What About My Design Context?: Exploring the Use of Generative AI to Support Customization of Translational Research Artifacts

Donghoon Shin
dhoon@uw.edu
University of Washington
Seattle, WA, USA

Tze-Yu Chen
alextyc@uw.edu
University of Washington
Seattle, WA, USA

Gary Hsieh
garyhs@uw.edu
University of Washington
Seattle, WA, USA

Lucy Lu Wang
lucylw@uw.edu
University of Washington,
Allen Institute for AI
Seattle, WA, USA

ABSTRACT

Despite the wealth of knowledge in research papers, practitioners struggle to apply research results to their work due to significant research-practice gaps. This study addresses the rigor-relevance paradox, where academic rigor can undermine the practical relevance of research for designers. Specifically, we explore the potential of large language models (LLMs) to customize translational research artifacts (*i.e.*, design cards) and improve relevance to specific designers' needs. In our preliminary study ($N = 15$), designers defined relevance as alignment between the content of the translational artifact and their design context—including target users, modalities/domains, and design stages. Based on these findings, we implemented an LLM-powered pipeline that allows designers to customize research papers into design cards tailored to their contexts. Our evaluation ($N = 20$) demonstrated that designers perceived customized artifacts as more relevant, actionable, valid, generative, and inspiring than those without customization—even for less topically related papers—indicating LLM-powered customization can be used to support research translation.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; **User centered design**; **Systems and tools for interaction design**.

KEYWORDS

translational science, rigor-relevance paradox, customization, generative AI

DIS '25, July 5–9, 2025, Funchal, Madeira

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

While many valuable insights are uncovered through scholarly research and presented in research papers, these papers are rarely read by practitioners [14]. This phenomenon, known as the research-practice gap [6, 52], has become a growing topic of discussion within HCI [8, 14, 15, 55], particularly regarding its implications for design practitioners. One noted reason why research papers are not effective in supporting design is the *rigor-relevance* paradox [58]. On the one hand, researchers focus on creating generalizable knowledge, which is centered on ensuring that their research is *rigorous* to meet the expectations of peer review. On the other hand, the practitioners seeking to apply the generated insights need this knowledge to be *relevant* to their specific use cases. For designers, this may result in limited consumption of published research due to a lack of perceived value. As one design practitioner quipped in Colusso *et al.*'s study on their use of research papers: *"Academic research goes so deep that it no longer is applicable for us (designers). Everything is pure theory and the real world doesn't work that way."* [14]

Recent research has shown that large language models (LLMs) can be used to translate research findings into translational communication artifacts (*e.g.*, design cards) for design practitioners [69]. While this work showcases LLMs' ability to scale the traditionally labor-intensive process of manual translation, a critical limitation remains: such AI-generated translational research artifacts were not perceived as significantly more actionable than simply reading the original papers. This limitation poses a significant barrier to the practical adoption of these artifacts for practitioners.

In this work, we hypothesize that LLMs are capable of addressing this limitation by making customized recommendations and connections to real-world design; in other words, by improving the *relevance* of the translational artifact to the needs of individual design practitioners and their specific design projects. To facilitate the customizing of translational communication artifacts to individual designers' needs, however, two questions must be answered. First, what does 'relevance' mean in the context of customizing translational artifacts? Second, how do we tailor the translational artifacts to improve relevance to the designer's specific design goals?

To address these questions, we conducted a preliminary interview study involving 15 designers to understand their perceptions of relevance when consuming design insights from research papers. Participants highlighted the importance of aligning research knowledge with the *who* (the target audience for whom they are designing) and *what* (the goal and design space of their design process; *e.g.*, modalities, pain points to address, client, metrics to enhance) of their design context. In addition, we surfaced several design implications for customizing the translational artifacts based on the designer's *design stage* (*e.g.*, research, ideation, and/or evaluation). Guided by these findings, we designed and implemented an LLM-powered pipeline that takes these individual designer needs as input, and automatically customizes insights from papers tailored to those specific needs. Presented in a design card format, research insights are displayed across a set of goal-specific components we designed based on these preliminary study findings.

To evaluate the efficacy of our customization pipeline, we conducted a user study with 20 designers. Participants were asked to reflect on one of their recent design projects, create customized design cards for this project with our pipeline, and compare customized design cards to those without customization. Participants found the insights communicated by the customized design cards to be significantly more relevant and actionable than cards without customization, while also being more generative, inspiring, and valid. Because LLMs can introduce hallucinations, we also conducted a manual evaluation of the customized cards generated in our user study to identify mismatches with the source paper and user inputs. Our evaluation found that the large majority of generated components accurately represented both the paper content and the user's provided design context without issue. However, we identified several types of mismatches, such as misinterpreting the paper's focus or offering overly generic solutions, and discuss these as potential areas for future improvement.

Our main contributions are as follows:

- We conducted a preliminary study with designers ($N = 15$) to understand how they define 'relevance' for translational research artifacts. Results revealed the importance of alignment with the designer's specific design context (*i.e.*, who/what they are designing for, and design stage) (§3);
- Based on these findings, we developed an approach that generates translational research artifacts tailored to the individual needs of designers with LLMs. We contributed a component library for delivering customized design insights, a pipeline for generating these components, and mechanisms for extraction and attribution (§4);
- We evaluated our customization pipeline with designers ($N = 20$), revealing the efficacy of the customized translational artifacts generated by our system and potential future enhancements. We also conducted an intrinsic evaluation of customized contents to assess the accuracy of AI-generated output; while error rates are low, we identified several classes of mismatches that should be considered in future model and pipeline development (§5, §6).

2 RELATED WORK

2.1 Rigor-Relevance Gap in Translational Science

Much of the prior literature on the research-practice gap has presented the contrast between the rigor of research with practitioners' perceived need for relevance as a main contributing factor (*i.e.*, rigor-relevance gap). As researchers and practitioners belong to different communities with different core activities [71], they have different "performing" aims in which they try to accomplish with their work. Researchers are tasked and incentivized to do "good" research, where they have to "define research problems precisely and investigate them carefully and deeply" [6]. Whereas, in practice, rigor may not be an important consideration, and can "be burdensome when speed, agility, and even creativity are paramount" [6]. This attribute to different languages, styles, and logics (*i.e.*, differences in defining and tackling problems) that exist between research and practice [38].

There is much ongoing discussion on whether this rigor-relevance gap needs to be mutually exclusive and whether the gap can be bridged. Some scholars have argued that this dichotomy is artificial, and that research should be both [29]. To facilitate that, one set of solutions focuses on bridging the gap during knowledge generation – exploring closer collaborations between scholars and practitioners so that the problems examined by researchers are relevant to practice [29]. However, some have criticized this approach as research and practice inherently serve different functions; having science that maximizes both rigor and relevance is unrealistic and naive, and may jeopardize the needs and functions of both communities [16].

In our work, we focus on bridging the gap during dissemination. Prior work, including research in HCI on research-to-design practice gaps, has suggested that research papers are too academic, inaccessible, and boring [14, 43, 64]. By focusing on dissemination, we may be able to preserve the functions of research so that scholars can still do the rigorous work that is needed within their academic communities. What we seek to facilitate, is on improving the boundary object between the research and practice communities. Instead of relying on research papers as is the current, which is ineffective, we posit that we may be able to develop more relevant (*i.e.*, customized) translational artifacts that can better satisfy the needs of practitioners.

2.2 Generative AI for Customizing Translational Communication

One of the inherent challenges with supporting research translation is who should be performing that work. As argued by the prior work, expecting researchers to do this and to do this effectively is unrealistic [69]—if they are sufficiently motivated and skilled to do so, they would be doing this already. Having translational workers or diffusion fellows [81] is a possibility, but this solution is costly, will not scale well, and may perpetuate inequalities due to unequal access to resources. Approaches leveraging expert crowdsourcing may address the problem of scalability, but do not address issues of motivation (*e.g.*, still relying on volunteerism) and coverage would likely be limited. Instead, the rise of LLMs offers a unique solution to this translational problem. For example, prior work demonstrated an LLM-powered pipeline that takes research papers as inputs and creates design cards on demand to communicate the embedded design implications [69].

While it presents an approach to scaling translational communication through automating the creation of design cards, which have a rich history in HCI for design communication, a critical yet challenging limitation persists: the approach does not account for the distinct design contexts that each individual designer has. Given the unlikelihood that a research paper has the exact same design context as each designer's, further customization to individual designer's context must be followed to meet their need for relevance. Indeed, the authors acknowledged that their approach of simply translating research papers into translational artifacts fails to augment the actionability of the communicated insights, and similarly, prior work has criticized such an indistinct framing as a limitation in current efforts to bridge the gap between research and practice [21]. In this work, we argue that, by increasing the relevance to each designer's context, we may further enhance the perceived value of these artifacts [21].

Our study also expands upon prior work adapting scientific content for non-target audiences such as interdisciplinary researchers, practitioners, and the public. Knowledge and disciplinary barriers such as jargon and terminology mismatch have been identified as critical challenges to translation across disciplines [12, 30, 47, 50]. Unfamiliar terminology that acts as a barrier to understanding can be predicted with some success using a person's social media posts [45]—or in the case of researchers, their prior publications [30]—as the contextual representation into a model. While it is possible to integrate a designer's context automatically to conduct customization, which we discuss in future directions, we focus in this initial study on identifying the customization inputs most likely to improve the relevance of our output translational artifacts.

Prior work has also explored interactive systems that simplify, augment, or otherwise allow different access to content in research papers to address these barriers [5, 20]. Chat-based paper question-answering capabilities have also been introduced in production systems like Elicit[1] and Semantic Scholar,[2] but while promising, these features are limited in being too general purpose (*i.e.*, do not aptly address the needs of a specific user group like design practitioners) and are not customized to individual needs (*i.e.*, do not address the core translation science communication problem of context awareness and relevance). In this work, we define the components of relevance identified as most important by design practitioners and implement an interactive customization pipeline targeted toward these components.

3 PRELIMINARY STUDY

We conducted an interview study with designers ($N = 15$) to understand (i) their definition of *relevance* when consuming research and (ii) how LLMs can tailor the presentation of translational research artifacts to improve relevance. To facilitate these discussions, we developed a simple LLM prompt that creates customized translational artifacts based on a user's brief description of their design context, and used the method and outputs as study probes.

3.1 Recruitment & Participants

We posted our study recruitment in three university groups and one online community focused on design. In this posting, we required the participant to be (i) currently working as a professional designer or (ii) pursuing a professional design degree. As a result, we recruited 15 designers; the participants had an average age of 26.5 years ($SD = 2.8$); 8 self-identified as female, 6 male, and 1 non-binary. Among them, 2 had less than 1 year of experience in design or studying a design-related major, 10 had 1 to 5 years of experience, and 3 had more than 5 years of experience.

3.2 Study Procedure & Analysis

To probe what designers want to see from the customized translational artifacts, we started by developing a customization interface (see Appendix A) that augments the AI-generated design card pipeline and format suggested by Shin *et al.* [69]. Their approach synthesizes the design implication discussed in HCI papers into a two-page card format, which includes a paper overview, a description of the design implication, supporting evidence, and citations.

To allow for the tailoring of these artifacts to the designers' context, we added a free-form input field to the interface that asks the user: "How do you want to customize these cards for your own design work?" Based on their response, the system generates a third card page with simple customized design knowledge using an LLM. This page combines the original design card content with the user's input, presenting a summary of the customization request and two tailored design guidelines meant to offer more customized insights. This initial pipeline is enabled by the following prompt with an LLM (Azure OpenAI GPT-4o [57]), which returns formatted JSON used to populate the third card page:

```
Based on the following design implication of the research
paper, generate two customized design insights tailored
to the designer's design work:
 - Design guideline: {{ design implication text used to
   generate design card }}
 - Designer's design work: {{ free-form description
   provided by the user }}
```

With this customization pipeline, we conducted study sessions remotely on Zoom. After consenting each participant to the study, we briefly introduced the goal of our study. We also asked several questions spanning their design background, prior experience in consuming/using knowledge from research materials in their design projects, and the way of determining relevance in consuming research knowledge. Each participant was then asked to describe details of one of their recent design projects, which lasted approximately 10 minutes.

Following, they were provided with a list of ten papers that included design implications/guidelines in the paper text. To curate the list, we first searched for papers using keywords indicating the presence of design implications (*e.g.*, "design implication," "design guideline"), and filtered for full papers published in ACM-sponsored conferences and sorted them by relevance. The two authors then collectively reviewed each paper manually to confirm the inclusion of explicit design implications or guidelines. To ensure topical diversity and minimize redundancy, we excluded papers with overlapping design focuses. As such, the final list of ten papers was selected across diverse domains—including accessibility, human-robot interaction, and collaborative authoring [2, 7, 22, 32, 51, 53, 59, 60, 75, 79]—with the number limited to ten to keep the selection manageable for participants.

[1] https://elicit.com
[2] https://webflow.semanticscholar.org/api-gallery/s2qa

Each participant was asked to describe a recent design project of their own, after which they were presented with the curated paper list (including titles and themes). They then selected the paper most relevant to their project and briefly explained their choice. Following, we moved on to the demo session, where participants were given a link to an online web interface. On this interface, they first encountered (i) the raw text of their chosen paper (in PDF format) and (ii) one of the design implications of the paper, which they were encouraged to read at their own pace to become familiar with the content. Once they had reviewed these materials, they were directed to the card page (see Figure 4), where they could view the initial design cards generated from the selected design implication of the paper. On this page, participants could input a description of their design work into the provided free-form input field to generate a customized card page. They were free to iterate on this process as many times as they wished. The demo session lasted approximately 15 minutes for each participant.

Lastly, each participant proceeded to the semi-structured interview, where they provided feedback on the content and format of the customized design cards, as well as the input and interaction for generating them. Additionally, they were asked to provide future enhancements along with rationales. As a result, each study lasted approximately one hour.

Participants were compensated 30 USD for their participation. All study procedures were reviewed and approved by the IRB of our university's human subjects division. To analyze the data, we began by transcribing the voice recordings. Next, we conducted a thematic analysis [10] with a bottom-up approach, where the researchers initially identified and categorized key themes from the responses. These preliminary themes were then discussed and refined collaboratively over four rounds, after which we reached a consensus on a final set of themes outlined in Section 3.3. We refer to participants in the preliminary study as P1–P15.

3.3 Preliminary Study Results

Of all 15 participants, 11 participants have previously used or attempted to use research papers in their design work, and the other 4 participants mentioned that they have read books or design articles as part of their design process. Reasons for engaging with research papers varied widely, ranging from class projects to corporate initiatives. However, participants encountered several challenges when trying to apply that knowledge directly to their design processes, confirming the challenges noted by prior work [14].

First, the complexity and time required to read research papers made it difficult for designers to engage effectively. In previous experiences, some participants found the language of research papers dense and the content challenging to understand, often necessitating frequent revisits to different paper sections and leading to significant time investment: *"I would say, it takes a long time. It takes like, at least an hour for one paper (...) the papers itself are very long. So it's very time-consuming for me."* (P1) Additionally, participants reported difficulties regarding the relevance and applicability of papers to their work. Participants found it challenging to align research closely with the context of their specific design projects, where papers may not address the precise needs of their enterprise design contexts: *"I think, when it comes to reading things, sometimes you can find different use cases, use cases like e-commerce, or sometimes software and complex web app, that are not perfectly aligned."* (P6)

In response, some participants mentioned that they collaborated with their design team to adapt the ideas from papers and synthesize them to better fit their own design contexts. However, some perceived the high-level insights gained from such literature reviews not to be highly useful: *"We share articles in Slack, and give a brief synopsis. (...) Often, to be honest, not even a sentence of the article relates (to the project). Sometimes it just be like, we should check this, sort of like that. Really abstract, and not very helpful sometimes."* (P10)

3.3.1 What does relevance mean in the context of consuming research papers for design work? From the interview and the preliminary exploration using our demo interface, participants generally described a desire to learn about the similarities and differences between the paper and their work. They were particularly interested in understanding how findings from the paper could be transferred to their specific projects, exemplifying the need for direct connections and comparisons between the tested designs in the paper and their own design challenges: *"I might be more interested (...) 'how does that pattern (from the paper) apply or not to my project?' And like, (...) looking at, like potential design crossover."* (P9) Based on study findings, we describe three key dimensions for making design knowledge from papers more relevant:

(i) 'Who' they are designing for — Participants mentioned that identifying the target audience for whom they are designing is critical for determining the relevance of design insights from research papers, as insights are most valuable when they align closely with the characteristics and needs of the intended users: *"Who am I designing for is probably the biggest one that you got (when evaluating relevance)."* (P15) Participants also offered suggestions on how their target audience could be provided and used to tailor research knowledge to their intended audience, such as by providing demographic information or personas representing their target users: *"Audience or, like, if I do have established personas that would be helpful"* (P9); *"A question I might want to ask (from the paper) is (...) okay, how can I design for this certain demographic?"* (P3)

(ii) 'What' they are designing for — In addition, participants mentioned that the goal and design space of their design process are crucial for determining the relevance of research papers. During the interview, participants reported several sub-dimensions that this design space encompasses, such as modality/domain, pain points, client/company, and metrics to enhance:

- *Modality and domain*: Participants reported that they often consider the modality or domain that the study focused on when evaluating the relevance of research. They mentioned that knowing if a paper pertains to the specific type of modality or domain relevant to their work helps in determining its applicability: *"I would definitely give like a one or two-liner description of the project, or the platform that I'm currently building"* (P8); *"(...) what is working over there (research paper) that we could potentially use in our work that would enhance our designs as a [P6's design domain] application."* (P6)

- *Pain point*: Also, they found research addressing specific pain points experienced by their target users or clients to be relevant. Understanding these pain points is reported to be crucial for identifying research that offers solutions to their particular challenges: *"The client was the {name of the client} where we talked to them and learned about their (target users) pain points."* (P4)
- *Client/company*: While design guidelines are valuable, participants reported that they often need to adjust these guidelines to fit their company's specific context, such as adapting the guidelines to the unique needs of their company or project: *"I think design guidelines can be helpful, but you'll need to relate, or twist a little bit and then relate your company's departments or your own designs."* (P7)
- *Metric*: Some participants mentioned that they would like to focus on specific metrics, such as key performance indicators (KPIs) or engagement levels, to determine the relevance of research:*"If you (designers) have a specific goal, (I want to provide) like I'm trying to maximize this certain KPI or I'm trying to increase engagement in this way. Having that would be the kind of the north star."* (P14)

(iii) 'Where' are they in the design process (i.e., design stage) — In addition to tailoring to the *who* and *what* of design, participants also reported that the insights they want, and how they may use the insights from the research paper change depending on which *design stage* they are in. Here, we describe some of the key stage-dependent needs reported by our participants:

- *Synthesizing for better understanding (research stage)*: During the research phase, participants mentioned that they would try to use scholarly papers to gain a comprehensive understanding of their design target, and help synthesize and compare information from various sources. This can lead to moments of tension, as exemplified by P15's experience of reconciling conflicting information from different sources, where designers must navigate these discrepancies to form a cohesive understanding that informs their design decisions: *"In the research phase, I read this paper, and this article that we chose told me this, but like in my own primary research, I found this thing. So how do I reconcile those things?"* (P15)
- *Suggesting specific design ideas (ideation stage)*: Participants also expressed a need for actionable and specific design ideas to guide their ideation process. At this stage, they sought examples derived from papers, tailored to their unique design contexts, to follow and apply during their design work: *"I'd have something that says expand upon it (high-level insight). Maybe that is like giving concrete examples of how I can use in addition (...) you might also consider it."* (P11)
- *Defining methods & metrics (evaluation stage)*: Another need identified during the interview was support in defining relevant methods and metrics for use in their design processes. Participants expressed a desire to understand the methodologies employed by the paper authors and the metrics used to assess the effectiveness of the design, which could potentially be adapted and applied to their own work: *"I will (see) their methods (...) and based on these, I'll kind of think about what it implies in my design."* (P7); *"(when reading papers) I'd look at the methods that they used, and how they conducted or designed a specific experiment. So from there, I can get some insights on how I should better conduct my experiments, my interviews, or my studies, based on the paper."* (P1)

4 DESIGN ITERATION & IMPLEMENTATION

Our preliminary study identified several dimensions for improving the relevance of translational communication artifacts. Building on these insights, we developed an LLM-powered pipeline (Figure 1) that generates customized design cards—consisting of multiple components (Figure 2) customized based on a designer's description of their current design goals and design stage. Below, we outline the pipeline's structure, design, and implementation. Examples of customized design cards can be found in Appendix C.

4.1 Overall Interaction & Structure of the Generated Cards

We iterated on the concept of AI-generated design cards [69] (see Figure 1-(a)) to build a pipeline that takes details of the designer's work as input to tailor these design cards. Rather than a single open-ended input field, we query the designer for specific dimensions of their design context. Based on this, our pipeline generates two additional card pages that present customized insights.

The first card page presents stage-independent components identified in our preliminary study, which we call *Inspiration and scope*, along with a summary of the inputs from the designer. In contrast, the second card page features components tailored to each stage of the design process (*e.g.*, *Understanding users*, *Design ideas*, *Methods for you*, *Metrics for you*, and *Figure*), allowing designers to adaptively consume the design knowledge (see Table 1). In Section 4.2, we provide a detailed discussion on the informational structure of components and their implementation, along with an example of the prompt used to generate components in Appendix B.

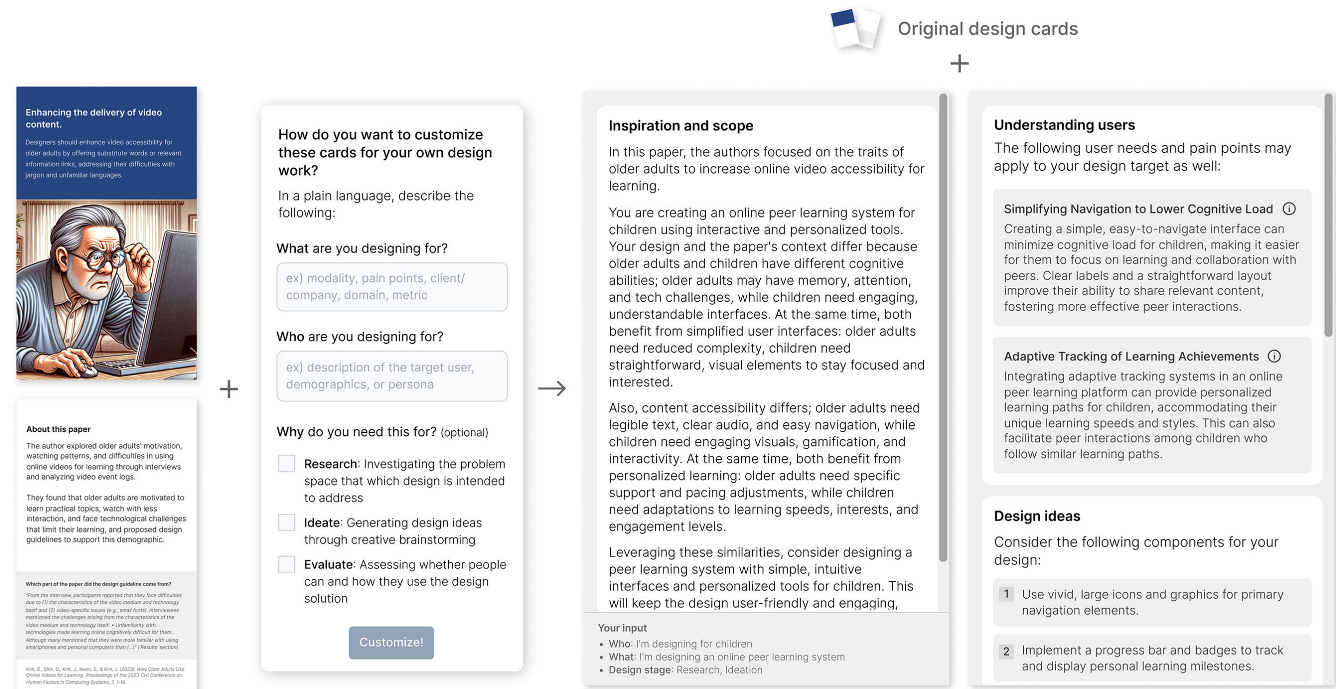
4.2 Components

4.2.1 Input. Our preliminary study revealed that an important aspect of relevance is to consider *who* and *what* designers are designing for. Thus, we ask designers to describe these two dimensions of their work as the starting point for customization. To assist them in articulating their work, we included several examples identified from our preliminary study as placeholder text for each input.

We also request the *design stage* as an optional input, which determines which components are included on the design cards. Of the five design stages supported by translational artifacts described in Hsieh *et al.* [33], our preliminary study identified distinct designer needs across three design stages—research, ideation, and evaluation. As a result, we focus on addressing the needs of these three stages, letting designers choose the stage(s) that suit their needs among these options (see Figure 1-(b)).

4.2.2 Stage-independent components. These components appear on all customized design cards, regardless of the design stage:

Inspiration and scope. The *Inspiration and scope* component provides an overview of how design insights from the paper are relevant to the designer's context (see Figure 2-(a)). The section begins with a brief one-line summary of the user's input. It then presents up



(a) Original design cards (b) Scaffolded description of their design work (c) Customized card pages

Figure 1: Pipeline for generating our customized cards. After being presented with (a) original design cards [69], each participant can describe their design work in a (b) scaffolded text entry consisting of what/who they are designing for and a checkbox for selecting design stage(s). Then, the system generates (c) customized card pages consisting of components corresponding to their design stage(s).

to two paragraphs highlighting the differences and similarities between the designer's design context and the paper, each focusing on a specific aspect of how the paper's context contrasts/compares to the designer's. Finally, to address designers' needs for using papers as a guiding resource, the section ends with a high-level design direction that the designer may consider, drawing on the similarities identified in the previous paragraphs.

To support the designer's needs to translate knowledge from the paper to their work, we utilized the concept of analogical idea representation as proposed by Yu *et al.* [91], by eliciting the underlying relationship between constructs. In essence, we composed an analogy between the paper's design insight and the designer's context. We found this approach to be beneficial, as it allows for some knowledge translation even when the paper's domain differs from that of the designer. Additionally, prior works have found efficacy and capabilities of LLMs in extracting analogies [85], including creative analogy mining (*e.g.* [9]), which highlights their potential to support creative reasoning and cross-domain knowledge transfer—an essential mechanism in our approach to tailoring insights.

Specifically, we first prompted the model to identify the *relation* between the paper's design target and its design guideline. Then, using this relation along with the designer's design target, we prompted the model to generate a corresponding design inspiration tailored to the designer's design target. This method enabled us to apply an analogy from the paper, even if the focus domains differ.

Methods used in this study. To support designers' overall needs for understanding the paper, our pipeline adds this section to the original design card page. This section extracts and showcases up to two methods used in the paper (see Figure 2-(g)).

4.2.3 Stage-dependent components. These components appear on the final customized card if the relevant design stage (in parentheses) is selected by the user:

Understanding users (research stage). The *Understanding users* component is designed to help designers reconcile their design goals with the design targets presented in the paper (see Figure 2-(b)). It highlights up to two user needs/pain points emphasized by the paper's authors, which can serve as motivation and potentially apply to the designer's specific design context.

To generate this section, the pipeline uses (i) the designer's input (who, what) and (ii) the original paper text. The model is prompted to return up to two user needs or pain points highlighted by the paper authors that may also help the designer understand their target users. Each need/pain point consists of: (i) a concise title (*i.e.*, keywords referring to the need/pain point), (ii) a description (*i.e.*, relevance of this need/pain point to the designer's target user), (iii) a source detail (*i.e.*, the definition as provided by the authors, if any, and why the paper's authors emphasized this need/pain point),

Table 1: Summary of how additional card components for the customized design card are constructed. In this case, we use GPT-4o as the LLM, GROBID [46] as the paper segmentation model, and PDFFigures 2.0 [13] as the figure localization model. For *Methods for you* and *Figure*, contents are tailored to the design stage(s) provided by the designer.

Design stage	Card component	Inputs used to construct the component	Where it appears	Model(s) used
All	Inspiration and scope	Designer's input (who, what); Original paper text	Customized card (first page)	LLM; Paper segmentation model
	Methods used in this study	Original paper text	Default card (second page)	LLM; Paper segmentation model
All (contents tailored to each stage)	Methods for you	Designer's input (who, what, design stage); Original paper text	Customized card (second page)	LLM; Paper segmentation model
	Figure	Designer's input (who, what, design stage); Localized figure data (from original paper text)	Customized card (second page)	LLM; Figure localization model
Research	Understanding users	Designer's input (who, what); Original paper text	Customized card (second page)	LLM; Paper segmentation model
Ideation	Design idea	Designer's input (who, what); *Inspiration and scope* component	Customized card (second page)	LLM
Evaluation	Metrics for you	Designer's input (who, what); Original paper text	Customized card (second page)	LLM; Paper segmentation model

(iv) the source paragraph (*i.e.*, where the user need/pain point is discussed in the original paper), and (v) its section title. Items (iii)-(v) are presented as a tooltip (see Figure 2-(f)) that appears on hover to provide a quick reference to the underlying rationale, as requested by participants in our preliminary study.

Design ideas (ideation stage). The *Design ideas* component (see Figure 2-(c)) is generated for the ideation stage, and shows specific design ideas that users could follow. To align these design ideas with the *Inspiration and scope* section, we prompted with both (i) the high-level insight presented on the *Inspiration and scope* component and (ii) the designer's input (who, what). We instructed the model to generate up to three specific one-line design recommendations that the designer may consider following, which builds on the high-level insight.

Methods for you (available for all stages but tailored to the chosen stage(s)). To assist designers in gaining insights into applicable methods for their work, the *Methods for you* component (see Figure 2-(d)) provides up to two methods that may be useful for the designer's design stage(s). While this component was initially designed to support the evaluation stage, we recognized that research papers often have methods across diverse design stages, and thus, other design stages could benefit from stage-specific methods as well. Therefore, we included tailored methods for each stage of the design process.

We used (i) the designer's input (who, what, design stage) and (ii) the original paper text to generate this section. The model is instructed to return up to two methods from the paper that may be used for the designer's design stage(s), each containing (i) a concise title (*i.e.*, the name of the methodology), and (ii) a description (*i.e.*, how the methodology may assist in achieving the design goals, along with considerations the designer should keep in mind when applying it to their target user, goal, or design stage). Additionally, similar to the *Understanding users* component, the model is instructed to return (iii) a source detail (*i.e.*, the definition of the method as provided by the authors, if any, and why the paper's authors used this method), (iv) the source paragraph, and (v) its section title, which are used to construct tooltips offering additional information.

This way, we aimed to provide designers with methods relevant to their stage, as well as the source from which the methods originate. If they did not provide any design stage, the pipeline presents methods that may be applied to any of the three design stages.

Metrics for you (evaluation stage). This component (see Figure 2-(e)) is presented when the user selects the evaluation stage. It offers up to two metrics tailored to their design context. Similar to *Methods for you*, the model is provided with (i) the designer's input (who, what) and (ii) the original paper text. It is instructed to return up to two metrics, each including (i) a concise title, (ii) a description, (iii) a source detail, (iv) the source paragraph, and (v) its section title.

4.2.4 Other components.

Figure from the paper (available for all stages but tailored to the chosen stage(s)). To enhance visual understanding of the paper's content, we extracted and included a figure that is most relevant to the designer's current stage of the design process (see Figure 2-(h)). First, we employed a figure localization model [13] to extract all figures and their corresponding captions from the paper. Then, by providing the LLM with the designer's design stage(s) and the figure captions, we prompted it to return a figure that would likely inspire the designer given the provided design stage(s) along with the rationale behind its selection.

It is then used to populate the figure at the bottom of the second customized card page, with the rationale serving as a caption. If the designer did not provide any design stage, the pipeline presents a figure that may inspire the designer in any of the three design stages. If no figure is available in the paper, or none is relevant to any of the designer-provided stages, this component is omitted.

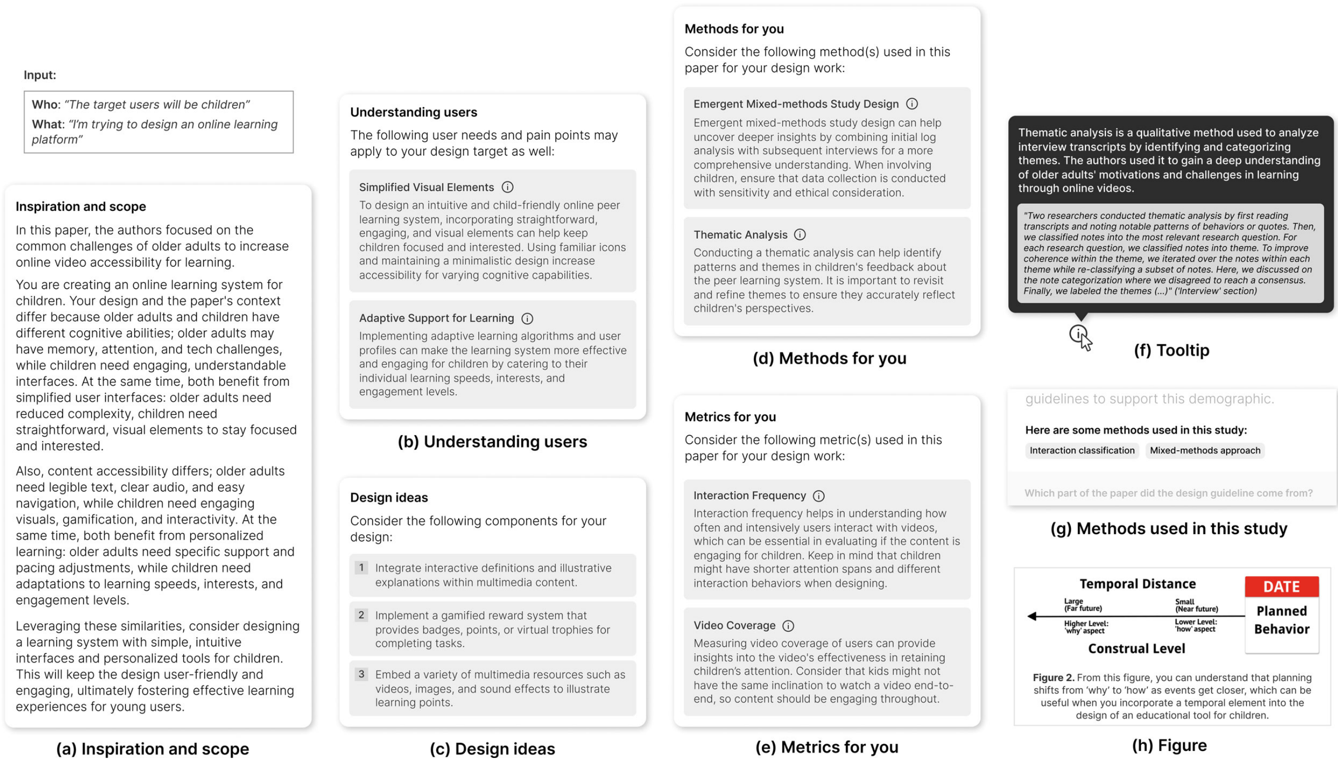


Figure 2: Components of our customized cards and the input used to generate the components. Based on the user input regarding the design stage, our system adaptively generates and assembles the components corresponding to the user's stage(s), creating tailored design cards.

4.3 Implementation of the Customization Pipeline

We implemented our pipeline as a web interface using SvelteKit [78], a JavaScript-based web app framework, which interfaces with a Python backend server hosted on an AWS EC2. The server performs several key functions: (i) executing LLM computations, (ii) segmenting paper PDFs using a paper segmentation model, (iii) using a figure localization model to identify and extract figures, and (iv) returning the generated design card to the user interface. We employed Azure OpenAI GPT-4o [57] as a model for natural language processing tasks, GROBID [46] for paper segmentation, and PDFFigures 2.0 [13] for figure localization. Since these components are modular, other alternative models can easily replace these models to provide similar capabilities.

5 EVALUATION METHODS

We conducted an interview study with designers ($N = 20$) to evaluate our customization pipeline. Participants were asked to identify a recent design project, and use its design scenario to generate several customized design cards using our system. Participants compared our customization results against design cards without customization to assess potential improvements to relevance and other qualities. Additionally, the research team reviewed all customized artifacts generated as a part of the evaluation study to assess the accuracy of AI-generated content.

5.1 Recruitment

To recruit designers, we posted our recruitment posting on three design-focused university groups and one online community. As a result, we recruited 20 participants, and they had an average age of 26.8 years ($SD = 3.9$). Of all, 14 of them self-identified as female, and 6 as male. On average, they had 3.7 years ($SD = 2.3$) of experience in design work or studies toward a professional design degree.

5.2 Study Setup

Our goal is to understand if the customized design cards generated from our pipeline successfully communicate relevant design insights from research papers with designers. To this end, we ran a within-subjects study that compares customized cards generated by our pipeline with those without customization (*i.e.*, default cards), using a mixed-methods approach of both a controlled quantitative survey and interviews to evaluate communication quality and gather qualitative feedback.

More specifically, participants were asked to customize the cards, and assess both customized and default cards. In this process, to better reflect real-world use of our pipeline during the evaluation, participants were asked to generate, review, and evaluate both card contents based on a recent design project they described during the survey and interview. This enabled a controlled evaluation of the communicative qualities of customized cards compared to those without customization, while ensuring that the feedback reflected

Table 2: Designer-provided queries we used to identify papers that are relevant to each designer's recent project and the papers we selected to generate customized cards for evaluation

ID	Designer-provided query	Papers	ID	Designer-provided query	Papers
P1	Augmenting campaign engagement	[54, 76]	P11	AR experiences using gyroscope and accelerometer	[92, 93]
P2	Usability of generative AI	[80, 87]	P12	Value-sensitive design practices in UX	[65, 94]
P3	Human-AI interaction in the cybersecurity space	[11, 74]	P13	Emerging entrepreneurship in the AI space	[24, 90]
P4	Design VR games for children	[27, 35]	P14	LLM-based tools, emotion-aware AI	[4, 67]
P5	AI for navigating, elders	[36, 84]	P15	Interface design of AI systems	[56, 62]
P6	STEM Education and VR for high school students	[49, 63]	P16	AI-generated information overload	[39, 42]
P7	Gesture-based interactions for mobile interfaces	[37, 77]	P17	VR games, kids, cognition	[48, 72]
P8	Audio navigation patterns, sensory aided navigation	[1, 34]	P18	Creating a social media platform for investors	[41, 83]
P9	AR education application for college students using AI	[17, 26]	P19	Geo-games	[18, 89]
P10	ADHD management strategies	[70, 73]	P20	UX in Blockchain	[23, 82]

their real-world design work. In addition, to assess the reliability of generative AI in customization, we ran an intrinsic evaluation of the contents of the customized cards to evaluate the accuracy of the content in the customized card components after the study.

To create translational research artifacts for our study, we chose to present each participant with 2 types of papers: (i) ones that are topically related to each designer's specific context and (ii) ones that are not directly related. This study design allowed us to explore the potential generalizability of our system in supporting designers to engage with papers that may not be obviously related to their design project.

For (i), participants signing up for our study were asked to provide a query that they might use when conducting a literature search to support one of their recent projects. Then, prior to starting our study, we searched using this query in the ACM Digital Library, sorting results by search relevance, and identified two papers including sections explicitly signaled by the authors as design implications (*i.e.*, sections that follow a passage containing keywords indicating design implications, or have such keywords in their heading, *e.g.*, implication, guideline, or recommendation). Designer-provided queries and the papers are given in Table 2.

For (ii), we selected two HCI papers that are more abstract and theory-oriented, specifically He *et al.* [31] and Suh *et al.* [75]—the former focuses on the application of the transtheoretical model in designing, and the latter focuses on the use of the theory of planned behavior. Both highlight cited social and behavioral science theories that are often used in behavior change designs/interventions [3, 61], and they are not directly connected to queries provided by any participant. From these 4 papers, we randomly selected one card each from the related set and the not-related set to augment with customization features, while the other two papers were used to generate default versions of our design cards.

Each session lasted approximately one hour. During the first 5 minutes upon the participant joining the study via Zoom, we provided an overview of the purpose and procedure of the study and introduced our system to the participants. Also, participants were asked to identify the design project they submitted as a query during the study signup, and use its design scenario to generate, view, and evaluate the cards. The system then presented a screen where they could generate four design cards (two customized cards, and two default cards) using one of the design implications present in each paper, with the presentation order randomized. After each card was shown, participants were directed to a survey screen, where they answered the survey questionnaires on a 7-point Likert scale, as detailed in Section 5.3. The process lasted approximately 40 minutes.

After viewing all cards, participants were also asked to indicate their preferred card format. Then, we conducted a semi-structured interview including questions about (i) the rationale for their preferences, (ii) perceptions of our customized cards, and (iii) suggestions for future improvements of the customized translational research artifacts, which took approximately 20 minutes. The interview was voice-recorded and later transcribed for our analysis. Upon completion, each participant was compensated with a 30 USD gift card. The study procedure was reviewed and approved by the IRB of our university's human subjects division.

5.3 Survey Measure & Hypotheses

The primary motivation of our study is to make insights from research papers more relevant to the individual designer's needs. Thus, in our quantitative survey, we first measured the relevance of the customized cards to their work and the inputs they provided. Additionally, prior literature [66] suggested six core dimensions for communicating design insights in research papers to practitioners (*i.e.*, generativity, inspirability, actionability, originality, generalizability, and validity), which have also been used to evaluate the quality of translational artifacts [69]. Following this prior work, we also measured these six dimensions in our survey. To ensure clarity and consistency in the evaluation metrics, the questionnaire incorporated the definitions from the original paper (*e.g.*, generalizability—the ability to be extended beyond the design context). We hypothesized that customization would enhance relevance and qualities associated with usefulness (*i.e.*, generativity, inspirability, and actionability), while preserving qualities associated with rigor (*i.e.*, originality, generalizability, and validity).

5.4 Analyses

5.4.1 Quantitative survey. To assess the quantitative outcomes, we utilized a linear mixed-effects model to examine the significant differences in relevance and perceived qualities between the two design card formats. In our model, we included the type of paper (*i.e.*, topically related vs. not topically related) as a control variable, while participant ID was treated as a random variable.

5.4.2 Interview. We conducted a thematic analysis of the responses from the user study to understand the user perception of the use of LLMs in supporting the customization of translational artifacts, with the following bottom-up approach: The individual authors reviewed the responses to become familiar with the data, and we identified emerging themes and grouped the key ones, where the initial themes were discussed and refined in four rounds. As such, we reached a consensus on the final set of themes, as described in Section 6.2, where we refer to each participant as P1 – P20, with participant order randomized.

5.4.3 Intrinsic evaluation. All customized design cards generated by participants in the study were reviewed to assess the quality of AI-generated content. We broadly defined a 'mismatch' as any component on the customized card where the model-generated content (i) was false or misleading, (ii) lacked grounding to the paper content and/or the user's design context, or (iii) made assumptions about the paper's content or the user's design context even when they were absent. The authors independently annotated design card components to identify mismatches. They then merged their annotations, identified classes of mismatches based on the merged annotation, discussed to refine key classes, and re-reviewed the design cards independently to identify mismatches based on the defined set of classes. This process was repeated for three rounds until annotations and classes were finalized.

6 EVALUATION RESULTS

6.1 Quantitative Survey

Our quantitative analysis (see Figure 3) revealed that the designers in our study perceived the customized design cards to be significantly more relevant to their design work ($M = 5.35$, $SD = 1.51$ vs. $M = 3.80$, $SD = 1.79$; $t = 4.87$, $p < 0.001$). This relationship is consistent for both types of papers we used (*i.e.*, topically related papers and not topically related papers); however, topical relatedness does improve the perceived relevance ($t = 2.35$, $p < 0.05$).

When analyzing the other measures, we found that the customized cards were reported to be more generative ($M = 4.83$, $SD = 1.41$ vs. $M = 4.23$, $SD = 1.64$; $t = 2.10$, $p < 0.05$), inspiring ($M = 5.50$, $SD = 1.62$ vs. $M = 4.55$, $SD = 1.83$; $t = 3.15$, $p < 0.01$), and actionable ($M = 5.33$, $SD = 1.40$ vs. $M = 3.60$, $SD = 1.72$; $t = 5.25$, $p < 0.001$), compared to the default design cards. Also, there was no interaction effect with topical relatedness.

With regards to the other dimensions, we found that there was no difference in originality ($M = 4.80$, $SD = 1.59$ vs. $M = 4.83$, $SD = 1.62$; $t = -0.11$, $p = 0.92$) and generalizability ($M = 5.40$, $SD = 1.43$ vs. $M = 5.05$, $SD = 1.45$; $t = 1.39$, $p = 0.17$) between customized cards and default cards. Interestingly, design cards customized by our system were viewed as more valid ($M = 5.53$, $SD = 1.38$), compared to default cards ($M = 4.78$, $SD = 1.70$; $t = 2.78$, $p < 0.01$). Similar to the above measures, there was also no observed interaction effect with topical relatedness.

Of the participants, 18 participants preferred the contents communicated through our customized cards, 1 participant responded that it depends on the purpose of consuming the design insights (*i.e.*, pure learning purpose vs. diving deeper into the paper), and 1 participant preferred being presented with the default cards only.

6.2 Interview

Below, we describe insights gained from our analyses of participants' qualitative responses:

6.2.1 Our pipeline helped designers to extend the notion of relevance beyond surface-level (e.g., topical) similarity. Participants noted that they often fixate too much on the specific domain of the papers, dismissing those that did not topically align as irrelevant: *"Initially, when I looked at it, I was like, I don't know how it can be relevant to my project or my study, because at first I thought that, okay, my product is not about like [topic] or something."* (P16)

Using our pipeline, however, participants could uncover underlying implications by identifying the similarities and differences between contexts, even across distinct domains. Being able to extract and leverage the potential applicability across design context and design implication highlighted in the papers, and apply analogies to identify similarities relevant to the designer's domain, participants found they could extend the notion of relevance beyond mere topical relatedness, and were able to uncover the underlying utility of these papers: *"Relevant to the problem trying to solve definitely (...) because (it said) both (paper and designer's context) can leverage on some sort of rewarding system. And then it kind of spewed out everything about the product that was designed for which I thought was funny, like, oh, you got tier system, and you can include badges to kind of get people to report."* (P3)

Similarly, participants highlighted the role of our customization pipeline in assisting them in pinpointing specific techniques (*e.g.*, methods, metrics) from the paper that they might have overlooked due to topic differences. Through our pipeline, participants were able to identify valuable concepts that could still be applicable to their projects but may have previously been overlooked due to topical differences, ultimately enhancing the utility of the paper: *"At first, it felt kind of far from my project. But there (customized cards) were still some nuggets in the hair that felt okay, I'll probably use this in my project, which is surprising (...) So if I was in the research phase (the stage that the participant provided) of my project, I'd probably like to look deeper into this."* (P11)

Based on this effectiveness, participants expected that our pipeline would help designers draw insights from related but distinct fields, potentially inspiring innovative designs: *"It's quite helpful in the sense that you might not always find the right papers, because maybe you're creating something that's not been created before (...) that's like one of the ways to create novel designs where you borrow things from different fields and apply to a field that's got nothing to do with it."* (P18) This aligns with our quantitative results,

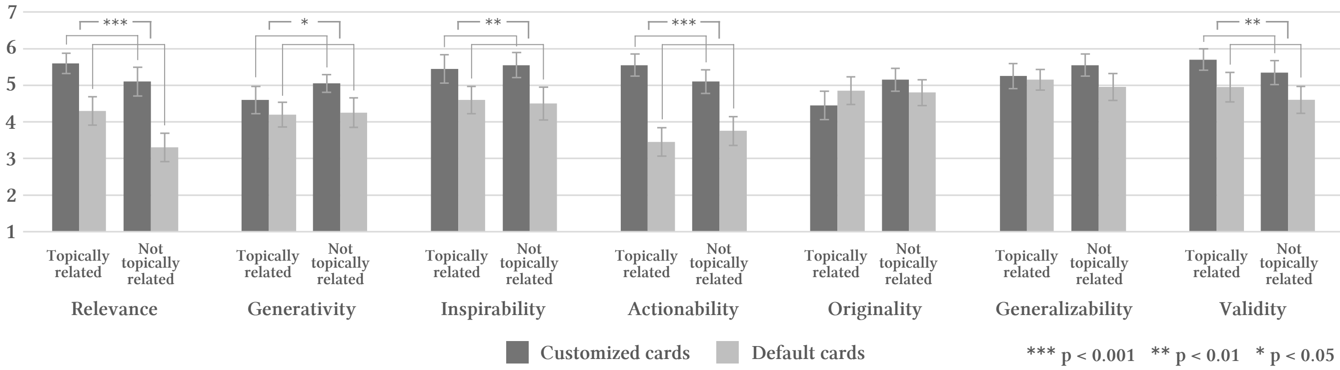


Figure 3: Perceived qualities of the design cards for each type. For each bar group, the result for cards built on a topically related paper is presented on the left, and the result for cards built on a not topically related paper is on the right. The significance level for each quality is based on the main effect (*i.e.*, card format; customized cards vs. default cards), with the bars indicating standard errors.

where the main effect was significant while no interaction effect was observed.

6.2.2 Information provenance supported the credibility of customized design insights. From our study, participants highlighted that the pipeline’s provenance features contributed to their perceiving the informational content as credible, especially after customization. Rather than presenting information in isolation, our pipeline allows users to view the specific section from which each informational component is derived through a hovering tooltip. Being able to trace the origins of the recommendations, participants were able to quickly capture their basis in this translational process: *“I think it’s super important to have it (source), because the provenance of where you’re building things from is super murky sometimes. So this is really, really good that you have on.”* (P19) This, in turn, enhanced the validity of the resources customized to their design work: *“(customized cards are) valid, because it basically extracted information from the paper, and it just adjusted the way of showing the information according to my prompt. So I really didn’t think that it made up the information or provided me with false information.”* (P13)

At the same time, participants desired enhanced cross-referencing by integrating ways to take users directly to the specific sections of the papers being referenced. Currently, our pipeline supports provenance only through quotes, but participants hoped to clarify concepts and access surrounding information to further enrich their understanding. For instance, they suggested having a side-by-side view where the original paper is displayed alongside the cards, automatically locating the relevant section to enhance understanding: *“So maybe, as I hover over this (understanding users), it would scroll me automatically to the section.”* (P14)

6.2.3 Prior familiarity with the presented information affects perceived relevance. One intriguing finding from our user study is that, participants’ prior familiarity with the concepts presented influenced their evaluation of relevance. When participants encountered methods or metrics they had not previously considered for their projects, they viewed our pipeline as generating applicable insights that they had overlooked: *“I think this was pretty useful, like the reminders and action items. I think like dashboard systems, it would be nice if it had reminders or action items. And entrepreneurs, they need to have goals. So this might actually be a feature that would be interesting to implement.”* (P13) On the contrary, when the pipeline presented methodologies they were already familiar with or had previously used, they perceived the insights as less relevant: *“It may be not that much relevant and helpful, because it did give me some very generic stuff (...) it’s something you know and you see everywhere.”* (P18)

6.2.4 Validity is enhanced by improved understandability through customization. Our quantitative data showed that validity was higher in our customized cards than in the default cards. This was surprising because our customization was only intended to help improve the usefulness of reported findings. Through our qualitative analyses, we found that the perceived improvement in validity may be attributed to the improved understandability afforded by our customization. When there was no clear connection between a paper and the designer’s work in default cards, participants struggled to grasp the insights enough to assess if the presented information was valid. This was most prevalent for translational artifacts from not topically related papers: *“I don’t know if I had to say it (default cards) is like, valid or not, because I don’t really fully understand that.”* (P16) Through our pipeline, designers were able to better comprehend design insights from the research papers through more contextualized understandings, such as comparing and contrasting the insights with their own work. This improved comprehension allowed participants to more easily gauge the validity of insights, and subsequently affirm the validity of information: *“It’s valid. It kind of shows you its comparisons and contrasts (...) It has a lot of built-in proofs which refer to specifics.”* (P3)

However, even when digesting insights from topically related research papers, the added contents from our customized cards helped elaborate key concepts (*e.g.*, providing definitions or contextual explanations included in the tooltip), helped participants understand the design insights, and made the insights seem more valid: *“Sometimes I encounter terms I’m not very familiar with, like I don’t understand. Now, by using tooltips, I get an idea which is a very

Table 3: Types of mismatches found in the customized cards created by participants using our pipeline. For each class of mismatch, we provide the components in which they appeared and an example.

Mismatch	Appearing component	Example
Misinterpreting the paper's focus	Inspiration and scope ($N = 5$)	Highlighting the paper's focus as designing solutions to assist *energy experts* in saving energy, yet its target audience is *any individual* aiming to save energy
Misinterpreting the user's design context	Inspiration and scope ($N = 2$) Design ideas ($N = 2$) Understanding users ($N = 1$)	Proposing that the participant's target users (*i.e.*, *teaching assistants*) may benefit from the automation of administrative tasks, while the participant wanted to design an AI-based interface to *assess* their performance and provide feedback to them
Offering overly generic solutions without sufficient specificity/rationale	Understanding users ($N = 3$) Design ideas ($N = 2$)	Returning *Understanding User-specific Contexts* as a user need without providing a compelling connection specific to the participant's input
Conceptual misclassification	Methods for you ($N = 5$) Metrics for you ($N = 2$) Understanding users ($N = 1$)	Proposing a theory cited by the paper (*i.e.*, *Construal Level Theory*) as a method to consider without specifying how it can be used as a design method

brief but clear explanation, and helps me to get like, what the [term] may be (...) I feel they are valid." (P15)

6.3 Intrinsic Evaluation

Post-hoc analysis of the customized card contents revealed that the generated content was generally faithful to both the contents of the source papers and the participants' design contexts. We reviewed and present results at the component level. Our customization pipeline is designed to generate 4 to 7 customized components per card depending on the user's specified design stage, and all together, the participants generated a total of 233 components through customization.

We identified no issues in 210 (or 90.13%) of the generated components. The remaining components ($N = 23$) contained mismatches categorized into four types: misinterpreting the paper's focus ($N = 5$), misinterpreting the user's design context ($N = 5$), offering overly generic solutions without sufficient specificity or rationale ($N = 5$), and conceptual misclassification (*e.g.*, treating theories as design methods; $N = 8$). Mismatches occurred less frequently when customizing topically related papers ($M = 0.37$ annotated mismatches per each customized card) compared to customizing not topically related papers ($M = 0.80$; $t = 2.40, p < 0.05$). In Table 3, we provide counts of appearing components and examples for each class.

7 DISCUSSION

This study advances research on translational science by exploring the use of LLMs to improve the relevance of research findings to an individual practitioner's goals. Specifically, our pipeline takes in key dimensions of perceived relevance for designers, extracts and integrates content from the research paper and designer's goals, and adapts the output to the designer's current design stage. Our empirical study showed that the resulting customized translational artifacts improved the actionability, validity, generativity, and inspirability of research to designers.

One unexpected finding from our study was that customization led to an improvement in the perceived validity of the design cards. Our qualitative results suggest that users' ability to better understand design insights generated from our pipeline, along with their ability to evaluate the information using provided provenance, contributed to an increase in perceived validity. By providing more specific and/or relevant information in the customized version of the design cards, the customized design cards helped designers more easily gauge validity when compared to the default cards. In other words, sufficient information from the paper may need to be provided in the translational artifact to help users assess validity.

Our study also revealed that a designer's prior familiarity with the research affects their perception of its relevance. While this aspect of relevance did not appear in our preliminary study, it has been noted in a prior discussion of rigor-relevance in management science, where the practitioners' perception of relevance is shaped by prior exposure to the information [21]. This also connects to Relevance Theory in communication, where people are more likely to find communication relevant if it yields a positive cognitive effect—such as filling a gap in knowledge [88]. This finding highlights the importance of considering familiarity in the customization process for design practitioners, in addition to the *who*, *what*, and *design stage* of the design process we have already identified and implemented. To facilitate this type of customization, one possibility is to integrate our pipeline within the designer's existing workflow and potentially with the designer's design history. For example, it is common for designers to organize insights using online canvas tools like Miro or Figma [19]. By leveraging these workplaces, our pipeline could better integrate a designer's prior experiences and current project details into the customization process.

During the evaluation, we observed that designers would modify customization inputs as a way of asserting their intentions. For example, designers would adjust the specificity of their inputs to influence the specificity of outputs, *i.e.*, if the output was not specific enough, they would provide more details in the inputs before regenerating. These input modifications sometimes resulted in the correct output behavior, but our pipeline could be enhanced to better address user intent in these cases. One possibility is to augment the pipeline to explicitly model multi-turn interactions, by providing customization history in prompts. Another solution may be to allow the user to provide direct feedback to the system on how to modify the output.

While generated card components were largely faithful to the research paper and participants' provided design contexts, our intrinsic evaluation did find potential mismatches in 9.87% of all generated components. In none of these cases did the model hallucinate false information, but rather misinterpret or make assumptions about the paper or designers' inputs. Despite this, these errors may impact the real-world design scenarios. Overly generic suggestions, for instance, might leave designers without actionable direction—particularly during early ideation or when navigating complex constraints—while misclassified concepts or misconstrued focus of the paper could lend undue credibility to strategies that appear research-grounded but are in fact misaligned, leading designers to overlook more context-appropriate approaches. To mitigate these risks, several NLP techniques could be incorporated; for instance, when the user's description of their design context lacks sufficient specificity to generate tailored insights for a target design context, the pipeline could prompt for details in a follow-up question. Techniques like chain-of-thought prompting [86] and LLM-as-a-Judge [28] could also be employed to enhance the alignment of the output with user intent, inputs, paper content, and instructions. Finally, to address conceptual misclassification, future iterations of the customization pipeline could incorporate retrieval mechanisms (*e.g.*, retrieval-augmented generation [44]) to connect the model with an external definition library of design methods.

Additionally, while the current pipeline does not have the paper authors in the loop for customizing artifacts, we believe there is an opportunity for them to be involved in this process. For example, as the system scales up, authors could optionally contribute to a shared translational research library by annotating their papers with potential applications or use cases they envisioned but were unable to evaluate within the scope of their study. These speculative extensions—grounded in the authors' deep understanding of their work—could serve as valuable signals for customization, helping the pipeline better align outputs with the underlying design intent while expanding the space of plausible adaptations. Authors might also highlight contextual assumptions or constraints that shaped their original study but could be relaxed or reimagined in different design scenarios. Such author-provided cues would not only reduce misinterpretations but also enable more grounded yet generative translations of research into practice.

Finally, with our approach, we found that designers could potentially benefit from an even broader set of literature beyond what is normally returned by search engines optimized for topical relevance. Papers that are not obviously topically related to a designer's project also gained relevance, generativity, inspirability, and actionability from customization, and the customized insights were perceived as valuable for the designers' work. This highlights the value of creating translational artifacts from papers that may not be directly topically related to the designer's design work, and supports previous research, which suggested the value of drawing analogical ideas from other domains to create design ideas [25].

While promising, there may be limits to how far out-of-domain a paper can be for customization to still improve relevance in this way. For example, it is unrealistic to expect papers on ergonomic design to significantly inspire AI-based web interface designers. Our intrinsic evaluation also showed that mismatches were more likely to occur when customizing papers that are not topically related. This trend would be expected to worsen as the domain gap increases, as there will be fewer analogical insights to draw from papers that are far removed from the target domain, potentially resulting in the LLM producing forced analogies or misinterpretations. Thus, while we are encouraged by the potential broadening of the scope of research available to support design practice, we do not believe that *any* research paper could be customized for *any* design context. Future work on research translation should instead aim to identify the optimal balance between helping designers identify potentially relevant papers *and* making those papers more relevant through customization.

8 LIMITATION & FUTURE WORK

Our work showed promise in using generative AI to customize design insights from academic papers to individual designers' needs, with potential impacts on their work. To better align the participants' evaluation of our pipeline with their real-world design scenarios, we asked participants to reflect on their recent design project to inform their customization and evaluation, guiding us to better understand their perceived usefulness of our pipeline given the context of their real-world designing. Still, these insights are limited in fully understanding how their actual design works may benefit from our pipeline, and future research should investigate how designers use these customized artifacts in real-world environments. Additionally, while we also observed benefits with our customization for the not topically related papers, we only tested two of these papers and there may be limits to how far out-of-domain a paper can be for our approach to still be useful. Thus, future work should examine these boundaries.

9 CONCLUSION

In this paper, we explored a method to overcome the rigor-relevance paradox in design. Based on preliminary interviews with designers, we identified key aspects of relevance, then proposed and developed a pipeline that leverages LLMs to customize papers to improve their relevance to a designer's specific design context. Results from our evaluation with designers showed that the customized design cards were more relevant, actionable, valid, generative, and inspiring compared to those without customization, and were largely faithful to the content of the source research papers and design contexts. Our work demonstrates the potential of automated customization to help close the research-practice gap in design practice. While our study focuses on design cards, our findings inform how customization might be operationalized for other types of translational artifact, such as playbooks or editable design templates. We encourage the extension of our methods to other formats of translational artifacts, as well as to other domains and disciplines beyond design.

ACKNOWLEDGMENTS

We would like to thank Raymond Fok and Tal August for their valuable insights into the feasibility of this direction and its implementation. We also thank Daniel Lee and the members of the UW LARCH Lab & Prosocial Computing Lab for their thoughtful feedback. This work was supported by the gift fund from the Allen Institute for AI, as well as the Accelerating Foundation Models Research (AFMR) grant program from Microsoft and the Academic Research Award from Google.

REFERENCES

[1] Ali Abdolrahmani, Maya Howes Gupta, Mei-Lian Vader, Ravi Kuber, and Stacy Branham. 2021. Towards More Transactional Voice Assistants: Investigating the Potential for a Multimodal Voice-Activated Indoor Navigation Assistant for Blind and Sighted Travelers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16. https://doi.org/10.1145/3411764.3445638

[2] Dragan Ahmetovic, Daniele Riboli, Cristian Bernareggi, and Sergio Mascetti. 2021. RePlay: Touchscreen Interaction Substitution Method for Accessible Gaming. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 1–12. https://doi.org/10.1145/3447526.3472044
[3] Icek Ajzen. 1991. The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes* (1991). https://doi.org/10.1016/0749-5978(91)90020-T
[4] Riku Arakawa and Hiromu Yakura. 2024. Coaching Copilot: Blended Form of an LLM-Powered Chatbot and a Human Coach to Effectively Support Self-Reflection for Leadership Growth. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–14. https://doi.org/10.1145/3640794.3665549
[5] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2022. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* 30 (2022), 1 – 38. https://doi.org/10.1145/3589955
[6] Pratima Bansal, Stephanie Bertels, Tom Ewart, Peter MacConnachie, and James O'Brien. 2012. Bridging the Research–Practice Gap. https://doi.org/10.5465/amp.2011.0140
[7] Amanda Baughan, Tal August, Naomi Yamashita, and Katharina Reinecke. 2020. Keep it Simple: How Visual Complexity and Preferences Impact Search Efficiency on Websites. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376849
[8] Jordan Beck and Hamid R Ekbia. 2018. The Theory-Practice Gap as Generative Metaphor. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3173574.3174194
[9] Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2023. CAM: A Large Language Model-based Creative Analogy Mining Framework. In *Proceedings of the ACM Web Conference 2023*. 3903–3914. https://doi.org/10.1145/3543507.3587431
[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706QP063OA
[11] Tianying Chen, Margot Stewart, Zhiyu Bai, Eileen Chen, Laura Dabbish, and Jessica Hammer. 2020. Hacked Time: Design and Evaluation of a Self-Efficacy Based Cybersecurity Game. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1737–1749. https://doi.org/10.1145/3357236.3395522
[12] Bernard C. K. Choi and Anita W. P. Pak. 2007. Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. Promotors, barriers, and strategies of enhancement. *Clinical and Investigative Medicine* 30, 6 (2007), E224–E232. https://doi.org/10.25011/cim.v30i6.2950
[13] Christopher Clark and Santosh Divvala. 2016. PDFFigures 2.0: Mining Figures from Research Papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. 143–152. https://doi.org/10.1145/2910896.2910904
[14] Lucas Colusso, Cynthia L Bennett, Gary Hsieh, and Sean A Munson. 2017. Translational Resources: Reducing the Gap Between Academic Research and HCI Practice. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 957–968. https://doi.org/10.1145/3064663.3064667
[15] Lucas Colusso, Ridley Jones, Sean A Munson, and Gary Hsieh. 2019. A Translational Science Model for HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300231
[16] Richard L. Daft and Arie Y. Lewin. 2008. Perspective—Rigor and Relevance in Organization Studies: Idea Migration and Academic Journal Evolution. *Organization Science* 19, 1 (2008), 177–183. https://doi.org/10.1287/orsc.1070.0346
[17] Ravit Dotan, Lisa S Parker, and John Radzilowicz. 2024. Responsible Adoption of Generative AI in Higher Education: Developing a "Points to Consider" Approach Based on Faculty Perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2033–2046. https://doi.org/10.1145/3630106.3659023
[18] John Dunham, Konstantinos Papangelis, Cati Boulanger, Nicolas Lalone, Evangelia Nika, Michael Saker, and David Schwartz. 2023. Building Positively Affective Location-Based Advertising: A Study of Pokémon GO Players. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3544548.3580939
[19] KJ Kevin Feng, Tony W Li, and Amy X Zhang. 2023. Understanding Collaborative Practices and Tools of Professional UX Practitioners in Software Organizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20. https://doi.org/10.1145/3544548.3581273
[20] Raymond Fok, Joseph Chee Chang, Tal August, Amy X Zhang, and Daniel S Weld. 2023. Qlarify: Recursively Expandable Abstracts for Directed Information Retrieval over Scientific Papers. *arXiv preprint arXiv:2310.07581* (2023). https://arxiv.org/abs/2310.07581
[21] Stephen Fox and Stefan N Groesser. 2016. Reframing the relevance of research to practice. *European Management Journal* 34, 5 (2016), 457–465. https://doi.org/10.1016/j.emj.2016.07.005
[22] Guo Freeman and Divine Maloney. 2021. Body, Avatar, and Me: The Presentation and Perception of Self in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27. https://doi.org/10.1145/3432938
[23] Michael Fröhlich, Maurizio Raphael Wagenhaus, Albrecht Schmidt, and Florian Alt. 2021. Don't Stop Me Now! Exploring Challenges Of First-Time Cryptocurrency Users. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. 138–148. https://doi.org/10.1145/3461778.3462071
[24] Anna Gausen, Bhaskar Mitra, and Siân Lindley. 2024. A Framework for Exploring the Consequences of AI-Mediated Enterprise Knowledge Access and Identifying Risks to Workers. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 207–220. https://doi.org/10.1145/3630106.3658900
[25] Karni Gilon, Joel Chan, Felicia Y Ng, Hila Liifshitz-Assaf, Aniket Kittur, and Dafna Shahaf. 2018. Analogy Mining for Specific Design Needs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3173574.3173695
[26] Philippe Giraudeau, Alexis Olry, Joan Sol Roo, Stephanie Fleck, David Bertolo, Robin Vivian, and Martin Hachet. 2019. CARDS: A Mixed-Reality System for Collaborative Learning at School. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*. 55–64. https://doi.org/10.1145/3343055.3359721
[27] Linda Graf, Katharina Emmerich, Stefan Liszio, and Maic Masuch. 2024. The Impact of Emotional Virtual Characters on Emotional State and Player Experience in VR Horror Games. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*. 1–13. https://doi.org/10.1145/3649921.3649942
[28] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594* (2024). https://arxiv.org/abs/2411.15594
[29] Ranjay Gulati. 2007. Tent Poles, Tribalism, and Boundary Spanning: The Rigor-Relevance Debate in Management Research. *Academy of Management Journal* 50, 4 (2007), 775–782. https://doi.org/10.5465/amj.2007.26279170
[30] Yue Guo, Joseph Chee Chang, Maria Antoniak, Erin Bransom, Trevor Cohen, Lucy Wang, and Tal August. 2024. Personalized Jargon Identification for Enhanced Interdisciplinary Communication. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 4535–4550. https://doi.org/10.18653/v1/2024.naacl-long.255
[31] Helen Ai He, Saul Greenberg, and Elaine M Huang. 2010. One Size Does Not Fit All: Applying the Transtheoretical Model to Energy Feedback Technology Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 927–936. https://doi.org/10.1145/1753326.1753464
[32] Nico Herbig, Gerrit Kahl, and Antonio Krüger. 2018. Design Guidelines for Assistance Systems Supporting Sustainable Purchase Decisions. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 1333–1344. https://doi.org/10.1145/3196709.3196726
[33] Gary Hsieh, Brett A Halperin, Evan Schmitz, Yen Nee Chew, and Yuan-Chi Tseng. 2023. What is in the Cards: Exploring Uses, Patterns, and Trends in Design Cards. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580712
[34] Gaurav Jain, Yuanyang Teng, Dong Heon Cho, Yunhao Xing, Maryam Aziz, and Brian A Smith. 2023. "I Want to Figure Things Out": Supporting Exploration in Navigation for People with Visual Impairments. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–28. https://doi.org/10.1145/3579496
[35] Qiao Jin, Saba Kawas, Stuti Arora, Ye Yuan, and Svetlana Yarosh. 2024. Is Your Family Ready for VR? Ethical Concerns and Considerations in Children's VR Usage. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. 436–454. https://doi.org/10.1145/3628516.3655804
[36] Yucheng Jin, Wanling Cai, Li Chen, Yizhe Zhang, Gavin Doherty, and Tonglin Jiang. 2024. Exploring the Design of Generative AI in Supporting Music-based Reminiscence for Older Adults. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3613904.3642800
[37] Anjali Khurana, Michael Glueck, and Parmit K Chilana. 2024. Do I Just Tap My Headset? How Novice Users Discover Gestural Interactions with Consumer Augmented Reality Applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–28. https://doi.org/10.1145/3631451
[38] Alfred Kieser and Lars Leiner. 2009. Why the Rigour–Relevance Gap in Management Research Is Unbridgeable. *Journal of Management Studies* 46, 3 (2009), 516–533. https://doi.org/10.1111/j.1467-6486.2009.00831.x
[39] René F Kizilcec. 2016. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–2395. https://doi.org/10.1145/2858036.2858402
[40] Yasmine Kotturi, Herman T Johnson, Michael Skirpan, Sarah E Fox, Jeffrey P Bigham, and Amy Pavel. 2022. Tech Help Desk: Support for Local Entrepreneurs Addressing the Long Tail of Computing Challenges. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3491102.3517708
[41] Yubo Kou, Sam Moradzadeh, and Xinning Gui. 2024. Trading as Gambling: Social Investing and Financial Risks on the r/WallStreetBets Subreddit. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3613904.3642768
[42] Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. Many: Comprehending Accurate

Information from Multiple Erroneous and Inconsistent AI Generations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2518–2531. https://doi.org/10.1145/3630106.3662681

[43] James J Leisenring and L Todd Johnson. 1994. Accounting Research: On the Relevance of Research to Practice. *Accounting Horizons* 8, 4 (1994), 74.

[44] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[45] Hsin-Ni Lin, Shu-Kai Hsieh, and Shiao-Hui Chan. 2012. Measuring Individual Differences in Word Recognition: The Role of Individual Lexical Behaviors. In *Proceedings of the Twenty-Fourth Conference on Computational Linguistics and Speech Processing*. https://aclanthology.org/O12-1007.pdf

[46] Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries*. Springer Berlin Heidelberg, 473–474. https://doi.org/10.1007/978-3-642-04346-8_62

[47] Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 6929–6947. https://doi.org/10.18653/v1/2023.findings-acl.433

[48] Divine Maloney, Guo Freeman, and Andrew Robb. 2020. A Virtual Space for All: Exploring Children's Experience in Social Virtual Reality. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 472–483. https://doi.org/10.1145/3410404.3414268

[49] Divine Maloney, Guo Freeman, and Andrew Robb. 2021. Stay Connected in An Immersive World: Why Teenagers Engage in Social Virtual Reality. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference*. 69–79. https://doi.org/10.1145/3459990.3460703

[50] Alejandro Martínez and Stefano Mammola. 2021. Specialized terminology reduces the number of citations of scientific papers. *Proceedings of the Royal Society B* 288 (2021). https://doi.org/10.1098/rspb.2020.2581

[51] Liam Mason, Kathrin Gerling, Patrick Dickinson, Jussi Holopainen, Lisa Jacobs, and Kieran Hicks. 2022. Including the Experiences of Physically Disabled Players in Mainstream Guidelines for Movement-Based Games. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3501867

[52] Donald McIntyre. 2005. Bridging the gap between research and practice. *Cambridge Journal of Education* 35, 3 (2005), 357–382. https://doi.org/10.1080/03057640500319065

[53] Brenna McNally, Priya Kumar, Chelsea Hordatt, Matthew Louis Mauriello, Shalmali Naik, Leyla Norooz, Alazandra Shorter, Evan Golub, and Allison Druin. 2018. Co-designing Mobile Online Safety Applications with Children. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–9. https://doi.org/10.1145/3173574.3174097

[54] Anika Nissen, Colin Conrad, and Aaron Newman. 2023. Are You Human? Investigating the Perceptions and Evaluations of Virtual Versus Human Instagram Influencers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3544548.3580943

[55] Donald A. Norman. 2010. The Research-Practice Gap: The Need for Translational Developers. *Interactions* 17, 4 (July 2010), 9–12. https://doi.org/10.1145/1806491.1806494

[56] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3173574.3174223

[57] OpenAI. 2024. *GPT-4o*. https://openai.com/index/hello-gpt-4o/

[58] Jill W Paine and Fred Delmhorst. 2020. A Balance of Rigor and Relevance: Engaged Scholarship in Organizational Change. *The Journal of Applied Behavioral Science* 56, 4 (2020), 437–461. https://doi.org/10.1177/0021886320961218

[59] Chris Parnin and Carsten Gorg. 2007. Design Guidelines for Ambient Software Visualization in the Workplace. In *2007 4th IEEE International Workshop on Visualizing Software for Understanding and Analysis*. IEEE, 18–25. https://doi.org/10.1109/VISSOF.2007.4290695

[60] David Porfirio, Allison Sauppé, Aws Albarghouthi, and Bilge Mutlu. 2018. Authoring and Verifying Human-Robot Interactions. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 75–86. https://doi.org/10.1145/3242587.3242634

[61] James O Prochaska and Wayne F Velicer. 1997. The Transtheoretical Model of Health Behavior Change. *American Journal of Health Promotion* 12, 1 (1997), 38–48. https://doi.org/10.4278/0890-1171-12.1.38

[62] Jeba Rezwana and Mary Lou Maher. 2022. Understanding User Perceptions, Collaborative Experience and User Engagement in Different Human-AI Interaction Designs for Co-Creative Systems. In *Proceedings of the 14th Conference on Creativity and Cognition*. 38–48. https://doi.org/10.1145/3527927.3532789

[63] Wendy Roldan, Xin Gao, Allison Marie Hishikawa, Tiffany Ku, Ziyue Li, Echo Zhang, Jon E Froehlich, and Jason Yip. 2020. Opportunities and Challenges in Involving Users in Project-Based HCI Education. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3313831.3376530

[64] Sara L Rynes, Jean M Bartunek, and Richard L Daft. 2001. Across the Great Divide: Knowledge Creation and Transfer Between Practitioners and Academics. *Academy of Management Journal* 44, 2 (2001), 340–355. https://doi.org/10.5465/3069460

[65] Malak Sadek, Marios Constantinides, Daniele Quercia, and Céline Mougenot. 2024. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20. https://doi.org/10.1145/3613904.3642810

[66] Corina Sas, Steve Whittaker, Steven Dow, Jodi Forlizzi, and John Zimmerman. 2014. Generating Implications for Design through Design Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1971–1980. https://doi.org/10.1145/2556288.2557357

[67] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20. https://doi.org/10.1145/3613904.3642152

[68] Donghoon Shin, Jaeyoon Song, Seokwoo Song, Jisoo Park, Joonhwan Lee, and Soojin Jun. 2020. TalkingBoogie: Collaborative Mobile AAC System for Nonverbal Children with Developmental Disabilities and Their Caregivers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3313831.3376154

[69] Donghoon Shin, Lucy Lu Wang, and Gary Hsieh. 2024. From Paper to Card: Transforming Design Implications with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3613904.3642266

[70] Lucas M Silva, Franceli L Cibrian, Clarisse Bonang, Arpita Bhattacharya, Aehong Min, Elissa M Monteiro, Jesus Armando Beltran, Sabrina Schuck, Kimberley D Lakes, Gillian R Hayes, et al. 2024. Co-Designing Situated Displays for Family Co-Regulation with ADHD Children. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3613904.3642745

[71] Wendy K Smith and Marianne W Lewis. 2011. Toward a Theory of Paradox: A Dynamic equilibrium Model of Organizing. *Academy of Management Review* 36, 2 (2011), 381–403. https://doi.org/10.5465/amr.2009.0223

[72] Jose Luis Soler-Dominguez, Samuel Navas-Medrano, and Patricia Pons. 2024. ARCADIA: A Gamified Mixed Reality System for Emotional Regulation and Self-Compassion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3613904.3642123

[73] Evropi Stefanidi, Johannes Schöning, Yvonne Rogers, and Jasmin Niess. 2023. Children with ADHD and their Care Ecosystem: Designing Beyond Symptoms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3544548.3581216

[74] Miriam Sturdee, Lauren Thornton, Bhagya Wimalasiri, and Sameer Patil. 2021. A Visual Exploration of Cybersecurity Concepts. In *Proceedings of the 13th Conference on Creativity and Cognition*. 1–10. https://doi.org/10.1145/3450741.3465252

[75] Minhyang Suh and Gary Hsieh. 2016. Designing for Future Behaviors: Understanding the Effect of Temporal Distance on Planned Behaviors. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1084–1096. https://doi.org/10.1145/2858036.2858591

[76] Kaiwen Sun, Jingjie Li, Yixin Zou, Jenny Radesky, Christopher Brooks, and Florian Schaub. 2024. Unfulfilled Promises of Child Safety and Privacy: Portrayals and Use of Children in Smart Home Marketing. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–29. https://doi.org/10.1145/3637422

[77] Zhida Sun, Sitong Wang, Chengzhong Liu, and Xiaojuan Ma. 2022. Metaphoraction: Support Gesture-based Interaction Design with Metaphorical Meanings. *ACM Transactions on Computer-Human Interaction* 29, 5 (2022), 1–33. https://doi.org/10.1145/3511892

[78] Svelte. 2024. *SvelteKit*. https://kit.svelte.dev/

[79] Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. 2021. User Expectations of Conversational Chatbots Based on Online Reviews. In *Designing Interactive Systems Conference 2021*. https://doi.org/10.1145/3461778.3462125

[80] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24. https://doi.org/10.1145/3613904.3642902

[81] Rajan Vaish, Shirish Goyal, Amin Saberi, and Sharad Goel. 2018. Creating Crowdsourced Research Talks at Scale. In *Proceedings of the 2018 World Wide Web Conference*. 1–11. https://doi.org/10.1145/3178876.3186031

[82] Artemij Voskobojnikov, Oliver Wiese, Masoud Mehrabi Koushki, Volker Roth, and Konstantin Beznosov. 2021. The U in Crypto Stands for Usable: An Empirical Study of User Experience with Mobile Cryptocurrency Wallets. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3411764.3445407

[83] Gang Wang, Tianyi Wang, Bolun Wang, Divya Sambasivan, Zengbin Zhang, Haitao Zheng, and Ben Y Zhao. 2015. Crowds on Wall Street: Extracting Value from Collaborative Investing Platforms. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 17–30. https://doi.org/10.1145/2675133.2675144

[84] Yiwen Wang, Mengying Li, Young-Ho Kim, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2024. Redefining Activity Tracking Through Older Adults' Reflections on Meaningful Activities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3613904.3642170

[85] Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour* 7, 9 (2023), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

[86] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837. https://doi.org/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[87] Justin D Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22. https://doi.org/10.1145/3613904.3642466

[88] Deirdre Wilson and Dan Sperber. 2006. Relevance theory. *The Handbook of Pragmatics* (2006), 606–632. https://doi.org/10.1002/9780470756959

[89] Jiangnan Xu, Konstantinos Papangelis, John Dunham, Cati Boulanger, Jin Ha Lee, Nicolas Lalone, and Michael Saker. 2023. Understanding Social Interactions in Location-based Games as Hybrid Spaces: Coordination and Collaboration in Raiding in Pokémon GO. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3544548.3581544

[90] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, et al. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3491102.3517491

[91] Lixiu Yu, Aniket Kittur, and Robert E Kraut. 2014. Searching for Analogical Ideas with Crowds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1225–1234. https://doi.org/10.1145/2556288.2557378

[92] Kexin Zhang, Brianna R Cochran, Ruijia Chen, Lance Hartung, Bryce Sprecher, Ross Tredinnick, Kevin Ponto, Suman Banerjee, and Yuhang Zhao. 2024. Exploring the Design Space of Optical See-through AR Head-Mounted Displays to Support First Responders in the Field. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3613904.3642195

[93] Yuhang Zhao, Elizabeth Kupferstein, Brenda Veronica Castro, Steven Feiner, and Shiri Azenkot. 2019. Designing AR Visualizations to Facilitate Stair Navigation for People with Low Vision. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 387–402. https://doi.org/10.1145/3332165.3347906

[94] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23. https://doi.org/10.1145/3274463

A DESIGN OF INITIAL CUSTOMIZED TRANSLATIONAL RESOURCE

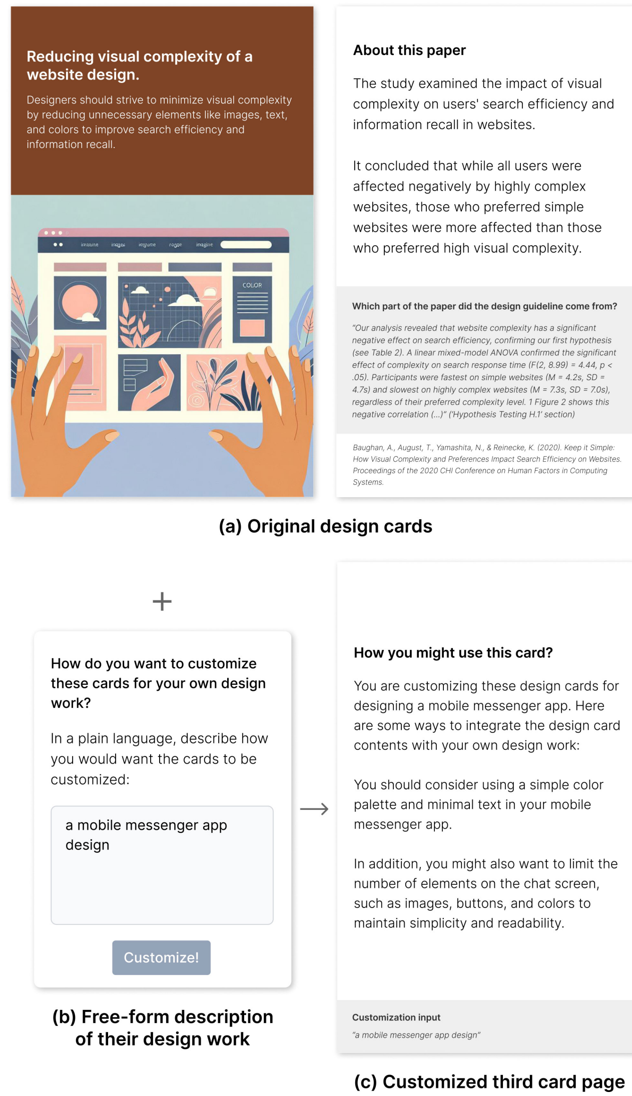


Figure 4: Pipeline for generating a customized card page for our preliminary study. After viewing (a) original design cards, each participant described their design work in (b) free-form text to (c) generate a customized third card page.

B PROMPT EXAMPLE

The following is an example of a prompt used for generating the *Understanding users* component. Our pipeline similarly generates other components using the following structure with a distinct set of inputs and output requirements, as detailed in Table 1 and Section 4.2.

System instruction

```
You will be provided (i) an academic paper and (ii) a
designer's design context, and you are a helpful assistant
who returns potential user needs or pain points that the
paper's authors identified, which may also help the designer
understand their design target.

[Rules]
  1. The output must be an array of JSONs, each of which
     contains the following entities:
     - "title": a concise title referring to the user
       need/pain point that the paper authors pointed out in
       the paper based on their target user, which may also
       be applicable or relevant to the designer's design
       target
     - "detail": an elaboration of how the need/pain point
       is relevant to the designer's design target
     - "source_detail": a definition of the user need/pain
       point as provided by the authors (if any), and why
       the paper authors emphasized this need/pain point
     - "source_paragraph": the original section paragraph in
       the paper where the user need/pain point is discussed
     - "source_section_title": the title of the section where
       the need/pain point was described in the paper content
  2. When returning needs/pain points from the
     paper contents, select those that may also be
     relevant/applicable to the designer's design target
     ("who", "what") in relevance order.
     - "who": the target audience for their design
     - "what": the goal and design space for their design,
       including the modality and domain for their design,
       target user's pain points they are already focusing on,
       client/company they are designing for, and/or metric
       that they would like to focus on
  3. Return up to two needs/pain points.
     - If there are not enough suitable needs/pain points to
       return, you may return less than two.

[Output format]
[
  {
    "title": String // keyword-based,
    "detail": String, // up to two lines
    "source_detail": String // up to two lines
    "source_paragraph": String,
    "source_section_title": String, // titlecased
  }, ...
]
```

User prompt

```
[Paper contents]
{{ original full paper text (in XML-parsed format generated
from GROBID) }}

[Designer's design target]
{{ designer's input }}
```

C EXAMPLES OF CUSTOMIZED DESIGN CARDS

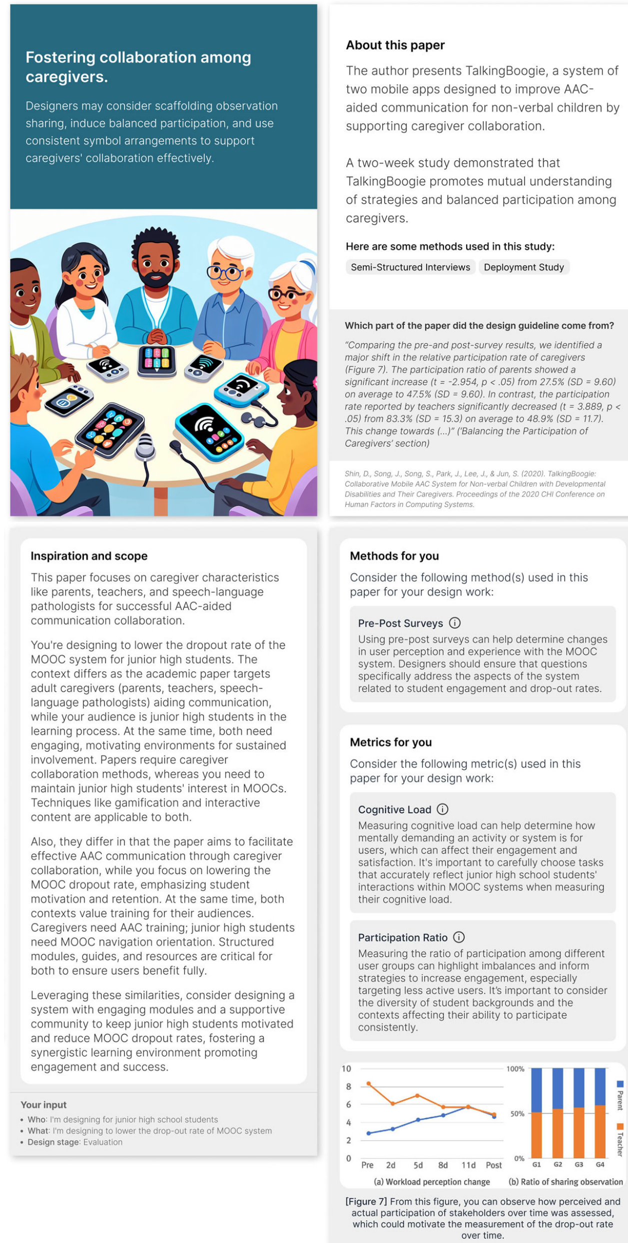


Figure 5: Customized design cards built on Shin *et al.* [68], with the following inputs – Who: *I'm designing for junior high school students*, What: *I'm designing to lower the drop-out rate of MOOC system*, Design stage: *Evaluation*

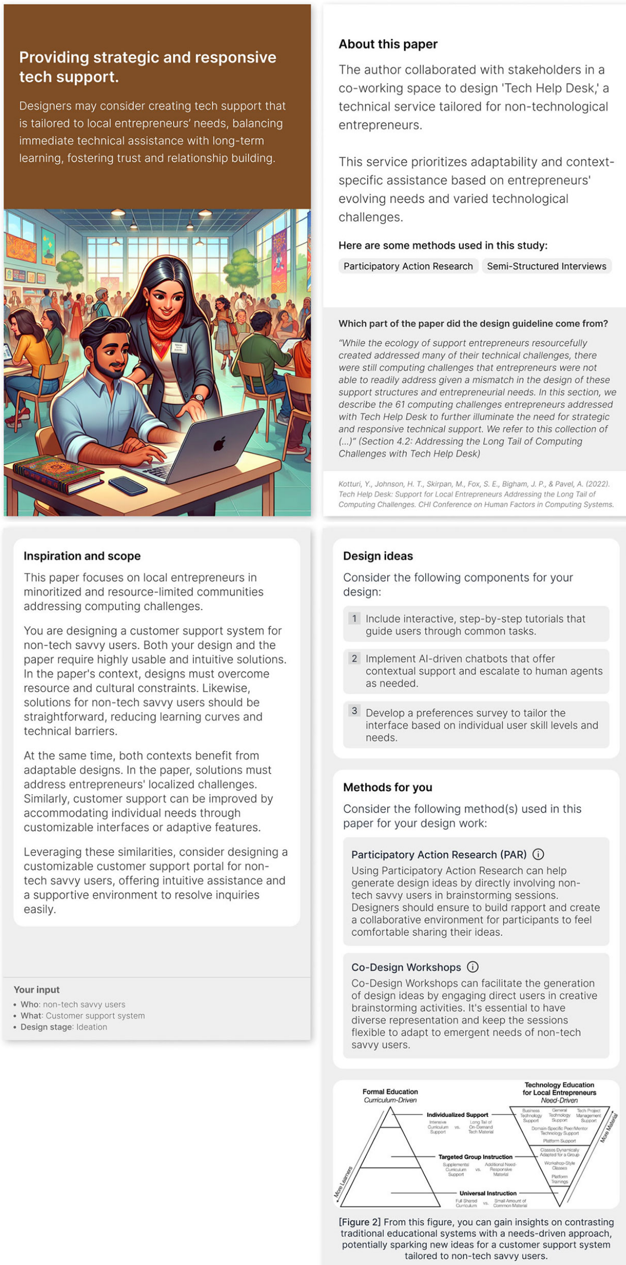


Figure 6: Customized design cards built on Kotturi *et al.* [40], with the following inputs – Who: *Non-tech savvy users*, What: *Customer support system*, Design stage: *Ideation*