ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240944

A Language Model as a Design Assistant for UI Design Recommendation and Evaluation

Lin Sheng[a], Cheng Deng[a], Junjie Zhang[b], Fangyuan Chang[a], Qinghua Sun[a], Hongyu Liu[a] and Zhenyu Gu[a,*]

[a]Shanghai Jiao Tong University, Shanghai, China
[b]Huawei Technologies CO.LTD, Shenzhen, China
ORCID (Lin Sheng): https://orcid.org/0000-0002-2357-2330, ORCID (Fangyuan Chang): https://orcid.org/0000-0003-2069-0507, ORCID (Qinghua Sun): https://orcid.org/0000-0002-8913-4608, ORCID (Hongyu Liu): https://orcid.org/0009-0002-6718-7976, ORCID (Zhenyu Gu): https://orcid.org/0000-0003-3921-5837

Abstract. In the digital era, the significance of design education is on the rise due to its ability to cultivate creativity. However, the disconnect between design practice and theory, coupled with the abundance of design knowledge, poses challenges to learning in this field. Despite the potential of large language models (LLMs) to integrate various data sources for facilitating design knowledge dissemination, they face obstacles such as the scarcity of design-related datasets and limited natural language representations. To overcome these challenges, we introduce DRELM, a design-centric language model that serves as an assistant providing UI design recommendations. We also offer corresponding resources to advance language modeling research in the design domain. Importantly, DesignInstruct stands out as a premier dataset for guiding user interface tasks, while DesignEvaluation significantly contributes to autonomous design evaluation and decision support. In our research, we utilize supervised data from DesignInstruct and DesignEvaluation to fine-tune pre-trained Qwen-7B models for design tasks. Experiments conducted on test data affirm the effectiveness of our dataset in enhancing knowledge comprehension, design execution, and evaluation. We commit to making all training data and DRELM models at https://github.com/sssala/DRELM-A-Language-Model-for-Design-Recommendation-and-Evaluation.

1 Introduction

In the digital age, the significance of digital design education is steadily on the rise. This surge is attributed to the fact that design education not only imparts specialized design skills but also nurtures a mindset and creativity essential for effectively applying innovative thinking across various fields [19]. However, the complexity of digital design extends beyond practical applications and academic research [25], resulting in the fragmentation and diversification of design knowledge. Moreover, the predominance of experiential learning in design has caused a gap between theory and practice. Consequently, individuals are compelled to invest a substantial amount of time in gathering and organizing multidisciplinary knowledge, often at the expense of practical experience. This study challenge becomes particularly daunting in situations where design resources are limited.

* Corresponding Author. Email: zygu@sjtu.edu.cn.

A design assistant that offers both instructional tutorials and evaluation feedback for normal users addresses the challenges in design education effectively. Yet, integrating cross-domain data for complex design tasks presents a formidable challenge due to variations in terminology, formats, and data structures [13]. These tasks include design search, layout, and code generation for user interfaces (UI) [9], visio-linguistic representations of UI screens [2], floor plan generation [21], and inferences about UI functionality and semantics [37]. The emergence of powerful large-scale language models, such as ChatGPT [28] and GPT-4 [14], has revolutionized Natural Language Processing (NLP), offering groundbreaking solutions for overcoming design challenges and delivering user-friendly interactive interfaces. Compared to traditional machine learning (ML) approaches based on design facts and pre-coded rules, it is capable of handling large amounts of cross-domain design data and generating creative solutions that go beyond predefined rules.

While Current Large Language Models (LLMs) excel in general domains [38][18], their applicability to the design domain remains uncertain. Limitations stem from pragmatic knowledge gaps in design problems, scarcity of exemplary design instances, challenges in translating design forms into natural language, and the subjective nature of judgment criteria linked to requirements [22]. Importantly, design-related data is sparse in widely used pre-trained corpora [39]. Translating macro, discrete, and subtle design requirements into specific or complex design tasks presents a significant challenge. Additionally, closed-source top LLMs like ChatGPT hinder research and progress in external domains. To address these challenges and advance research and applications in the design domain, we propose a language model as a design assistant named DRELM. DRELM, a language model with 7 billion parameters, builds upon the pre-trained Qwen-7b model [36], specifically tailored for the intricacies of the design domain. Along with DRELM, this paper introduces the creation of supervised data for design and evaluation tasks.

We conducted instruction tuning [10][7] to enhance DRELM's ability to follow design instructions, creating two instruction-tuning datasets named DesignInstruct and DesignEvaluation. DesignInstruct integrates examples from nine distinct design tasks, covering areas such as color recommendation, font suggestions, animation design, symbol design, layout design, web component design, and general web design. DesignEvaluation is a dataset for evaluating design

tasks, aiding users in making design decisions. It includes single-value evaluation and perceptual evaluation, compiled by crawling and processing a collection of web cases for design task evaluation and manual labeling. Finally, we invited 10 designers to assess 600 questions, including 300 subjective questions, 100 objective questions, 100 design tasks, and 100 assessment tasks using different metrics to evaluate the usability of DRELM.

Our contributions are summarized as follows:

• Introducing DRELM, a language model in the design field excelling in answering design questions, executing specific design tasks, and evaluating design work as a proficient design assistant.

• Constructing DesignInstruct, a dataset for supervised design instructions. And DesignEvaluation, a dataset for design task evaluation, which potentially serves as a reward model to enhance DRELM.

• Comparing DRELM with similar-sized baseline models for design assignments. And evaluating the design model against normal users' evaluations to verify its usability.

The remainder of the paper is organized as follows: Section 2 discusses related work on LLM for Design, Section 3 outlines the details of data collection and supervised instruction data construction, Section 4 provides insights into parameter-efficient instruction tuning processes, Section 5 evaluates DRELM, and finally, Section 6 explores the application of DRELM as a design assistant.

2 Related Work

Foundation Language Models: Since the introduction of ChatGPT, we've witnessed the development of several related models like GPT-4[14], and models including open-source ones like internLM[34], Qwen[1], ChatGLM[12], Baichuan [41], CodeGen [27], LLaMA[36], LLaMA2[36], BLOOM, and various instruction-tuned LLMs such as Alpaca[33], Vicuna[6], Dolly[3], Guanaco[11], Wizardmath[23]. The evolution of these Large Language Models has significantly contributed to the rapid emergence of knowledge models across diverse domains.

Domain Language Models: A multitude of models tailored to specific domains have emerged to address domain-specific challenges. For instance, K2 [10] offers research assistance and knowledge reasoning in geoscience, while Chatlaw [8] specializes in legal counseling within the field of law. In the Chinese financial domain, xuanyuan [43] provides accurate and contextually appropriate responses. Life science benefits from models like Med-PaLm [31], PaLM 2 [32], BioGPT [24], focusing on pharmaceutical design and disease diagnosis. Various models from other disciplines, such as Code llama [29], VeriGen [35], have been developed. In the design field, some LLMs like I-Design are for interior designers [4], and LLMs enable generative collaborative design in a mixed reality environment [40]. However, the UI design domain is notably absent.

Parameter Efficient Fine-tuning: The challenge in fine-tuning Large Language Models (LLMs) for design arises from the intricate interplay of knowledge and the manipulation of structured data types, which designers need to tune. This necessitates efficient parameter tuning on LLMs. Random approaches, such as Random and Mixout[20] models, operate independently of task-specific data. Rule-based strategies, including BitFit [42], MagPruning, Adapter[16], and LoRA[17], Qlora[11], etc, involve leveraging prior knowledge to identify crucial features, mitigating the limitations of random approaches. Projection-based techniques, exemplified by DiffPruning[30], ChildPruning[15], were developed to aid in selecting the models tunable parameters.

3 Data Collection and Curation

To train DRELM, we collected a diverse corpus of design texts and design-oriented data from various sources, organizing them into two functional datasets: DesignInstruct and DesignEvaluation (refer to Figure 1). Subsequently, we structured this data into signals, as detailed in [10], providing valuable insights into design-related knowledge, requirements, and task forms. These signals were then reorganized into design instructions.

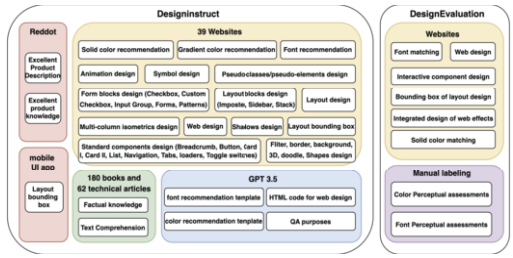


Figure 1. Instruction tuning data.

3.1 Instruction Tuning Data for design task

We compiled crucial instruction-tuning data to customize pre-trained models for non-professional "designers." Using a semi-manual pipeline, we created a dataset named DesignInstruct, tailored to diverse design tasks, including color recommendation, font suggestion, animation design, symbol design, layout design, web component design, and general web design. This dataset was closely aligned with the expertise of design professionals. To ensure domain-specific accuracy, we involved designers in reviewing knowledge-intensive data. Following the principles of reconstructive pretraining, signals were selectively gathered from specialized design websites, datasets, and books. These signals were then structured into <instruction, input, output> pairs, forming the foundation for instruction-tuned samples. The databases considered are listed in Figure 1.

Professional Design Websites: We gathered valuable data from 39 renowned web/UI design websites, capturing themes, tags, and functional elements such as hexadecimal color codes, font names, CSS, and HTML codes. These websites cover diverse aspects including Color Matching, Background Design, Font Matching, Layout Design, Kinetic Design, Symbol Design, Interactive Component Design, Web Design, and Integrated Design of Web Effects. Refer to Table 1 for the numbered website addresses, which are presented for statistical convenience, along with the corresponding count indicating the quantity of cleaned data (38,047 pairs).

Design Question-and-Answer (QAs): We compiled text from 180 UI design books and 62 technical articles on components. Using OpenAI [28] for template generation and input from designers, we created a concise and precise dataset comprising 36,810 pairs of UI specification Q&As.

Self-instruct: With Alpaca-GPT4 [10], we employed GPT3.5 to generate a dataset of 51,000 design-based questions and answers(17.81MB) for design explanation. Additionally, GPT3.5 was used to create descriptions for web pages, including font recommendation templates (10,574), color recommendation templates (7,973), and HTML code implementation (21,390). All data underwent verification by designers after a thorough cleaning.

Reddot: This dataset contains information about products that have won the Red Dot Design Award, an esteemed international

Table 1. 39 Professional design website list.

Num	**Color Matching**	Count
A1	http://zhongguose.com/	526
A2	https://www.colorsandfonts.com/color-gradients	380
A3	https://colordrop.io/	787
A4	https://colorhunt.co/	2680
A5	https://mybrandnewlogo.com/color-palette-generator	9600
A6	https://www.happyhues.co/	85
A7	https://uigradients.com/	334
A8	https://www.materialpalette.com/	19
A9	http://gradientsguru.com/strong-gradients	170
A10	https://webgradients.com/	115
A11	https://nipponcolors.com/	250
	Background Design	
B	https://www.magicpattern.design/tools/css-backgrounds	22
	Symbol Design	
C1	https://icons.getbootstrap.com/	2050
C2	https://iconsvg.xyz/	57
C3	https://www.toptal.com/designers/htmlarrows/	96
	Layout Design	14
D1	https://layout.bradwoods.io	484
D2	https://js.design/	1952
D3	https://www.mockplus.cn/	4
D4	https://cloud.protopie.io/	14
D5	https://css-tricks.com/snippets/css/complete-guide-grid/introduction	45
	Animation Design	
E	https://www.transition.style	50
	Font Matching	
F1	https://www.5ifont.cn/font	6048
F2	https://www.googlefonts.cn/	997
	Interactive Component Design	
G1	https://navnav.co/	1531
G2	https://bem-cheat-sheet.9elements.com/	28
G3	https://10015.io/tools/css-loader-generator	330
G4	https://uiverse.io/all	1757
G5	https://vant-contrib.gitee.io/vant/zh-CN/button	581
	Web Design	
H1	https://codemyui.com/	1569
H2	https://codepen.io/	4205
H3	https://htmlpage.cn/builder/	31
H4	https://www.30secondsofcode.org/css	140
	Integrated Design of Web Effects	
I1	https://csscoco.com/inspiration/	188
I2	https://html-css-js.com/css/generator/box-shadow	84
I3	https://web.dev/patterns?hl=zh-cn	125
I4	https://lhammer.cn/You-need-to-know-css/	45
I5	https://www.w3schools.com/howto/	70
I6	https://c.runoob.com/examples/	499
I7	https://www.kaggle.com/datasets/olgabelitskaya/html-recipes	85
Total		38047

design competition. We extracted product, category, and description details to reconstruct DesignInstruct. This dataset is available at https://huggingface.co/datasets/xiyuez/red-dot-design-award-product-description (21,183 entries).

Mobile UI app: This dataset includes images and object detection boxes with class labels and location information in mobile UI designs. We extracted width, height, and bound-ing box details of objects to reconstruct DesignInstruct. This dataset is accessible at https://huggingface.co/datasets/mrtoy/mobile-ui-design (7,846 entries).

For a comprehensive understanding of design signals, we outline key signals below and detail the restructuring process:

Q1: Factual Knowledge. This includes design facts, such as the functionality of buttons and descriptions of products awarded by Reddot. These were sourced from design-related QAs platforms and are valuable for QAs and fact verification.

Q2: Text Comprehension. Found in design application platforms and textual content featuring question-and-answer pairs, this signal aids in facilitating QAs processes.

Q3: Design Requirements and Tips. Encompassing theme descriptions, requirements, and user favorites, these inputs were sourced from design websites and designers' labels. They are evident in the requirements described in the Self-instruct for laymen's design application.

Q4: Font/Color Name. Font signals, denoted as F1 for Chinese characters and F2 for English characters, were sourced from Font Matching websites. Similarly, Color signals, represented by A1 for Chinese color and A11 for Japanese color, were sourced from color-matching websites. These signals are utilized in the Self-instruct for font and color recommendations.

Q5: Color Hexadecimal Coding. Color codes serving various design needs were sourced from color-matching websites. These include solid color matches (A3-6, A8, and Self-instruct for color recommendation), gradient color matches (A2, A7, A9-10), and color name translation (A1, A11).

Q6: Component/Block Layout Bounding Box. Found in various websites (D2-D4) and the Mobile UI dataset, this signal involves the extraction of bounding box coordinates for components or blocks: [top-left point coordinate x, top-left point coordinate y, top-right point coordinate x, top-left point coordinate y].

Q7: SVG Encoding. SVG, used for image encoding, is found in C1, C2, I1, I4, and I5.

Q8: CSS Encoding. CSS, used for page feature encoding, is found in websites on B, H1-4, E, C1-3, D1, D5, G1-5, and I1-7.

Q9: HTML Encoding. This signal, used for page encoding, is found in C3, G1, D1, H1-4, I1, I3, I5, and I6, as well as in the Self-instruct for HTML code.

To effectively utilize these signals, we structure the data into <instruction, input, output> pairs, tailored for tasks such as component interpretation, question answering, color and font suggestions, animation, symbol, layout, and web component design, as well as other web effects and web design. This formatting aligns with the preferences of non-professionals, emphasizing DRELM as a tool focused on meeting user requirements. To enhance transparency, we plan to release all scripts as open source in the final version.

Component Explanation: To train the model in word explanation skills, we digitized 62 technical text components, extracting all words and their explanations (Signal Q1). Additionally, we augmented the dataset with related entries from design books (Signal Q2).

Question Answering: We restructured two types of supervised data as follows:

• Curating question-answer pairs by separating product names (input) and descriptions (output) in the Reddot dataset.

• Generating and answering questions with ChatGPT, validated by designers to ensure accuracy (Signal Q2).

Color Recommendation: We restructured two types of supervised data to address solid and gradient color matching recommendations, and translation of Chinese and Japanese color names:

• Self-instructing ChatGPT 3.5 on color matching, utilizing page features as input and generating page component color matching as output (Signal Q5).

• Extracting data from 11 color-matching websites to provide input on color names (Signal Q4) or usage definitions (Signal Q3), and output hexadecimal coding (Signal Q5).

Font Recommendations: We restructured two types of supervised data for font recommendations:

• Self-instructing ChatGPT 3.5 on font pairings, utilizing page features as input and generating page component font pairings as output (Signal Q4).

• Extracting data from specialized font recommendation websites to provide font names (Signal Q4) as input and definitions (Signal Q3) of font usage as output.

UI Animation Design: Gathering animation names and effects (Signal Q3) from CSS or HTML code from E, G1, H4, I1, I4, and

I5, using them as input to generate corresponding CSS/HTML code (Signals Q8, Q9) as output.

Symbol Design: Extracting SVG (Signal Q7) and HTML code (Signal Q8) from F3, using tips and image features as input to generate SVG/HTML code as output.

Layout Design: Structuring supervised data for layout design, using layout names or themes as input to generate CSS code (Signal Q8) from D1, D5, I1, and I3-5 or bounding box information (Signal Q6) as output.

Web Component Design: Web components encompass standard blocks (Breadcrumb, Button, Card I, Card II, List, Navigation, Tabs, Loaders, Toggle Switches), form blocks (Checkbox, Custom Checkbox, Input Group, Forms, Patterns), and layout blocks (Imposter, Sidebar, Stack). Extracting names and effects of web components from G1-5, I1, and I4-5 as input, generating CSS or HTML code as output (Signals Q3, Q8, Q9).

Other Web Effect Design: Web effects encompass the implementation of multi-column isometrics, shadows (box-shadow, drop-shadow), the use of pseudo-classes/pseudo-elements, filters, borders, backgrounds, 3D effects, doodles, shapes, etc. Extracting these elements from I1-7 as input to generate CSS (Signal Q8) or HTML code (Signal Q9) as output.

Web Design: Extracting case names and themes from the H1-4 as input, generating CSS (Signal Q8) and HTML code (Signal Q9) as output.

After completing the procedure mentioned above, we obtained an extensive dataset. Prioritizing quality over quantity, we sampled and cleaned the data to create the DesignInstruct instruction tuning dataset presented in Table 2.

Table 2. DesignInstruct lists.

Tasks	Records	Total(cleaned)
Component explanation	50,000	8,900
Question Answering	100,000	49,093
Alpaca_ch		51,000
Color Recommendation	24,000	22,919
Font recommendations	32,011	17,619
Animation design	1,020	615
Symbol design	2,300	2,203
Layout design	3,100	10,345
Web component design	1,3231	3,815
Other Web effect design	1,222	1,246
Web design	2,8300	26,721
Total		194,476

3.2 Instruction Tuning Data for Design Evaluation

To assess design learning and empower non-professionals in independent design decision, we curated a robust evaluation dataset called DesignEvaluation (refer Table 3). This dataset amalgamates website examples and sentiment assessments, covering a spectrum of design tasks including font and color matching, interactive component design, layout bounding boxes, web design, and web effects integration. To mitigate data collection variations and outliers, we applied log transformations to the evaluation set. Each dataset entry follows a structured format <instruction, input, output>, with input representing the design outcome and output indicating evaluation quality.

Additionally, we solicited perceptual evaluations from designers to refine design elements. Ten designers collaborated to define emotional categories and manually evaluate color composition and typeface. Each assessment was scored on a scale of 0 to 1. This collaborative endeavor culminated in a comprehensive manual annotation comprising 242,040 items.

Table 3. DesignEvaluation lists.

Evaluation type	Count(cleaned)
Font matching	997
Solid color matching	3,467
Interactive component design	1,757
Bounding box of layout design	484
Web design	4,532
Integrated design of web effects	188
Perceptual assessments	242,040
All	253,465

4 Training

4.1 Design Domain Adaptation Recipe

Given the creative and subjective nature of UI design, there's a notable absence of language models tailored to this field. However, leveraging advanced natural language models and tools can significantly aid users in knowledge discovery, improve design efficiency, and foster creativity in their work. Hence, the acquisition of a language model capable of understanding design concepts, generating UI, and executing color/font commands is crucial. The design domain offers abundant resources, including specifications, papers, web page libraries, and specialized design websites, providing a solid foundation for training large-scale language models. Leveraging this wealth of design-centric data, we developed DRELM, utilizing instruction tuning to enhance Qwen-7b's ability to generate design content based on user instructions. This approach allows the model to extract additional insights from the refined domain knowledge [10].

4.2 Instruction Tuning

To ensure compliance with design instructions, we implemented multitask training, incorporating knowledge-intensive instructions like DesignInstruct and DesignEvaluation. In the instruction learning phase, we introduced parameter-efficient fine-tuning (PEFT) to facilitate training in a low-resource setting. Following LoRA [17], with a hidden layer $h = W_0x$, a pre-trained weight matrix $W_0 \in R^{d\times k}$, where $B \in R^{d\times r}$, $A \in R^{r\times k}$, and the rank $r \ll \min(d, k)$, α is a constant in r. When optimizing with Adam, adjusting α is akin to tuning the learning rate. The modified forward pass is as follows:

$$h = W_0x + \alpha/rBAx \tag{1}$$

In the instruction tuning of Qwen-7b, our objective is to align the model with expert knowledge, achieved through Low-Rank Adaptation (LoRA). Specifically, we employ a learning rate of 1e-5 with a global batch size of 128. In the LoRA configuration, $lora\_rank$ is set to 8, while $lora\_alpha$ is set to 32. Based on our experimental observations, we designate $lora\_target\_modules$ as k, q, v. LoRA-based instruction tuning trains only 6M parameters on two A30 GPUs over 120 hours.

5 Evaluation and Results

This section is about evaluating two types of tasks based on two datasets: design recommendation and design assessment. Design recommendation encompassed design explanation tasks (e.g., Q&As) and design tasks. Design assessment included both single numerical evaluations and perceptual evaluations (perceptual words and degrees). Each type had distinct evaluation metrics.

5.1 Evaluation of Design recommend task

To evaluate the effectiveness of language models in providing design explanations and addressing design-related queries, we systematically gathered data from various open-source knowledge platforms for our QAs tasks, including websites like https://wenku.baidu.com/, https://blog.csdn.net/, https://max.book118.com, and https://www.docin.com/. This meticulous process resulted in a dataset comprising 300 subjective questions and 100 objective questions, focusing on topics most relevant to designers, including UI design, user research, and design psychology. We then conducted a human evaluation involving 100 distinct design tasks, such as color and font recommendations, web component generation, and web page creation, along with 100 design evaluations from the retained training set. Given the inherent creativity and subjectivity in design tasks, this evaluation was essential [5]. Following the methodology outlined in [26], we engaged 10 professional designers to perform a comparative analysis of Qwen-7B[1], Baichuan-7B[41], ChatGLM-6B[12], internLM-7B[34], and DRELM. Moreover, to assess the model's ease of use and usability, we also engaged 10 normal users to evaluate DRELM, rated on a scale of 1 to 7.

The evaluation criteria for the design explanations and design tasks differ in the correctness metric because design knowledge can have objective accuracy, whereas design tasks do not. Therefore, correctness, consistency (uniformity across different instances), and rationality (logical soundness and reasoning) were used for evaluating explanation tasks [10]. Similarly, design tasks were assessed based on satisfaction, consistency (uniformity across different instances), and rationality (including adherence to design principles, user preferences, and task context), using a rating scale from 1 (poor) to 3 (good). Table 4 presents the results of professional designers, demonstrating DRELM's strong performance in QAs task consistency and satisfaction and consistency for design tasks. It remains competitive in terms of rationality. Figure 2(a) shows that normal users think the model is usable, but the ease of use should be improved in the image output rather than the text. These findings underscore our model's enhanced ability to understand and fulfill design tasks.

Table 4. Human evaluation.

Baselines	QAs			Design tasks		
	correctness	consistency	rationality	satisfaction	consistency	rationality
Qwen7B	2.1	2.2	**2.3**	2.1	2.3	**2.3**
internLM7B	1.3	1.6	1.4	1.1	1.2	1.2
Baichuan7B	1.5	1.6	1.7	1.8	1.9	2.1
ChatGLM6B	**2.2**	2.3	2.2	2.2	2.0	2.2
DRELM	2.1	**2.4**	2.1	**2.3**	**2.4**	2.1

5.2 Evaluation of Design evaluation task

The output of design assessment includes two formatsingle numerical for all design task and perceptual evaluation (perceptual words + degrees) for font and color recommend. Single scoring employs Euclidean distances, while sensory scoring encompasses two aspects: sensory vocabulary accuracy, determined by the ratio of correctly generated sensory words (0/1) to the total number of senses generated, and sensory rating accuracy, assessed by comparing the generated rating with the test rating using Euclidean distance. As depicted in Figure 2(a), the mean error for single ratings was 0.047. However, notable bias was observed in web functional design execution, potentially attributed to the presence of icons in the training data. In Figure 2(b), the mean error of perceptual rating values was 0.179. And both dimensions are sufficiently accurate when 1-2 perceptual ratings are generated at the same time, with a large bias when higher than three.

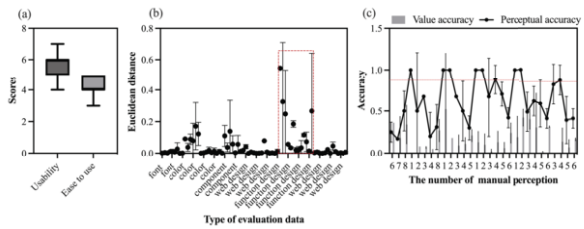


Figure 2. (a) Normal users' evaluation of DRELM.(b) Accuracy of design task evaluation with single ratings.(c) Accuracy of design evaluation with varying numbers of perceptual ratings conducted simultaneously.

6 Application

This section shows our exploration regarding the potential applications of DRELM, focusing on its use cases in color recommendation, font recommendation, component generation, and design evaluation in Table 5. All prompts begin with "As an experienced designer".

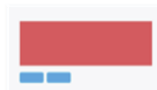
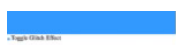
7 Conclusion and future work

This study introduces DRELM, a robust language model specialized in the field of UI design for normal users. DRELM excels in responding to design queries and executing layman's instructions, rendering it a valuable design assistant. We curated two datasets, namely DesignInstruct for design tasks and DesignEvaluation for evaluating design decisions, and utilized them to fine-tune training on Qwen-7B, resulting in the development of DRELM. DRELM showcased its proficiency and efficacy, surpassing other models of similar size. We illustrate DRELM's versatility through applications such as color recommendation, font recommendation, and component and web effect generation. Furthermore, we will make all of data and model at https://github.com/sssala/DRELM-A-Language-Model-for-Design-Recommendation-and-Evaluation. There are still many issues to be addressed in future work. For instance, in terms of usability, improvements in generating user-friendly visual designs or conducting more user tests should be considered. Regarding model capabilities, creating more complex pages or editable prototype pages for professional designers is still constrained by layout limitations and requires integration with other technologies.

Ethical Statement

There are no ethical issues.

Table 5. case study lists.

Functionality	Prompt	Output	Visualization
Solid color matching	We invite your recommendations for a solid color scheme in hex code that aligns with the page theme and its associated tags: Sunset Dream, with tags such as Bright, Bold, Dark, Light, Orange, and Pink.	#FFCC00, #0B3059, #334257, #F5F6F9	
Gradient color matching	We invite your recommendations for a gradient color scheme in hex code specifically aligned with the page theme 'Pinot Noir.	#833466 → #774087	
Component-specific solid color matching	Please help me design the color scheme of the areas and components of the pages of this website based on the description of the website I have provided, and the colors are displayed in hex code. This different business website is an online sales platform for fashion cosmetics. The site offers various brands of cosmetics and beauty products for consumers to purchase. In addition to selling products, the site also provides the latest fashion and beauty information, makeup tutorials, and esthetician consultations, and is dedicated to giving users an all-around beauty experience. The site's style is fashionable, sophisticated and modern, focusing on showing elegance and high quality.	{ Main Color #FF5A5F, // The main color used, using bright red to highlight the site's brand image and attract the user's attention Background Color #F8F9FA, // The background color of the page is light grey, to create a relaxing and comfortable shopping environment Button Color #00AEEF // Button colors Choose bright blue color to increase interactivity and add vitality to the overall page design}	
Color name lookup	Kindly provide the hexadecimal codes for the following color: Don Calamus Red.	#ea2c42	
Font recommendation	Please recommend fonts based on the provided usage scenarios, types, and styles. The usage scenarios include e-commerce, advertising, gaming, video and film, poster promotion, print and packaging, layout reading, and media & H5. The types are brush calligraphy, and the style is Chinese, powerful, handsome, and unrestrained.	Han Yi Shangwei's handwritten note	
Component fonts recommendation	Please help design the typography for a website based on the provided description. This is a business website for wedding planning, specifically catering to couples seeking wedding planning and services. The website focuses on a theme of simplicity and romance, emphasizing a happy and warm atmosphere.	Navigation bar font: Song font, Page title font: Microsoft YaHei, Body text font : Regular script, Button font: Chinese running script	
Simple website	please design the web page and present it in html code according to the following web requirement. The web requirement is that the web page is an online bookstore web page where users can browse books, search for books, view book details and add books to the shopping cart. The style of the web page should be simple and clear, focusing on the presentation of the books and the user experience.	<!DOCTYPE html> \n <html> \n <head> \n <meta charset='UTF-8'> \n <!– Page title and description –> \n <title> Online Bookstore</title>\n </head> \n <body> \n <!– navigation bar –>\n <header>\n Online Book Store \n <nav>\n \n < a href='#'> home page \n <form action='#' method='GET'>\n <input type='text' name='q' placeholder='search'>\n <button type=submit> search</button>\n </form>\n \n Shopping cart contains: \n <!– Shopping Cart Contents –>\n \n \n </nav> \n </header>\n </body>\n</html>	
Effect generation	Please design the component and render it in html code according to the name of the following interactive component. Glitch effect in LESS	<style> \n #glitch {\n position: relative;\n padding: 20px 400%;\n background-color: #3399FF;\n }\n .glitched {\n display: inline-block;\n position:absolute;\n}\n @import url (https://fonts.googleapis.com/css?family=Roboto:30,4,70); \n</style> \n< div id=glITCH> \n GLITCH \n \n<input type=checkbox checked=checked /> \n <label for=toggle class=container style=font-size:2em> Toggle Glitch Effect</label>\n< label id='toggle' for='glitter'></label>	
Font perception	You have been asked to determine 1 emotional color for the following font and give it a score between 0 and 1. The closer to 1, the more adequate the emotional color. Aju Bubble Body	{'robust': '0.4'}	
Solid color matching evaluation	Please rate the following solid color combinations on an aesthetic scale of 0-1, with the higher score proving the more popular the solid color combination. #3b5998,#FFFFFF,#000000,#FFC107	0.7093	

Acknowledgements

This study was supported by Huawei Technologies CO.LTD.

References

[1] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[2] P. Banerjee, S. Mahajan, K. Arora, C. Baral, and O. Riva. Lexi: Self-supervised learning of the ui language. *arXiv preprint arXiv:2301.10165*, 2023.

[3] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. OBrien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[4] A. Çelen, G. Han, K. Schindler, L. Van Gool, I. Armeni, A. Obukhov, and X. Wang. I-design: Personalized llm interior designer. *arXiv preprint arXiv:2404.02838*, 2024.

[5] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[6] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[7] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models, 2022.

[8] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023.

[9] B. Deka, Z. Huang, C. Franzen, J. Hibschman, D. Afergan, Y. Li, J. Nichols, and R. Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 845–854, 2017.

[10] C. Deng, T. Zhang, Z. He, Y. Xu, Q. Chen, Y. Shi, L. Fu, W. Zhang, X. Wang, C. Zhou, Z. Lin, and J. He. K2: A foundation language model for geoscience knowledge understanding and utilization, 2023.

[11] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

[13] P. Duan, J. Warner, and B. Hartmann. Towards generating ui design feedback with llms. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–3, 2023.

[14] J. Fu, L. Lin, X. Gao, P. Liu, Z. Chen, Z. Yang, S. Zhang, X. Zheng, Y. Li, Y. Liu, et al. Kwaiyiimath: Technical report. *arXiv preprint arXiv:2310.07488*, 2023.

[15] Z. Fu, H. Yang, A. M.-C. So, W. Lam, L. Bing, and N. Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12799–12807, 2023.

[16] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[18] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen. Time-llm: Time series forecasting by reprogramming large language models, 2024.

[19] B. Kye, N. Han, E. Kim, Y. Park, and S. Jo. Educational applications of metaverse: possibilities and limitations. *Journal of educational evaluation for health professions*, 18, 2021.

[20] C. Lee, K. Cho, and W. Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*, 2019.

[21] S. Leng, Y. Zhou, M. H. Dupty, W. S. Lee, S. C. Joyce, and W. Lu. Tell2design: A dataset for language-guided floor plan generation. *arXiv preprint arXiv:2311.15941*, 2023.

[22] D. Lockton, D. Harrison, and N. Stanton. The design with intent method: A design tool for influencing user behaviour. *Applied ergonomics*, 41:382–92, 10 2009. doi: 10.1016/j.apergo.2009.09.001.

[23] H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, and D. Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.

[24] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

[25] M. W. Meyer and D. Norman. Changing design education for the 21st century. *She Ji: The Journal of Design, Economics, and Innovation*, 6 (1):13–49, 2020.

[26] S. Narayan, S. B. Cohen, and M. Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

[27] E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, and C. Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

[28] L. OpenAI. Introducing chatgpt. 2022.

[29] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[30] Y. Shulman. Diffprune: Neural network pruning with deterministic approximate binary gates and l_0 regularization. *arXiv preprint arXiv:2012.03653*, 2020.

[31] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[32] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[33] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[34] I. Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

[35] S. Thakur, B. Ahmad, H. Pearce, B. Tan, B. Dolan-Gavitt, R. Karri, and S. Garg. Verigen: A large language model for verilog code generation. *ACM Transactions on Design Automation of Electronic Systems*, 2023.

[36] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[37] J. Wu, S. Wang, S. Shen, Y.-H. Peng, J. Nichols, and J. P. Bigham. Webui: A dataset for enhancing visual ui understanding with web semantics. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2023.

[38] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. Next-gpt: Any-to-any multimodal llm, 2023.

[39] L. Xu, X. Zhang, and Q. Dong. Cluecorpus2020: A large-scale chinese corpus for pre-training language model, 2020.

[40] S. Xu, Y. Wei, P. Zheng, J. Zhang, and C. Yu. Llm enabled generative collaborative design in a mixed reality environment. *Journal of Manufacturing Systems*, 74:703–715, 2024.

[41] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[42] E. B. Zaken, S. Ravfogel, and Y. Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

[43] X. Zhang and Q. Yang. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4435–4439, 2023.