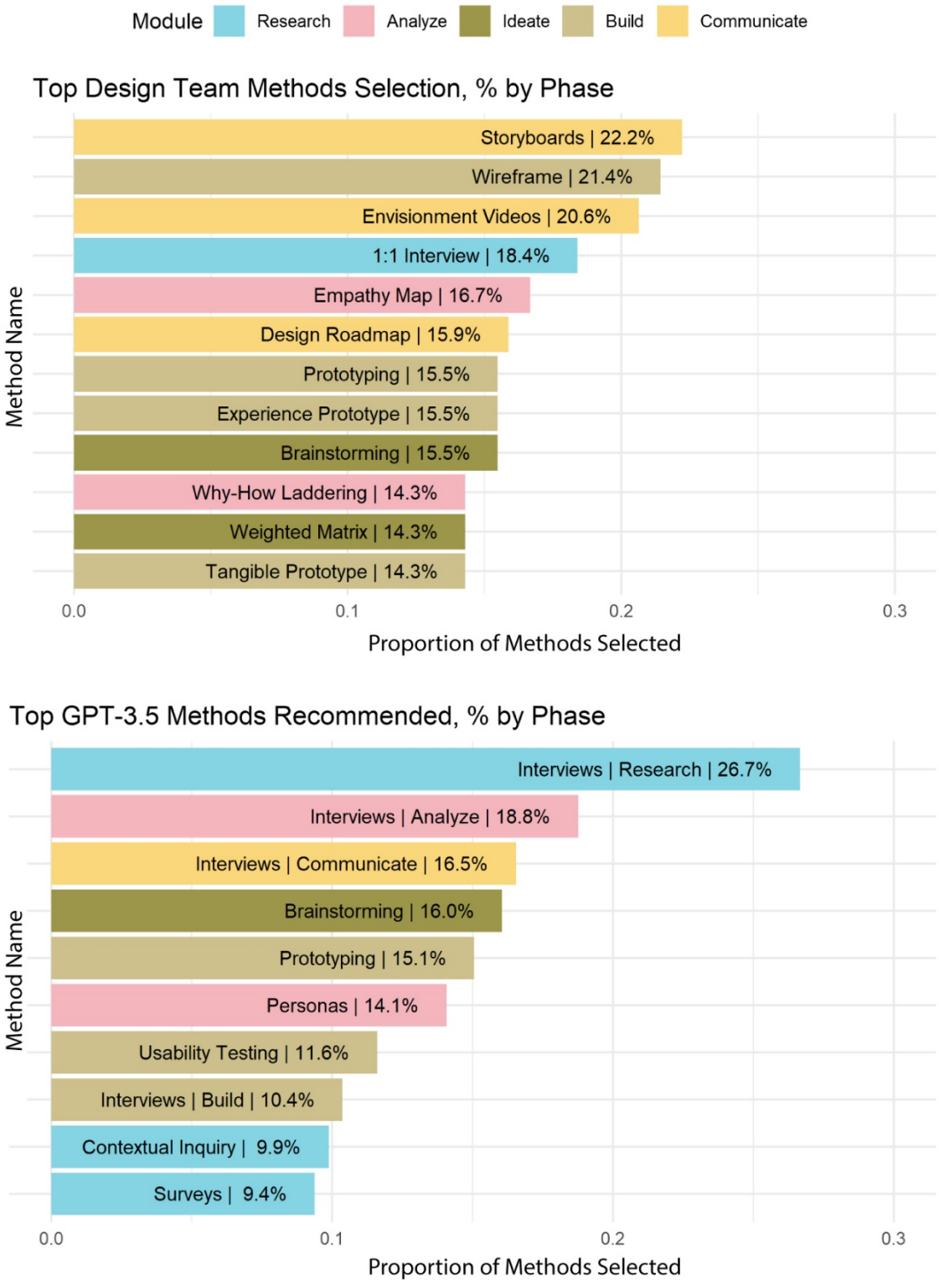

*Figure 3 Most-selected methods by design teams (**top**) and most-recommended methods by LLM (**bottom).** Bars are color-coded to correspond to the relevant design phase (legend at top), and proportion %'s represent the percentage of methods selected in a given phase. So, for*

the human team plot, Storyboards represent 22.2% of methods selected by all teams during the Communicate phase.

Third, we do note that although much of the LLM's advisory may be generic when it comes to phase, for 3/5 phases there is good alignment between top methods recommended by the LLM and human design teams. That suggests that LLM's, for design phases that are more immediately tangible– the Research, Ideate, and Build phases – their recommendations could be relevant to design teams in an advisory capacity. However, as mentioned in the previous insight, this does not necessarily mean that the LLM operates as an effective advisor – rather, that by sheer probability, relevant methods are associated with more tangible phases.

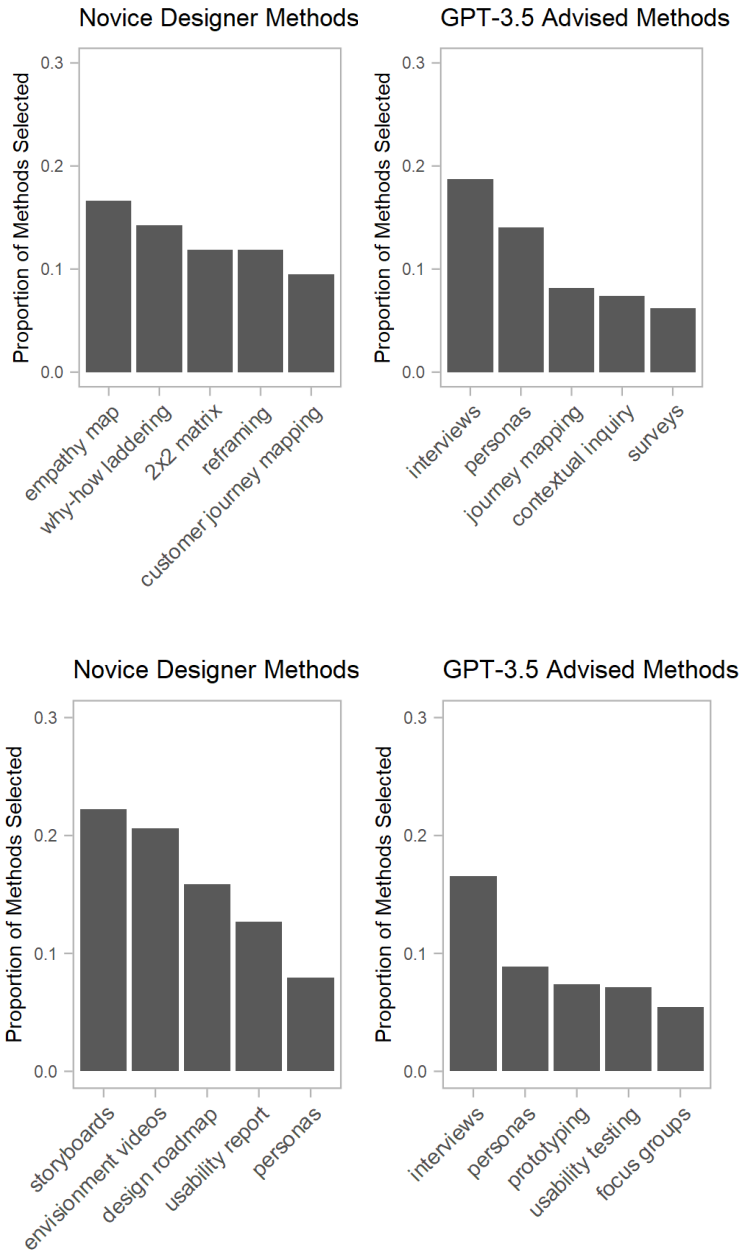

*Figure 5 Most-selected in the Analyze phase (**top panel**) and the Communicate phase (**bottom panel**).*

4.3. Design teams may more clearly distinguish between project types based on methods selected than the LLM distinguishes between them

Design teams may select different methods based on project types (Fig. 5). For example, we observe that teams working on projects classified in the Product type choose the empathy map most frequently (6.1% of all methods selected); this does not appear as a top method for the Product-Service or Spatio-Social project types (3.0% and 2.0% of all methods chosen, respectively). While no significant difference between the 'empathy map's' prevalence among the three project types was found ($p > 0.05$, holm-adjusted pairwise proportion test),

the frequency was notably different and warrants further investigation. Qualitatively, we observe a diversity of methods as the most popular across project types: with the exception of the '1:1 interview' and 'wireframe,' repeated for Product-Service and Spatio-Social projects, no top methods are shared among project types. In contrast, the LLM has two methods that appear for all three project types: interviews and prototyping. While this is driven by the insight from 4.2, that the LLM repeatedly suggests the same method across phases – leading to higher prevalence of a few methods – it also suggests that the LLM's recommendations may be less sensitive to project type than design team's selections.

These findings suggest further insight about LLM advisory in design support. Human design teams appear to select methods in ways that could be project type-specific. Examining top methods selected by design teams more closely, Product teams have three prototyping approaches among their top methods chosen; the other project types have none. While further study is needed to statistically validate this finding, it does stand in anecdotal contrast to LLM-recommended methods – which appear to be far less sensitive to project type. This possibly suggests that designers' immersion in their project offers them project-specific insight that can inform more relevant method selection than LLM-sourced recommendations.

4.4. Human design teams' justifications appear to focus on the design team, process, and outcome, while the LLM's justification focuses on the design process

Coherence modeling on design team justifications revealed the optimal number of LDA topics to be five. As we used unsupervised LDA, these five topics and their top ten words each (Table 4) were to be interpreted by our research team. Examining the LLM-sourced team justifications, coherence modeling revealed the optimal number of LDA topics to be six. As with the human design team data, these topics and their top ten words each (Table 4) required interpretation. We note that the ratio (%) indicates the average percentage of words in each topic's justification text represented by the top 10 keywords.

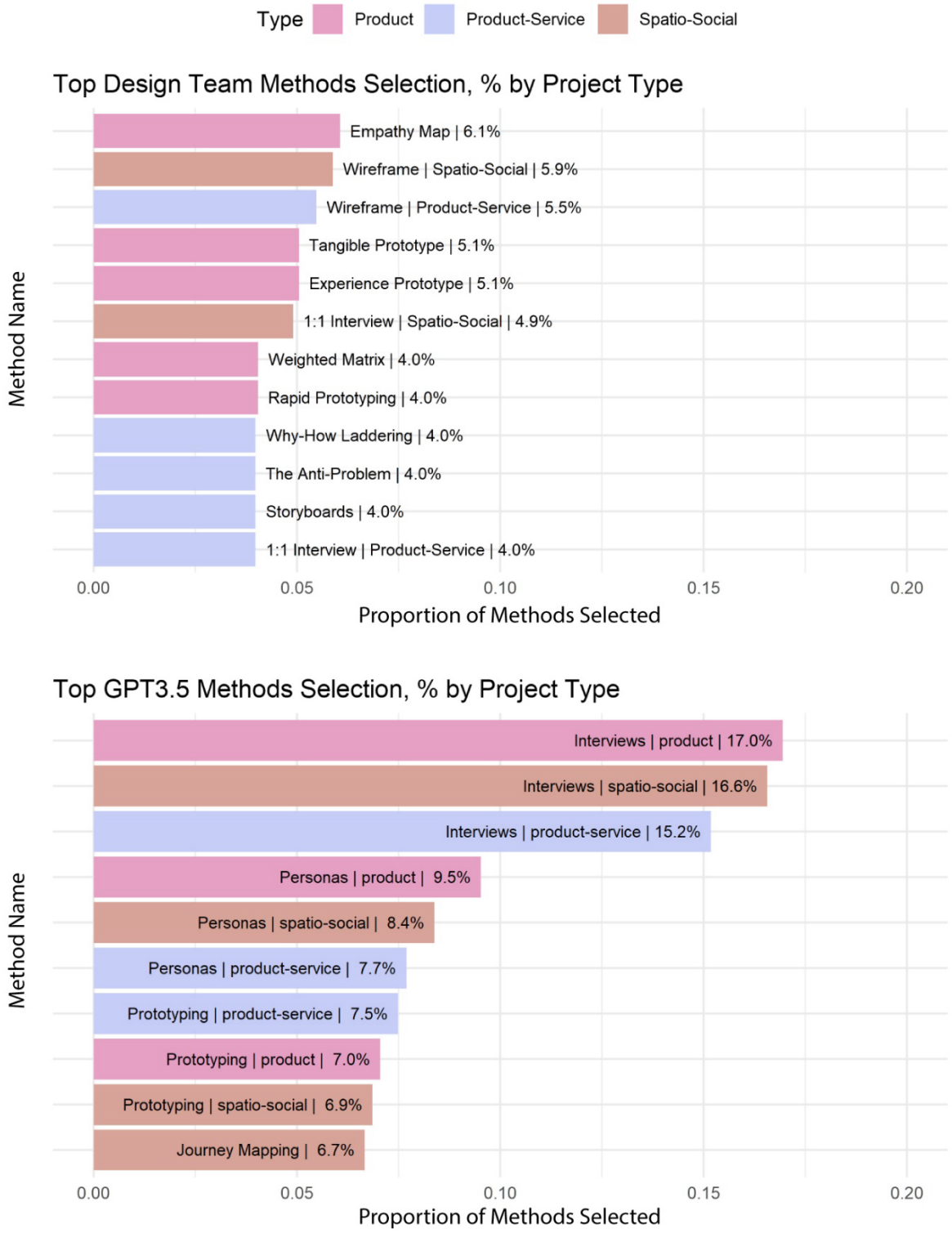

*Figure 6 Most-selected methods by design teams (**top**) and most-recommended methods by LLM (**bottom).** Bars are color-coded to correspond to the relevant design project type (legend at top), and proportion %'s represent the percentage of methods selected among selections of a given project type.*

Interpretation of LDA topics is a highly subjective process. However, our findings suggest one insight about LLM design advisory. We note that of the interpreted topics, human topics appear to address the design *process*, but also address how to enhance design *outcomes* and how to support the design *team*. Meanwhile, the LLM-sourced justifications appear to focus more exclusively on the design *process.*

Table 4 LDA Topics and Interpretations for Method Selection Justifications.

Category	Topic Interpretation	Topic #	Keywords	Ratio (%)
Machine	*methods chosen for helping address users' needs*	1	'user', 'help', 'need', 'secur', 'design', 'feel', 'contact', 'understand', 'solut', 'trace'	31.5
	unknown	2	'help', 'come', 'secur', 'way', 'user', 'variou', 'inform', 'servic', 'disclosur', 'explain'	29.6
	methods chosen for identifying users and opportunities	3	'identifi', 'user', 'help', 'address', 'team', 'improv', 'journey', 'map', 'project', 'point'	39.9
	methods chosen for understanding user needs	4	'user', 'understand', 'need', 'design', 'use', 'effect', 'context', 'help', 'product', 'system'	26.4
	methods chosen for aligning solutions with stakeholders	5	stakehold', 'solut', 'allow', 'group', 'process', 'creativ', 'idea', 'involv', 'togeth', 'bring'	32.1
	methods chosen for enhancing designs to help users and improve usability	6	'design', 'user', 'help', 'ensur', 'usabl', 'test', 'persona', 'allow', 'build', 'technolog'	34.2
Human	*methods chosen for prototyping purposes*	1	'method', 'us', 'prototyp', 'user', 'allow', 'servic', 'label', 'experi', 'decid', 'storyboard'	12.5
	methods chosen to enable discussion	2	'method', 'us', 'would', 'user', 'discuss', 'idea', 'design', 'help', 'use', 'agre'	15.4
	methods chosen for brainstorming and ideation	3	'idea', 'brainstorm', 'method', 'gener', 'team', 'discuss', 'product', 'decid', 'also', 'time'	13.4

	methods chosen for desired user outcomes	4	'method', 'us', 'user', 'use', 'design', 'product', 'idea', 'help', 'team', 'need'	19.1
	methods chosen for desired product outcomes	5	'use', 'product', 'method', 'prototyp', 'user', 'would', 'usabl', 'agre', 'us', 'help'	16.7

5. Implications for design research, practice & education

Human-AI teaming holds great potential to aid design method selection. While LLM's appear to capture design method knowledge, advisory of methods requires further nuance, either through prompt engineering, user education, or deeper human-AI collaboration.

5.1. Limitations

There are several limitations to this work; we focus on four most salient for the design research community. First, design teams' method selections were constrained to a prescribed set of 12-18 methods that were particular to a given design phase (e.g. the Research phase). Those methods were not shared with other phases, as tDX presents methods as being often phase-specific.

Second, no sequential prompt engineering or few-shot learning was undertaken: that is, zero-shot LLM prompting was used, with no follow-up questions. For example, asking "what kind of prototyping?" when recommended a generic prototyping method could have yielded greater specificity. However, to model novice designer behavior, for whom providing a 'correct example' of a method in their project context could be prohibitive, a zero-shot approach is informative. Subsequent work on crafting few-shot prompts commensurate with novices' experience, and approaches using retrieval-augmented generation to provide LLM's more context on a given query, could help better generalize these preliminary findings.

Third, LLM's are constantly evolving, presenting both opportunities and risks. Since the beginning of this work, GPT-4 has superseded GPT-3.5, and promises more effective performance – and may limit some of the findings in this work. It is unclear whether the results presented here will persist as LLMs continue to develop. An aforementioned risk with LLMs is bias that may emerge from its training, which will also develop. In our study, we observed the model produced method recommendations that were concentrated on certain methods, and not as diverse as human group. While a thorough bias analysis was out of scope of this work, future work could examine the role of LLM's bias in specific engineering design applications.

5.2. Implications for design research

This work suggests two follow-on research questions related to human-AI teaming in design. First, how can we blend the 'best' of design team- and LLM-driven method selection? Design teams could, for example, incorporate a nuanced understanding of their project into LLM

prompts, that could then more rapidly generate a range of possible methods. Second, what does the performance of an LLM, presumably with more design method data than a novice designer, say about the contribution of human design cognition in problem-solving? This work suggests that the contextual nuance that is readily available to designers is difficult for an LLM, despite its data sources, to interpret without explicit instruction. Understanding more the division between general design knowledge and design team decision-making in a project could contribute to a deeper understanding of design cognition.

5.3. Implications for design practice & education

Our findings suggest that while LLM's hold great promise to operate as a 'coach' or another advisor, they may require a substantial amount of prompt engineering, nuance and insight to deliver meaningful design outcomes. Practitioners and instructors seeking to leverage LLM's in design may consider operationalizing the types and phrasing of context delivered to an LLM to guide decision-making.

Similarly, the differences observed between human- and LLM-sourced design methods suggest that there is potential for human-AI teaming in combining the more general recommendations of an LLM with the more specific method selections and rationale of a human designer, supported by an instructor or design-specific prompts. Simultaneously, LLM method recommendations could lead human designers to over-rely on the more popular methods – something that practitioners and educators should be aware of. Future research could explore how to integrate greater topical design expertise into LLM operation, through few-shot learning, as discussed earlier, or through retrieval augmented generation (Lewis et al., 2020).

6. Conclusions

In this work, we examine how design teams' selections of HCD methods in a project contrast with methods recommended by OpenAI's GPT-3.5 LLM. Examining more than 400 team method selections and more than 2000 LLM-sourced method recommendations, we observe that human designers appear to readily bring project-specific nuance to their method selections, and appear to choose methods for a diversity of motivations that span the project outcome, collaboration, and design process. In contrast, the LLM appears to struggle to distinguish between specific design phases and project types, and appears to justify its recommendations primarily for design process purposes. We close with a discussion of implications for design research and practice.

5. References

Allison, J. T., Cardin, M.-A., McComb, C., Ren, Y., Selva, D., Tucker, C. S., Witherell, P., & Zhao, Y. F. (2021). Artificial Intelligence and Engineering Design. *Journal of Mechanical Design*, 1–6.

Avle, S., Lindtner, S., & Williams, K. (2017). How Methods Make Designers. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 472–483. https://doi.org/10.1145/3025453.3025864

Bazzano, A. N., Martin, J., Hicks, E., Faughnan, M., & Murphy, L. (2017). Human-centred design in global health: A scoping review of applications and contexts. *PLOS ONE*, *12*(11), e0186744. https://doi.org/10.1371/journal.pone.0186744

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). *Large language models in machine translation*.

Brown, T. (2008). Design thinking. *Harvard Business Review*, *86*(6).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Camburn, B., He, Y., Raviselvam, S., Luo, J., & Wood, K. (2020). Machine learning-based design concept evaluation. *Journal of Mechanical Design*, *142*(3), 031113.

Cash, P., Daalhuizen, J., & Hekkert, P. (2023). Evaluating the efficacy and effectiveness of design methods: A systematic review and assessment framework. *Design Studies*, *88*, 101204. https://doi.org/10.1016/j.destud.2023.101204

Ceschin, F., & Gaziulusoy, I. (2016). Evolution of design for sustainability: From product design to design for system innovations and transitions. *Design Studies*, *47*, 118–163. https://doi.org/10.1016/j.destud.2016.09.002

Chan, J., & Schunn, C. D. (2017). A Computational Linguistic Approach to Modelling the Dynamics of Design Processes. In *Analysing Design Thinking: Studies of Cross-Cultural Co-Creation*. CRC Press.

Chasanidou, D., Gasparini, A., & Lee, E. (2014). Design thinking methods and tools for innovation in multidisciplinary teams. *Workshop Innovation in HCI. Helsinki, Finland: NordiCHI*, *14*(2014), 27–30.

Chen, C., Mullis, J., & Morkos, B. (2021). A topic modeling approach to study design requirements. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *85383*, V03AT03A021.

Daalhuizen, J., & Cash, P. (2021). Method content theory: Towards a new understanding of methods in design. *Design Studies*, *75*, 101018.

Daalhuizen, J., Timmer, R., van der Welie, M., & Gardien, P. (2019). An architecture of design doing: A framework for capturing the ever-evolving practice of design to drive organizational learning. *International Journal of Design*, *13*(1), 37–52.

Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies*, *26*(5), 445–461.

Feuerriegel, S., & Pröllochs, N. (2021). Investor Reaction to Financial Disclosures across Topics: An Application of Latent Dirichlet Allocation*. *Decision Sciences*, *52*(3), 608–628. https://doi.org/10.1111/deci.12346

Fuge, M., & Agogino, A. (2014, August 17). User Research Methods for Development Engineering: A Study of Method Usage With IDEO's HCD Connect. *Proceedings of the ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Buffalo, NY. https://doi.org/10.1115/DETC2014-35321

Fuge, M., & Agogino, A. (2015). Pattern Analysis of IDEO's Human-Centered Design Methods in Developing Regions. *Journal of Mechanical Design*, *137*(7). https://doi.org/10.1115/1.4030047

Geis, C., Bierhals, R., Schuster, I., Badke-Schaub, P., & Birkhofer, H. (2008). Methods in practice–a study on requirements for development and transfer of design methods. *DS 48: Proceedings DESIGN 2008, the 10th International Design Conference, Dubrovnik, Croatia*, 369–376.

Gericke, K., Eckert, C., & Stacey, M. (2022). Elements of a design method – a basis for describing and evaluating design methods. *Design Science*, *8*, e29. https://doi.org/10.1017/dsj.2022.23

Gericke, K., Kramer, J., & Roschuni, C. (2016). An Exploratory Study of the Discovery and Selection of Design Methods in Practice. *Journal of Mechanical Design*, *138*(10). https://doi.org/10.1115/1.4034088

Gerrike, K., Eckert, C., & Stacey, M. (2017, August 21). What do we need to say about a design method? *Proceedings of the 21st International Conference on Engineering Design (ICED 2017)*. 21st International Conference on Engineering Design (ICED 2017), Vancouver, Canada. http://oro.open.ac.uk/50445/

Goel, T., Shaer, O., Delcourt, C., Gu, Q., & Cooper, A. (2023). Preparing Future Designers for Human-AI Collaboration in Persona Creation. *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, 1–14.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235.

Jagtap, S., Warell, A., Hiort, V., Motte, D., & Larsson, A. (2014). Design methods and factors influencing their uptake in product development companies: A review. *DS 77: Proceedings of the DESIGN 2014 13th International Design Conference*.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*, 15169–15211.

Jones, J. C. (1992). *Design methods*. John Wiley & Sons.

Joung, J., & Kim, H. M. (2021). Automated Keyword Filtering in Latent Dirichlet Allocation for Identifying Product Attributes From Online Reviews. *Journal of Mechanical Design*, *143*(084501). https://doi.org/10.1115/1.4048960

Kiatkawsin, K., Sutherland, I., & Kim, J.-Y. (2020). A Comparative Automated Text Analysis of Airbnb Reviews in Hong Kong and Singapore Using Latent Dirichlet Allocation. *Sustainability*, *12*(16), Article 16. https://doi.org/10.3390/su12166673

Kim, E., Beckman, S. L., & Agogino, A. (2018). Design roadmapping in an uncertain world: Implementing a customer-experience-focused strategy. *California Management Review*, *61*(1), 43–70.

Kim, E., Rao, V., Bluemink, B., Klitsie, B., & Santema, S. (2022, November 11). *Examining a Trajectory of Complex System Design Processes: Airport Eco-System Case Studies by Novice Student Teams*. ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. https://doi.org/10.1115/DETC2022-89901

Kolisko, S., & Anderson, C. J. (2023). Exploring Social Biases of Large Language Models in a College Artificial Intelligence Course. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(13), Article 13. https://doi.org/10.1609/aaai.v37i13.26879

Kramer, J., Agogino, A. M., & Roschuni, C. (2016, December 5). *Characterizing Competencies for Human-Centered Design*. ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. https://doi.org/10.1115/DETC2016-60085

Kramer, J., Poreh, D., & Agogino, A. (2017). Using TheDesignExchange as a knowledge platform for human-centered design-driven global development. *DS 87-1 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 1: Resource Sensitive Design, Design Research Applications and Case Studies, Vancouver, Canada, 21-25.08.2017*. https://www.designsociety.org/publication/39509/Using+TheDesignExchange+as+a+knowledge+platform+for+human-centered+design-driven+global+development

Kwon, E., Huang, F., & Goucher-Lambert, K. (2021). Multi-Modal Search for Inspirational Examples in Design. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *85420*, V006T06A020.

Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences*, *120*(49), e2309350120.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Lloyd, P., Chandrasegaran, S., Kim, E., Cagan, J., Yang, M., & Goucher-Lambert, K. (2022). Designing Dialogue: Human-AI Collaboration in Design Processes. *DRS Biennial Conference Series*. https://dl.designresearchsociety.org/drs-conference-papers/drs2022/editorials/28

Ma, K., Grandi, D., McComb, C., & Goucher-Lambert, K. (2023, November 21). *Conceptual Design Generation Using Large Language Models*. ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. https://doi.org/10.1115/DETC2023-116838

Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, *151*(2), 127–138.

Method Cards. (n.d.). Retrieved November 30, 2019, from https://www.ideo.com/post/method-cards

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. In R. Barzilay & M. Johnson (Eds.), *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Association for Computational Linguistics. https://aclanthology.org/D11-1024

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1–40.

OpenAI Platform. (n.d.). Retrieved March 15, 2024, from https://platform.openai.com

Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2022). *How to train your stochastic parrot: Large language models for political texts*. Working Paper.

Panchal, J. H., Fuge, M., Liu, Y., Missoum, S., & Tucker, C. (Eds.). (2019). Special Issue: Machine Learning for Engineering Design. *Journal of Mechanical Design*, *141*(110301). https://doi.org/10.1115/1.4044690

Picard, C., Edwards, K. M., Doris, A. C., Man, B., Giannone, G., Alam, M. F., & Ahmed, F. (2023). From Concept to Manufacturing: Evaluating Vision-Language Models for Engineering Design. *arXiv Preprint arXiv:2311.12668*.

Poreh, D., Kim, E., Vasudevan, V., & Agogino, A. (2018, August 26). Using "Why and How" to Tap Into Novice Designers' Method Selection Mindset. *Proceedings of the ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. https://doi.org/10.1115/DETC2018-85997

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raina, A., McComb, C., & Cagan, J. (2018, August 26). Design Strategy Transfer in Cognitively-Inspired Agents. *Proceedings of the ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. https://doi.org/10.1115/DETC2018-85599

Raina, A., McComb, C., & Cagan, J. (2019). Learning to Design From Humans: Imitating Human Designers Through Deep Learning. *Journal of Mechanical Design*, *141*(11). https://doi.org/10.1115/1.4044256

Rao, V., Kim, E., Kwon, J., Agogino, A. M., & Goucher-Lambert, K. (2021). Framing and Tracing Human-Centered Design Teams' Method Selection: An Examination of Decision-Making Strategies. *Journal of Mechanical Design*, *143*(3), 031403.

Rao, V., Kwon, E., & Goucher-Lambert, K. (2022). "Like a Moodboard, But More Interactive": The Role of Expertise in Designers' Mental Models and Speculations on an Intelligent Design Assistant. *International Conference On-Design Computing and Cognition*, 749–765.

Roschuni, C., Agogino, A. M., & Beckman, S. L. (2011). The DesignExchange: Supporting the Design Community of Practice. *DS 68-8: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 8: Design Education, Lyngby/Copenhagen, Denmark, 15.-19.08.2011*.

Roschuni, C., Kramer, J., Zhang, Q., Zakskorn, L., & Agogino, A. (2015, July 27). Design Talking: An Ontology of Design Methods to Support a Common Language of Design. *The Design Society - a Worldwide Community*. International Conference on Engineering Design ICED15, Milan, Italy.

Song, B., Gyory, J. T., Zhang, G., Soria Zurita, N. F., Stump, G., Martin, J., Miller, S., Balon, C., Yukish, M., McComb, C., & Cagan, J. (2022). Decoding the agility of artificial intelligence-assisted human design teams. *Design Studies*, *79*, 101094. https://doi.org/10.1016/j.destud.2022.101094

Song, S., Dong, A., & Agogino, A. (2003). Modeling Information Needs in Engineering Databases Using Tacit Knowledge. *Journal of Computing and Information Science in Engineering*, *2*(3), 199–207. https://doi.org/10.1115/1.1528921

Tomiyama, T., Gu, P., Jin, Y., Lutters, D., Kind, Ch., & Kimura, F. (2009). Design methodologies: Industrial and educational applications. *CIRP Annals*, *58*(2), 543–565. https://doi.org/10.1016/j.cirp.2009.09.003

Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W. S., Cheung, N.-M., Nguyen, H. L. T., Ho, C. S. H., & Ho, R. C. M. (2019). Characterizing Artificial Intelligence Applications in

Cancer Research: A Latent Dirichlet Allocation Analysis. *JMIR Medical Informatics*, *7*(4), e14401. https://doi.org/10.2196/14401

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Verganti, R. (2009). *Design driven innovation: Changing the rules of competition by radically innovating what things mean*. Harvard Business Press.

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). Kelly is a Warm Person, Joseph is a Role Model: Gender Biases in LLM-Generated Reference Letters. *EMNLP-Findings*.

Yang, Q. (2018). Machine Learning as a UX Design Material: How Can We Imagine Beyond Automation, Recommenders, and Reminders? *2018 AAAI Spring Symposium Series.*, 6.

Yuan, B., Ye, J., Wu, X., & Yang, C. (2023). Applying Latent Dirichlet Allocation and Support Vector Regression to the Aesthetic Design of Medical Nursing Beds. *Journal of Computing and Information Science in Engineering*, *23*(051014). https://doi.org/10.1115/1.4062350