This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3326522

NL2Color: Refining Color Palettes for Charts with Natural Language

Chuhan Shi, Weiwei Cui, Chengzhong Liu, Chengbo Zheng, Haidong Zhang, Qiong Luo, and Xiaojuan Ma

Abstract— Choice of color is critical to creating effective charts with an engaging, enjoyable, and informative reading experience. However, designing a good color palette for a chart is a challenging task for novice users who lack related design expertise. For example, they often find it difficult to articulate their abstract intentions and translate these intentions into effective editing actions to achieve a desired outcome. In this work, we present NL2Color, a tool that allows novice users to refine chart color palettes using natural language expressions of their desired outcomes. We first collected and categorized a dataset of 131 triplets, each consisting of an original color palette of a chart, an editing intent, and a new color palette designed by human experts according to the intent. Our tool employs a large language model (LLM) to substitute the colors in original palettes and produce new color palettes by selecting some of the triplets as few-shot prompts. To evaluate our tool, we conducted a comprehensive two-stage evaluation, including a crowd-sourcing study (N=71) and a within-subjects user study (N=12). The results indicate that the quality of the color palettes revised by NL2Color has no significantly large difference from those designed by human experts. The participants who used NL2Color obtained revised color palettes to their satisfaction in a shorter period and with less effort.

Index Terms—chart, color palette, natural language, large language model

1 Introduction

Charts are extensively adopted as an effective visual format for conveying data. The choice of color is a crucial aspect of chart design, as it greatly contributes to the aesthetics, engagement, and memorability of charts [12]. In practice, users often need to refine the color palettes (i.e., a set of colors used to represent existing data in a chart [11]) of their charts to align with various factors such as data types, application contexts, and user preferences. For example, users may not be satisfied with the initial color palettes they design and hope to improve the palettes. In addition, if users wish to apply their previously created charts in a slide deck with a certain template, they need to modify the charts' color palettes to match the template.

Unfortunately, users' requests for refining color palettes are often vague and abstract [24], such as "I hope the color palette of the chart be more modern" and "The overall color scheme of the slides is dark blue, and I would like the charts to match it". In a formative study with six participants, we found that changing color palettes to meet such requirements is challenging, particularly for novice users in chart design. First, novices struggle to translate abstract refinement needs into specific color editing actions since they lack the necessary domain knowledge to identify the colors that could achieve their desired effects. Moreover, they may be uncertain about the details of their requests, such as how to select colors that complement the primary color of the presentation. While tools like Adobe Color [1] have been developed to aid in color palette selection, users may find it time-consuming and tedious to locate the appropriate palette among a large number of options. Even though they can find the right one, novice users may also struggle with mapping colors in the new palette to elements in the chart [55]. This becomes even more complex when users need to revise a well-designed color palette for purposes such as matching presentation themes while preserving certain intrinsic color relationships, such as maintaining originally intended sequential color scales.

Natural language interfaces (NLI) enable inexperienced users to complete complex tasks by simply expressing their intentions [23]. To make visualization authoring and modification more accessible, various natural-language-based visualization authoring tools [49] have emerged. However, these methods are primarily designed to support visualization manipulation based on concrete NL editing requests and may not fully accommodate vague or abstract color palette refinement intents. When users just have vague or abstract requests (e.g., "more modern" or "cleaner"), they still need to translate them into effective and executable low-level editing actions before employing these tools to achieve their desired modifications.

In this paper, we propose a tool, NL2Color, that allows novice users to refine the color palettes of charts by expressing their vague or abstract requests in natural language. Specifically, we first collected a dataset of 131 triplets, each consisting of the original color palette of a chart, a vague or abstract request for palette refinement, and a new color palette designed by human experts to fulfill the request. When a user inputs a chart and a chart palette refinement request, NL2Color automatically extracts the chart's color palette and retrieves the triplets in our dataset whose refinement requests are similar to the user's request. Using these triplets as few-shot prompts, our tool employs a large language model (LLM), GPT-3, to refine the original color palette of the chart to meet the user's requirement. We validated NL2Color through a comprehensive two-stage evaluation. First, we carried out a crowd-sourcing study to compare expert-designed and NL2Color-refined palettes and examine the effectiveness of our prompt design. The results indicate that the quality of the color palettes recommended by NL2Color has no significantly large difference from those designed by human experts. Second, we conducted a within-subjects user study with 12 participants. Both qualitative and quantitative results show that compared to manually refining color palettes with online resources, NL2Color supported users to obtain satisfactory modified color palettes in less time and with less effort.

In summary, the major contributions of this paper are as follows:

- NL2Color, a tool that enables users to refine chart color palettes by expressing their vague or abstract requests in natural language.
- A two-stage evaluation study that demonstrates the usefulness and effectiveness of NL2Color in helping users refine the color palettes of charts.

- C. Shi is with Southeast University and the Hong Kong University of Science and Technology. E-mail: cshiag@connect.ust.hk
- W. Cui and H. Zhang is with Microsoft Research Asia. E-mail: {weiweicu, haizhang}@microsoft.com
- C. Liu, C. Zheng, and X. Ma are with the Hong Kong University of Science and Technology. E-mail: {chengzhong.liu, cb.zheng}@connect.ust.hk and mxj@cse.ust.hk
- Q. Luo is with the Hong Kong University of Science and Technology and the Hong Kong University of Science and Technology (Guangzhou). E-mail: luo@cse.ust.hk

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2 Related Work

2.1 Natural Language Interfaces for Visualization

In recent years, there has been a growing interest in developing natural language interfaces (NLIs) for visualization systems. These works

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3326522

mainly focus on facilitating visual analytics and visualization creation.

The NLIs for visual analytics aim to assist users in the tasks of information discovery, search, and query [47]. For example, FlowSense [54] applied a semantic parser to understand user queries and accordingly manipulate the visualizations produced by a dataflow diagram to help users with visual data exploration within a dataflow system. Liu et al. [27] proposed ADVISor, a pipeline to automatically generate charts for tabular data to answer users' natural-language questions. Luo et al. [29] developed an end-to-end deep learning model, ncNet, which translates natural language queries raised by users to Vega-Lite to generate visualizations. In such research, how to address vague and underspecified natural language expressions is a key challenge and different solutions have been proposed. For instance, Hearst et al. [19] carried out an empirical study and proposed design guidelines for how an NLI should respond to vague modifiers in natural language queries. Setlur et al. [39] presented a system called Sentifiers to infer the data attributes involved in vague utterances.

Another line of research explores visualization creation based on natural language descriptions. For example, Cui et al. [15] designed an approach that automatically generated infographics according to natural language statements containing proportion facts. Rashid et al. [36] focused on chart production and explored an approach to generate bar, line, or pie charts for user-input natural language text.

However, there is still limited research on how NLIs can enable users to author and modify visualizations by expressing their desired outcomes in natural language. While some studies (e.g., [49]) have focused on authoring-oriented tasks, they are only applicable to natural language requests containing specific editing intents, such as "make the United States bar red". Hence, these systems do not fully support refinement requests that are vague and abstract.

2.2 Color Palette Design Tools

A variety of works have proposed different methods to facilitate color palette design. As AI-driven approaches gained attention, Peng and Chou [33] utilized sentiment analysis to help designers understand stakeholders' color palette requirements, while Bahng et al. [10] developed Text2Colors, a tool that employs input text semantics for grayscale image colorization. Qiu et al. [34] proposed a masked color model for recommending colors for different components in landing pages. These studies demonstrate that AI techniques can greatly enrich color palette generation tools.

In the field of visualization, there is a long history of studying tools for color palette design. Early works include ColorBrewer [17] for map coloring and the techniques proposed by Wijffelaars et al.'s work [50] for generating univariate palettes based on easily-understood perceptual-property parameters. Later research, such as Lin et al. [26] as well as Setlur and Stone [40], delved into color assignment based on concept-color associations. Shugrina et al. [41] introduced Color Builder, an innovative interface that integrates operations like swatches and smooth for enhanced visualization coloring. Yuan et al. [55] began to incorporate AI for color palette design by developing InfoColorizer, an interactive tool to recommend color palettes for infographics using a deep learning model trained on a large-scale infographics dataset. Wang et al.'s visualization authoring pipeline allows users to change chart colors using natural language, although explicit instructions like "Set the color of the Ford bar to red" are required [49].

Despite these advancements, existing research still exhibits limitations, such as a lack of consideration for the relationship between recommended palettes and users' initial palettes. Additionally, users may still struggle to effectively apply general color palettes to their charts. Our work addresses these gaps by offering a more practical and user-friendly tool for color palette design in visualization authoring.

3 FORMATIVE STUDY

3.1 Participants and Procedure

We invited 6 participants (3 female, 3 male; FP1-6) by word-of-mouth. They all have no background in design but have the need to use charts in their daily life (Table 1).

Table 1: Demographics of all participants in the formative study, including each participant's ID, gender, and chart usage scenarios.

ID	Gender	Chart Usage Scenarios
FP1	Male	Research Paper, Technical Report, Presentation Slides
FP2	Male	Research Paper, Presentation Slides, Storyboard, Web UI Design
FP3	Female	Research Paper, Presentation Slides, Product Analysis Report, User Behavior Analysis Report
FP4	Female	Research Paper, Presentation Slides
FP5	Female	Presentation Slides, Course Report
FP6	Male	Research Paper, Work Report

We conducted semi-structured interviews with these participants. After signing the consent form, they were first asked to recall and describe their latest experience of refining the color palettes of charts, including but not limited to whether there were some refinement requests that are vague or abstract, what vague or abstract requests they have, and how they modify the charts to satisfy such requests. Then we asked about the difficulties they faced in the refinement process and the need for facilitating the color palette refinement. Finally, we invited the participants to envision what services they would like to have for the color palette refinement and what expectations and concerns they had for such services.

3.2 Results

3.2.1 Refinement Requests

All participants reflected that they had vague or abstract requests for chart color palette refinement during the chart designing process. Based on the interview results, we identified two common types of refinement requests: descriptive-word-based and chart-topic-based (Table 2).

Descriptive-word-based. We observed that all participants utilized descriptive words or phrases to specify their desired color palettes in the refinement requests. Specifically, such requests could be further divided into those with detailed references and those without (Table 2). For without-reference requests, participants only expressed their vague feelings, such as desired styles (e.g., "cyberpunk style") and change directions (e.g., "more professional"), about the original charts but did not know what specific colors can achieve their desired outcomes. For example, FP2 stated that sometimes he felt the color palette of a chart was not professional enough, yet he could not imagine in his mind what kind of color palette was professional. An interesting finding is that with the same without-reference requests, participants may desire different refinements in different usage scenarios, either brand-new or fine-tuned color palettes. Four participants pointed out that if the original color palettes largely meet their needs, they just need a fine-tuning to the palettes, which means that the hues of the colors in the original palettes do not need to be revised but only other small changes (e.g., increasing the lightness of the colors) are required. In other cases, all participants reported that they desire brand-new color palettes when they have without-reference refinement requests.

Furthermore, two participants mentioned that sometimes they remember a reference chart in their minds but cannot find it. Therefore, they can only describe the reference in natural language (e.g., "I would like a fresh and lovely color scheme, specifically a yellow and green palette."). When having such requests, both of them sought to gain a brand-new set of color palettes.

Chart-topic-based. Three participants reflected that they may have refinement requests about making the charts' color palettes align with specific topics, such as "environmentally friendly". For instance, FP3

Table 2: Refinement requests for chart color palettes.

		Brand-new	Fine-tuned
Descriptive-word-based	Without-reference (e.g., styles)	✓	✓
	With-reference	✓	✗
Chart-topic-based		✓	✗

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3326522

said that she always makes presentation slides about different topics and she would like the color palettes of the charts to match her topics when she creates the charts for the slides. When having chart-topic-based requests, the three participants all expressed that they hope to obtain brand-new color palettes.

3.2.2 Challenges and Needs for Color Palette Refinement

Four participants indicated that while they would like the color palettes of charts to be revised, they request that the direction of the color scales embedded in the original palettes should be preserved in the refined palettes. For example, FP5 said that *"if some components of a chart are denoted by sequential/diverging colors, the revised color palettes should also contain corresponding sequential/diverging colors to represent these components"*. However, they were concerned that maintaining the existing representations of colors makes the palette update rather difficult. FP4 complained that it is challenging to identify complex color scales when the original color palettes contain many colors. *"Sometimes I cannot distinguish whether the several colors are a set of sequential colors or categorical colors with similar hues"* (FP4). Even though such color scales of the original palettes can be identified properly, a tedious and time-consuming manual color mapping is commonly required because the consistency of the direction of the color scales between refined palettes and original ones is rarely satisfied. For instance, FP1 shared a personal experience that he wanted to revise a color palette containing a set of sequential colors, while he could only search for a satisfactory new categorical palette and needed to manually extend a color of it into a set of sequential colors based on the original sequential colors.

In addition, we found that individuals may have different preferences when refining the color palettes of charts. When different participants have the same refinement request, their expected new palettes may be different. Even for the same request from a single user, different palettes may be selected in different scenarios. Therefore, all participants suggested that it would be helpful if our tool could provide multiple possible options that fulfill the refinement requests so that they could select one based on their preferences.

Table 3: The chart types contained in the dataset.

Chart Type		N	Percent
Line chart		5	8.3%
Bar chart	Grouped bar chart	12	20.0%
	Stacked bar chart	9	15.0%
Pie chart		8	13.3%
Area chart		8	13.3%
Scatter chart		7	11.7%
Box chart		11	18.3%

3.2.3 Design Requirements

Based on the qualitative results from the participants, we concluded three design requirements for our system design.

- **DR1: Support both brand-new and fine-tuned color palette refinement requests.** From the formative study, we observed that users often desire brand-new or fine-tuned color palettes in various usage scenarios. Therefore, our tool should enable the production of both brand-new and fine-tuned color palettes in response to users' refinement requests.

- **DR2: Accommodate users' refinement intents while preserving the direction of color scales in the original color palettes.** When refining the color palettes of charts, users typically do not want to break the well-designed color scales in the original palettes. Thus, it is important to maintain the direction of color scales of the original color palettes while fulfilling users' refinement requests.

- **DR3: Provide multiple options for a single input refinement request to cater to various user preferences.** According to our formative study, individual users may have different preferences and needs for color palettes despite having the same color palette refinement requests. Hence, for each input refinement request, the tool should provide multiple potential revised color palettes to cater to diverse user preferences.

4 NL2COLOR

Based on the design requirements, we presented a tool, NL2Color, that enables novice users to refine the color palettes of charts by expressing their vague or abstract requests and intents in natural language. In this section, we introduce the implementation details of NL2Color, including data collection and a pipeline that automatically refine chart color palettes based on users' requests.

4.1 Data Collection

The data for NL2Color were collected through two methods. First, we invited eight novice users who do not have backgrounds in design and introduced the types of vague or abstract chart palette refinement requests identified in our formative study to them. For each type of refinement request, we presented one or two examples mentioned by the participants in the formative study, helping the novice users understand the concept of vague or abstract refinement requests. After the introduction, we asked the eight novices to write down potential vague or abstract requests they may make in their day-to-day life when wanting to change the color palettes of charts. We removed duplicated requests and finally obtained 41 unique requests. Then we randomly collected 84 SVG-based charts of various common types from the Plotly Chart Studio[1], a widely used website that allows users to manually create and share charts [20]. For the subsequent model training in the color palette refinement module (Section 4.2.2), we only kept the charts whose color palettes are categorical colors and discarded the others, resulting in 60 charts in our dataset (Table 3). We showed these 60 charts to the eight novices and asked them to match the requests they provided to each chart. For the type of without-reference requests, we also asked them whether they wanted to get a brand-new or fine-tuned color palette. We found that each chart corresponded to at least four requests. Subsequently, for each chart, we randomly selected two matched requests, resulting in 120 pairs of a chart and a corresponding abstract refinement request for the chart color palette. Among them, 80 pairs require brand-new color palettes, and 40 pairs require fine-tuned ones. As the other method to collect data, we asked the participants in the formative study (Section 3) to provide the charts they mentioned in the interview that they were not satisfied with, as well as their corresponding refinement requests. Through these two methods of data collection, there are a total of 131 pairs of original charts and requests in our dataset, 85 pairs for brand-new requests and 46 pairs for fine-tuned requests **(DR1)**. We then invited nine design experts to design a new color palette for the chart in each pair based on the refinement request. These experts all have more than five years of design experience and often create charts in their daily design work (Table 4).

We extracted the color palettes of the charts in our dataset. Specifically, for each chart, we identified all the unique colors in the SVG document while excluding those employed for texts, background, and axes. After this, we finalized our dataset of 131 triplets of (1) the original color palette of a chart, (2) a vague or abstract color palette refinement request, and (3) a new color palette designed by human experts according to the request.

4.2 Color Palette Refinement Pipeline

We developed a pipeline that refines the color palette of the input chart based on the user's vague or abstract refinement requests. The pipeline consists of two modules: an original color palette extraction module and a color palette refinement module.

4.2.1 Original Color Palette Extraction Module

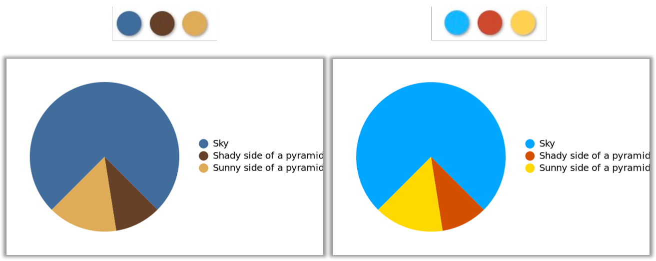
When a user uploads an SVG-based chart, our module automatically extracts its color palette. For this purpose, we first identified all the

[1] https://chart-studio.plotly.com/feed/

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3326522

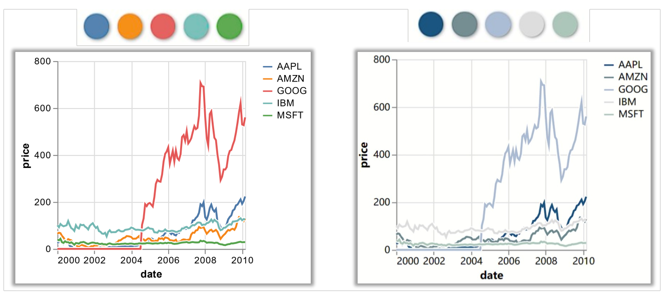
Table 4: Demographics of the nine design experts, including each expert's ID, gender, design experience (in years), and previous design activities.

ID	Gender	Design Experience.	Previous Design Activities
1	Female	5	Graphic Design, Interaction Design, Industrial Design, Service Design
2	Prefer not to say	5	Mobile App UI Design, Activity Poster Design, Industrial Design
3	Female	6	Web UI Design, Activity Poster Design, Product Advertisement Promotion Design
4	Male	6	Visualization Design, Product Design, Industrial Design
5	Female	5	Mobile App UI Design, Activity Poster Design, Industrial Design
6	Female	7	Mobile App UI Design, Illustration Design, Product Advertisement Promotion Design, Game Design, Service Design, Pavilion Design
7	Female	5	Mobile App UI Design, Activity Poster Design, Game Design
8	Male	5	Mobile App UI Design, Interaction Design, Interior Design
9	Male	5	Mobile App UI Design, Industrial Design

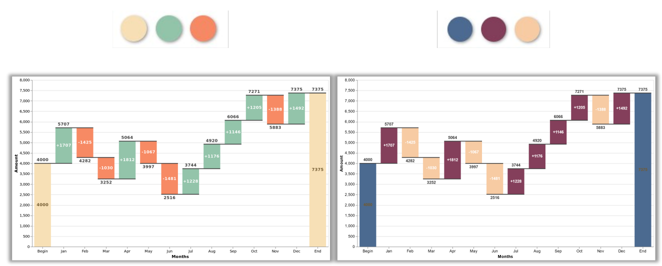


(a) Fine-tuned request: "Please use more vibrant colors that create contrast and add energy to the chart."

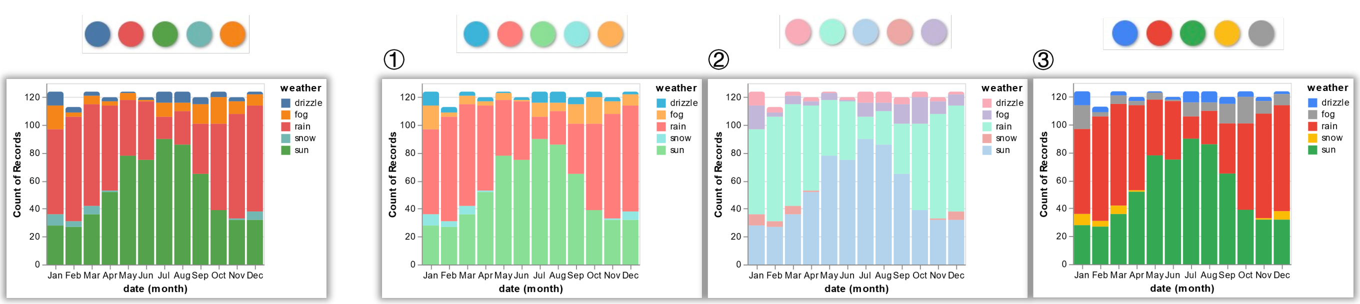
(b) Brand-new request: "I think this chart is too fancy."



(c) Brand-new request: "I want the colors to show a sense of silence."



(d) Brand-new request: "I would like a professional chart. I remember that many professional charts would use navy blue and maroon colors."



(e) ① Fine-tuned request: "I think the chart should have a more cartoon style but not industrial."; ② Brand-new request: "I think the chart should have a more cartoon style but not industrial."; ③ Brand-new request: "Please use our company's (Google) colors and ensure that the chart design is consistent with our overall visual identity."

Fig. 1: Examples of color palette refinement by NL2Color. (a)-(d) show four pairs of an original chart (left) and a new chart (right) refined by NL2Color according to the request. (e) shows an original chart (left) and three new charts (right) NL2Color generated according to three refinement requests. The color palette of each chart is displayed above the chart. The original charts are collected from Vega-Lite [9].

unique colors in the SVG document in a similar way we extracted the color palettes of the charts in our dataset (Section 4.1).

Color scales, i.e., categorical, sequential, and diverging colors, are commonly used by chart users to effectively communicate data [42]. To preserve the color scales in the original palette, we identified all sets of sequential colors (i.e., a gradation of colors that go from light to dark or dark to light [38]) and diverging colors (i.e., two sequential colors based on two different hues that meet in a neutral midpoint [13]) from the set of unique colors we extracted **(DR2)** and only kept their primary colors in the original color palette for the subsequent model training (Section 4.2.2). Specifically, we first obtained all sets of sequential colors by identifying the groups of colors that have the same hue and have a linearly monotonic sequence of increasing (or decreasing) lightness [13]. Based on [57], the luminance trajectory should be balanced between the two sets of sequential colors in a set of diverging colors. Thus, we transferred all these sets of sequential colors into HCL (Hue-Chroma-Luminance) color space. If there are two sets of sequential colors with equal intervals in luminance, they would

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3326522

combine to form a set of diverging colors. Then we identified the primary color(s) of each set of sequential and diverging colors. We defined the primary color of sequential colors as the color closest to the center of the gradient and the primary colors of diverging colors as the two primary colors of the pair of sequential colors they contain. Therefore, we arranged the colors in each sequential color group from lightest to darkest and regarded the color located in the middle or, if the sequential color group contains an even number of colors, the lighter of the two middle colors as its primary color. Finally, we combined these primary colors with the remaining categorical colors in the set of unique colors other than sequential and diverging color groups as our extracted original color palette.

4.2.2 Color Palette Refinement Module

This module employed OpenAI's GPT-3 model [14] to refine the original chart color palettes based on users' requests. The GPT-3 model has been demonstrated to have high performance on various tasks using a small number of examples and a well-crafted prompt [43, 46, 53]. We followed the principles and techniques proposed by [31, 37] to design prompts for GPT-3. Specifically, we crafted two few-shot prompts, respectively, for the brand-new and fine-tuned color palette refinement requests **(DR1)**. In each prompt, we first described our task, including the task goal, the input, and the output, as well as the definition of brand-new or fine-tuned color palettes (please see the supplementary material). Then we applied the text embedding API[2] of OpenAI to get the embeddings of the natural language request input by the user and the vague or abstract requests in the triplets in our dataset. If the user requires a brand-new (fine-tuned) color palette, we would solely consider brand-new (fine-tuned) requests in our dataset. We calculated the cosine similarity - a commonly used effective measure of text similarity [22, 48] - between the user-input request and each request in the dataset based on their embeddings. The five requests in our dataset with the highest similarity were selected, and the corresponding triplets were added to the prompt for the GPT-3 model to perform few-shot learning. Finally, we concatenated the original color palette extracted from the user-uploaded chart (Section 4.2.1) and user-input refinement request to the prompt and passed it to the model to refresh the color palette. The model's output was controlled to present ten alternative palettes to provide multiple options to users **(DR3)**.

Once the updated color palettes were produced, we extended the new primary colors in them into new sequential or diverging colors to match the original palette **(DR2)**. Specifically, for each color in a set of sequential colors we identified through the original color palette extraction module (Section 4.2.1), we computed the difference value between the primary color and it in the luminance channel. Then the new color corresponding to this color can be obtained by adding this difference value to the luminance value of the new primary color in the new palette. In the same way, the two groups of sequential colors are obtained respectively, thereby extending a new set of divergent colors.

Note that our pipeline only supports SVG-based charts for easy color extraction. We acknowledge that this may limit users' flexibility in the chart design process. Our main goal in this work is not to develop a full-fledged system but to propose a basic method for automatically refining the color palette of a chart according to the user's intent expressed in natural language. Future work could improve our methods for original color palette extraction to make it applicable to diverse chart formats.

5 EXAMPLE OUTPUT

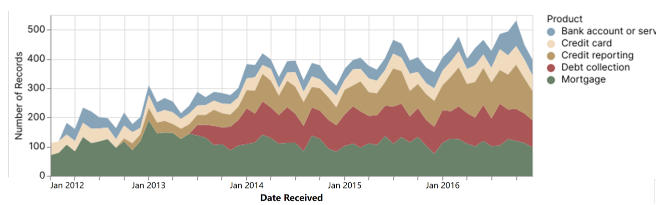
Fig. 1 shows several examples of color palette refinement by NL2Color. As shown in Fig. 1(a), we used our tool to fine-tune the color palette of the original chart (left) to make it more vibrant and energetic. The modified color palette keeps the hues of the colors in the original palette (i.e., blue, brown, and yellow) but makes the colors brighter. The contrast between the colors is also more pronounced following the refinement request. For the second pair of examples (Fig. 1(b)), the new palette does not maintain the same hues as the original palette since the input request indicates a desire for a brand-new palette. Even if we do not directly state what kind of palette we want in the refinement request but simply describe the problem with the palette (i.e., "I think the chart is too fancy"), NL2Color still successfully understands our needs and returns a palette with plain colors. To satisfy the requirement in Fig. 1(c), NL2Color changes the original colors to a cooler tone to create a calm and serene feeling. Fig. 1(d) showcases an example of NL2Color processing with-reference requests. It not only applies the colors from the description of the reference palette but also makes sure that the other NL2Color freely complemented colors are harmonious with these colors and create the professional feel the user wants. Fig. 1(e) displays the new charts that NL2Color recommends based on the same original chart but according to different refinement requests. The colors of 1 are more vivid and playful while maintaining the same hues as the original palette. In 2, NL2Color directly uses some candy colors to generate a brand-new chart in cartoon style. As for 3 which is requested to keep consistent with Google's overall visual identity, NL2Color directly applies the four main theme colors of Google company in the new palette and complements an extra grey color to ensure that the number of colors in the new palette is consistent with the original one without affecting the expression of Google's visual identity. These examples showcase how NL2Color handles the different types of color palette refinement requests (Table 2) and confirm that NL2Color meets our **DR1**.

6 EVALUATION

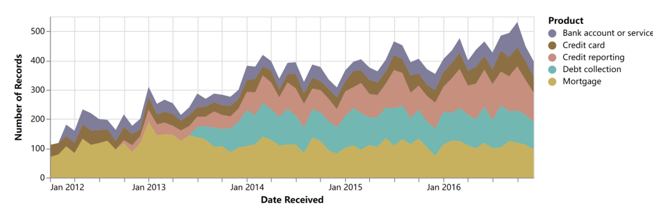
To assess the effectiveness and usefulness of NL2Color, we conducted a two-stage evaluation, including a crowd-sourcing study and a within-subjects user study. In this section, we present our evaluation study design and the findings regarding the performance of our tool and whether and how it would influence novice chart users' color palette revision process.

6.1 Study1: Crowd-sourcing Study

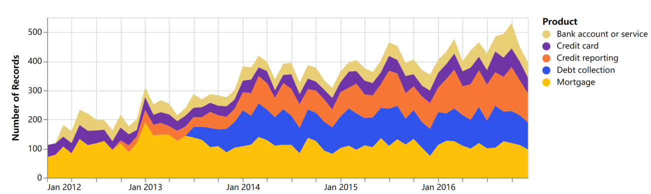
In this study, we evaluated how well NL2Color refines color palettes from the perspective of chart readers. Since we designed two models



(a) Expert-designed. Rating: 4.4



(b) NL2Color-refined. Rating: 4.0



(c) Zero-shot-model-refined. Rating: 3.3

Fig. 2: An example of the new charts refined in the three conditions of our crowd-sourcing study, along with the ratings they received from the participants. The original chart is Fig. 5(a) and the refinement request is "The chart should have a cultural or historical theme".

[2] https://platform.openai.com/docs/guides/embeddings

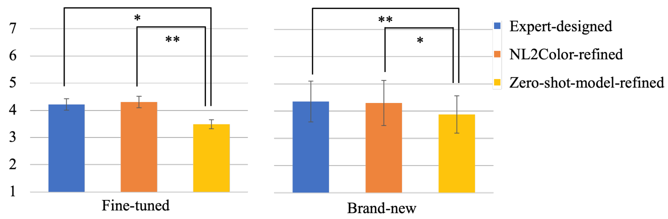


Fig. 3: Means and standard errors of the participants' ratings on the revision quality respectively for fine-tuned requests (left) and brand-new requests (right) on a 7-point Likert scale (1 - absolutely not meet the refinement request, 7 - absolutely meet the refinement request; *: $p < .05$, **: $p < .01$).

respectively for the fine-tuned and brand-new requests, we evaluated their performance separately. On the one hand, we compared the quality of the new color palettes refined by NL2Color with those designed by human experts. On the other hand, to validate our prompt design (Section 4.2.2), we further crafted two zero-shot prompts (i.e., the prompt only contains task description) for GPT-3 corresponding to the two few-shot prompts we designed for NL2Color to respectively generate fine-tuned and brand-new color palettes. Overall, we, for each type of request (i.e., brand-new or fine-tuned), compared three conditions: (1) expert-designed, (2) NL2Color-refined, and (3) zero-shot-model-refined color palettes (Fig. 2).

6.1.1 Study Setup

Using the first way of data collection we mentioned in Section 4.1, we collected 60 pairs of (1) a chart with an original color palette and (2) a vague or abstract refinement request for the crowd-sourcing study. Among them, 40 pairs consist of requests for brand-new color palettes and 20 pairs for fine-tuned ones. To acquire expert-designed palettes, we invited eight professional designers to revise the color palette for each chart according to the corresponding request. For the NL2Color-refined and zero-shot-model-refined conditions, we leveraged the results returned by NL2Color and the models with zero-shot prompts. As each of these models would provide ten alternative color palettes for a given pair of data, we randomly selected a palette from the options as the model-refined palette.

We developed crowdsourcing questionnaires on Prolifc[3]. In each questionnaire, six problem sets (two about fine-tuned requests and four about brand-new requests) were randomly assigned to each participant and evaluated one by one. Each problem set contains a pair of an original chart and a refinement intent, as well as three new charts colored with the three color palettes refined in the three conditions. We referred to these three new charts as “Chart 1”, “Chart 2”, and “Chart 3” to eliminate potential bias in human designers and the models. For each problem set, we asked participants to respectively rate how each new chart can meet the refinement request on a 7-point Likert scale (1 - absolutely not, 7 - absolutely meet).

We filtered out unreliable responses if the answers met any of the following criteria: 1) unreasonably completed the questionnaire too quickly and 2) had consistent patterns in ratings. We also made sure that each pair of data in our dataset for the crowdsourcing study received ratings from at least five participants. For those that remain less than five valid responses, we repeated the aforementioned crowdsourcing steps until five valid responses were obtained. Finally, we recruited 71 participants in total and averaged the ratings from different users per new chart. Each of these participants is given 0.76 USD as a reward for the valid questionnaire completion and the average duration for each questionnaire completion is about five minutes.

6.1.2 Data Analysis

For each type of request (i.e., brand-new or fine-tuned), we first performed the Shapiro-Wilk test on the ratings of the refined color palette in the three conditions. The results show that they all followed the normal distribution ($p > .05$). Therefore, we ran the Friedman test with post-hoc Wilcoxon signed-rank tests with Bonferroni correction [25] to assess the difference in the participants' ratings on the quality of the revised color palettes across the three conditions.

6.1.3 Results

The results indicate significant differences ($\chi^2(2) = 16.30$, $p < .01$) between the quality of the fine-tuned charts in the three conditions (Fig. 3 (left)). The pairwise comparisons showed that the charts refined by human designers (4.21, [3.78, 4.65] 95% CI) and NL2Color (4.31, [3.87, 4.74] 95% CI) received significantly higher ratings (expert-designed: $Z = -2.02$, $p < .05$; NL2Color-refined: $Z = -3.92$, $p < .01$) than those revised by the zero-shot model (3.49, [3.14, 3.84] 95% CI). However, the difference between the expert-designed condition and the NL2Color-refined condition is not significant.

We also found a significant difference between the participants' ratings on the charts with brand-new color palettes in the three conditions ($\chi^2(2) = 6.87$, $p < .05$). As shown in Fig. 3 (right), participants gave significantly higher ratings (expert-designed: $Z = -2.93$, $p < .01$; NL2Color-refined: $Z = -2.40$, $p < .05$) to the charts modified by human experts (4.37, [4.13, 4.61] 95% CI) and NL2Color (4.29, [4.03, 4.56] 95% CI) compared to those revised by zero-shot model (3.84, [3.63, 4.05] 95% CI); no statistical difference is found between the brand-new color palettes designed by human experts and NL2Color.

These results proved that the color palettes refined by NL2Color, regardless of whether they are fine-tuned or brand-new, have no significantly large difference from those designed by human experts. Furthermore, the prompts we designed for NL2Color are demonstrated to be effective in facilitating the GPT-3 model to handle the color palette refinement tasks.

6.2 Study2: User Study

In this study, we evaluated NL2Color with real users and explored its influence on users' color palette refinement process. We conducted a within-subjects study with 12 participants, where the participants completed chart palette refinement under two conditions. In the control condition, the participants are allowed to use any tools and websites they commonly use in their routine practices (e.g., Adobe Color [1], Color Hunt [4]) to revise the color palettes. In the experiment condition, participants were allowed to use NL2Color only. We did not choose any specific tool, such as Adobe Illustrator [2], as the baseline in the user study since we found from the formative interviews (Section 3) that each novice has his/her own way of exploring new palettes and we did not reach a conclusion regarding widely-used tools.

6.2.1 Experimental Website Design

We developed an experimental website for the experimental condition based on NL2Color (Fig. 4) which involves five panels. Specifically, the Original Chart Panel (Fig. 4C) allows a user to upload an original chart for refinement and displays this chart. Once the chart is uploaded, NL2Color would automatically extract its color palette. All extracted sets of sequential colors, diverging colors, and the color palette are shown in the Color Palette Panel (Fig. 4A). After the chart submission, the user could input the refinement intent in natural language in the Refinement Request Panel (Fig. 4B). There are a Brand-new/Fine-tuned button group and a Get Refined Palettes button under the input box. The Brand-new/Fine-tuned button group is used for users to specify whether they want a brand-new color palette or a fine-tuned one. Once the user clicks on the Get Refined Palettes button, NL2Color would recommend new color palettes according to the user's request. The returned new palette alternatives and the thumbnails of the charts colored with them are listed in the Refined Color Palettes Panel (Fig. 4E). The user can select the new color palette of interest to view it in the Refined Chart Panel (Fig. 4D).

6.2.2 Participants and Procedures

We recruited 12 participants (7 males and 5 females) with diverse academic backgrounds through word-of-mouth. They all self-reported

[3]https://www.prolific.co/

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3326522

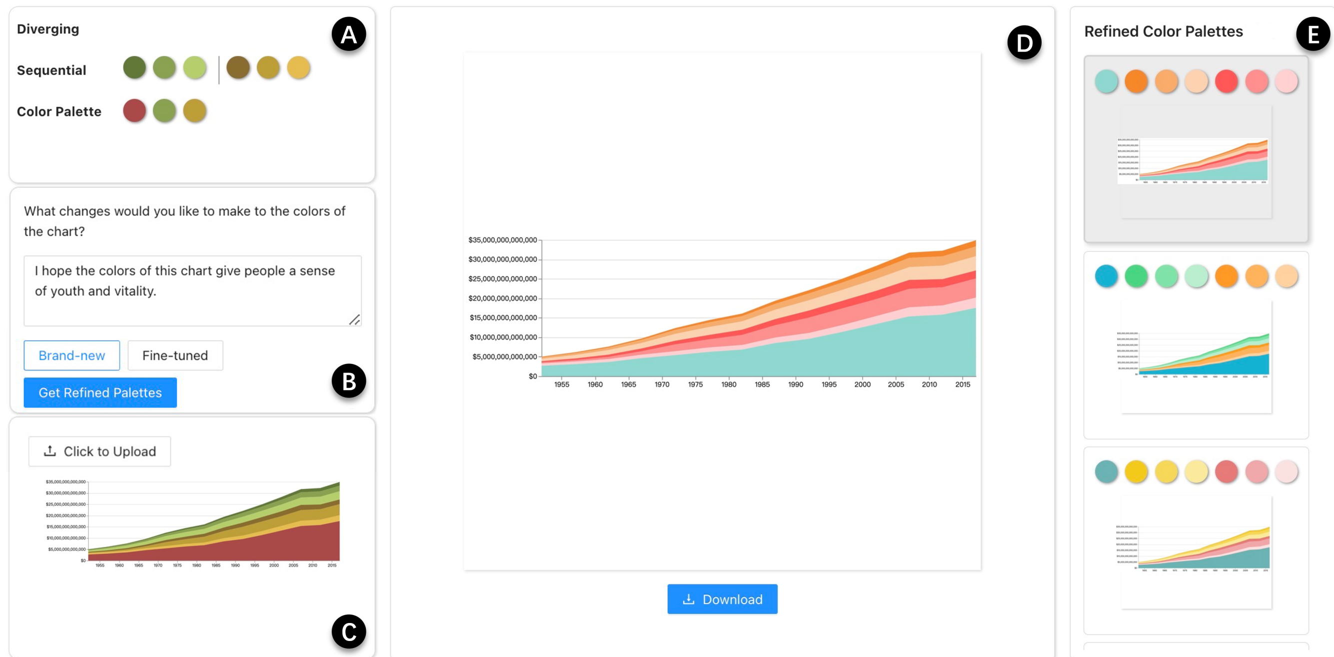


Fig. 4: Screenshot of the experimental website. The example of color palette refinement showcases that NL2Color meets our **DR2** and **DR3**.

having no domain knowledge in chart design but have the need to create and modify charts in their daily life.

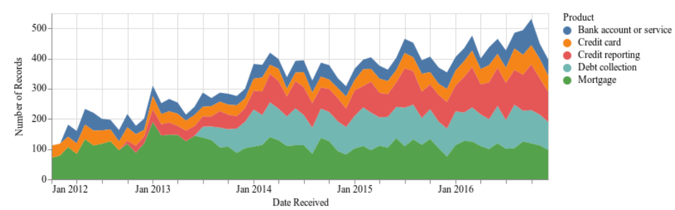
We discussed with two professional designers who helped us refine color palettes in the expert-designed condition of the crowd-sourcing study (Section 6.1) and designed two chart color palette refinement tasks for our within-subjects user study. To ensure that the two tasks are of similar difficulty, we chose two charts respectively for the two tasks whose color palettes contain the same number of colors. Also, as most of the formative study participants mentioned the without-reference requests that indicate the change directions of an original palette, we selected such requests in our task design. The two color palette refinement tasks are as follows:

- *T1:* Refine the color palette of the chart (Fig. 5(a)) to make it more edgy and bold.
- *T2:* Refine the color palette of the chart (Fig. 5(b)) to make it more playful and fun to appeal to a younger audience.

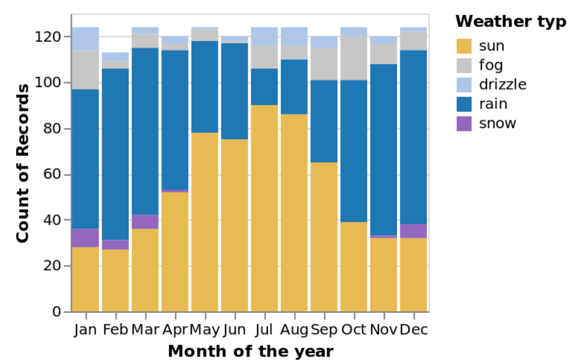
After obtaining the participants' consent, we asked each of them to complete these two tasks separately in the control and experiment conditions. For each task, participants need to refine the color palette until they were satisfied with the results without a time limitation. Before the tasks using NL2Color, we carefully introduced the experimental website to the participants and gave them 5 minutes to familiarize themselves with NL2Color. To alleviate the potential order effect, we counterbalanced the task assignment and the order of the two conditions. We recorded the video of these two user study sessions. At the end of each session, we asked the participants to fill out a questionnaire on a 7-point Likert scale to rate (1) user confidence in the final refined new palettes [52]; (2) the cognitive load during the tasks, measured using the NASA Task Load Index (NASA-TLX) [18]. In the in-task survey in the experiment condition, participants were also asked to rate their perceptions of NL2Color, including the usability, usefulness, and user satisfaction with the tool [28]. To better understand participants' ratings and behavior, we further conducted a semi-structured interview with them upon the completion of the two sessions.

6.2.3 Data Analysis

As a series of Shapiro-Wilk tests showed that all quantitative measures (i.e., user behavior data coded from video recordings and participants' responses on the questionnaires) have significant departures from the normal distribution, we conducted Wilcoxon signed-rank tests to compare the two conditions regarding each measure. As for the qualitative data, two authors of this paper conducted a thematic analysis on the transcripts of the post-study semi-structured interview and identified key themes in participants' feedback.



(a) The original chart for refinement in *T1*.



(b) The original chart for refinement in *T2*.

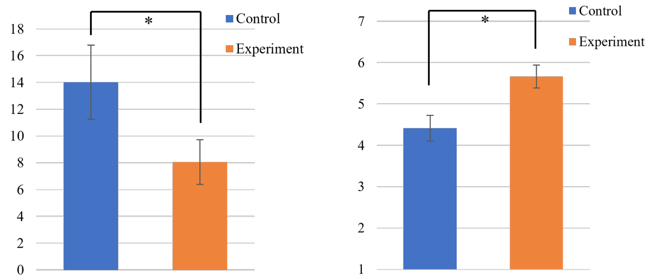
Fig. 5: The original charts for refinement in the user study.

6.2.4 Results

Here we summarize the quantitative results regarding participants' task completion time, user confidence in the refined color palettes, and perceived cognitive load during the tasks, as well as qualitative findings from the user study.

Completion time. To inspect how well NL2Color helps users revise chart color palettes, we performed a statistical analysis of the participants' task completion time. As shown in Fig. 6(a), participants using NL2Color (8.07, [4.40, 11.73] 95% CI) spent significantly less time

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3326522



(a) Task completion time. (b) User confidence

Fig. 6: Means and standard errors of the participants' task completion time and user confidence in their refined color palettes (*: $p < .05$.)

($Z = -2.00$, $p < .05$) completing the color palette refinement tasks than when they followed their routine practices (14.02, [7.92, 20.12] 95% CI). In the control condition, the participants mainly use three categories of tools: 1) six participants applied manual tools, including Adobe Photoshop [3], Adobe Illustrator [2], Inkscape [6], and PowerPoint, where they selected colors from the color panels to refine the original color palettes; 2) two participants used the color palette generation support tool, Adobe Color [1]; 3) four participants searched for color schemes on color palette recommendation websites (e.g., Material UI [7], Coolors [5], Palettable [8]) to guide their palette refinement. Compared to utilizing these tools to complete color palette revision, all participants reflected that they preferred to use NL2Color in the interview. With NL2Color, they did not need to spend a lot of time and effort to learn the complex functions of the aforementioned professional software (P1, P6-7), search for pre-designed color palettes online for charts to be refined (P4-5, P11), and manually map and substitute the original colors with those in the refined palettes (P1, P9, and P11). P3 added that *"Even though sometimes the results returned by NL2Color still need to be manually fine-tuned, it cost much less time than designing a new color palette from scratch, especially when the color palettes contain many colors"*.

User confidence. The results reveal that compared with the control condition (4.42, [3.73, 5.11] 95% CI), participants reported to have significantly higher confidence ($Z = -2.49$, $p < .05$; Fig. 6(b)) in their final refined color palettes in the experiment condition (5.67, [5.04, 6.29] 95% CI). For one thing, in the control condition, participants typically revised the colors in the original palette one by one, or searched for a pre-designed palette and then fine-tuned it to get the final new palette. Hence, they often could only come up with one new palette after the color palette refinement process. In contrast, NL2Color provides various refined palette alternatives and users can choose the satisfactory one after carefully comparing them, which makes participants feel more confident in their final palette design. For another, three participants claimed that they usually compromise on the quality of refined palettes due to their limited capability in chart design. On the contrary, when using NL2Color, users would pursue higher standards (e.g., color harmony, visual appeal) on palette refinement and they found that NL2Color could help them well meet their standards, improving their confidence in refined color palettes (P3).

Cognitive load. Using Wilcoxon signed-rank tests, we analyzed participants' cognitive load during the color palette refinement process on each related dimension in the control condition and experiment condition. We found significant differences in the Mental Demand ($Z = -2.29$, $p < .05$), Physical Demand ($Z = -2.17$, $p < .05$), Temporal Demand ($Z = -2.77$, $p < .01$), Performance ($Z = -2.12$, $p < .05$), and Effort ($Z = -2.10$, $p < .05$) dimensions of cognitive load and marginally significant difference in the Frustration ($Z = -1.84$, $p = .07$) dimension (Fig. 7). Participants explained that they perceived less cognitive load when using NL2Color because they only needed to think about how to accurately communicate their palette revision requirements with our tool instead of being overwhelmed by low-level technical issues they encountered in the control condition, such as what exact colors could achieve their desired effects, how to assign the colors in the refined palettes to the chart elements, and how to preserve color relationships in the original palettes.

Usability and usefulness. Participants generally gave positive feedback on the usability and usefulness of NL2Color (Fig. 8). In the post-study interview, they praised our tool as it is *"satisfactory"* (9/12), *"convenient"* (5/12), *"intuitive"* (3/12), and *"user-friendly"* (2/12). All participants appreciated that NL2Color makes it easier to know where to start modifying color palettes and greatly simplifies the palette refinement process. P2 explained, *"Since I often only have vague and abstract needs for palette refinement, I do not know what colors I want at the beginning. However, the tool provides me with a variety of modified palettes from which I can choose one directly or use one as a basis for simple adjustments to obtain my desired new palettes"*. In addition, three participants reported that NL2Color facilitated them to examine and compare alternative refined color palettes intuitively. Without our tool, users need to choose and try different colors based on subjective feelings until find the proper colors to form the refined palettes, which *"causes lots of trial-and-error tweaking"* (P12). For example, as P5 complained, *"In my chart palette revision routine, I always find palettes online that seem to meet my refinement requests but actually do not work well after being applied in the charts. In such cases, I need to go back to search for other palettes that may satisfy my requirements"*. In comparison, NL2Color enables users to compare alternative palette designs in the charts adopting them. Moreover, P11 pointed out that NL2Color stimulated her creativity and inspired the palette refinement process since it recommended designs that she never thought of.

7 DISCUSSION

In this section, we discuss the generalizability of our work. Built upon the key findings in our user study, we then derive several design considerations and implications for NLIs for visualization. We also discuss the failure cases and limitations of our research.

7.1 Generalizability

Although our system, NL2Color, is designed for common charts, our proposed system design and pipeline could be easily extended to other types of visualizations (e.g., pictorial visualization and node-link diagram). Our approach could be adapted to help refine the color palettes of these visualizations based on users' vague or abstract requests (e.g., "more vivid" or "softer") by 1) adjusting the task description in prompts and 2) gathering and applying visualization-specific training data for few-shot learning. Although the prompt should be tailored to the specific type of visualization, our prompt design provides guidance for this and the template of our prompt (Section 4.2.2) could be applicable to other visualizations, which contains the task goal, the descriptions of input and expected output, the explanation of special constraints, and few-shot learning examples.

Meanwhile, color coordination influences the quality of visualizations' color palettes a lot [45]. The attributes related to color coordination can be easily incorporated into our system to improve the quality of color palette refinement and make our tool applicable for more complex charts (e.g., heat maps, and cartograms) where color coordination is especially important. For example, we could add additional constraints, such as color harmony [30] and visual consistency [35], into our prompts to enhance the quality of the refined color palettes.

7.2 Design Considerations and Implications

7.2.1 Balance Automated and Manual Visualization Editing

Although all the participants in our user study appreciated the convenience of NL2Color, five of them expressed their demand for manual chart editing. This is because of the inherent uncertainty of vague or abstract natural language requests. With such refinement requests, participants sometimes may not obtain satisfactory outcomes from NL2Color even after going through multiple iterations and hope to make more fine-grained adjustments on the basis of the returned results (e.g., manually fine-tuning the brightness parameter of a certain component of the returned chart). Hence, we suggest balancing automated

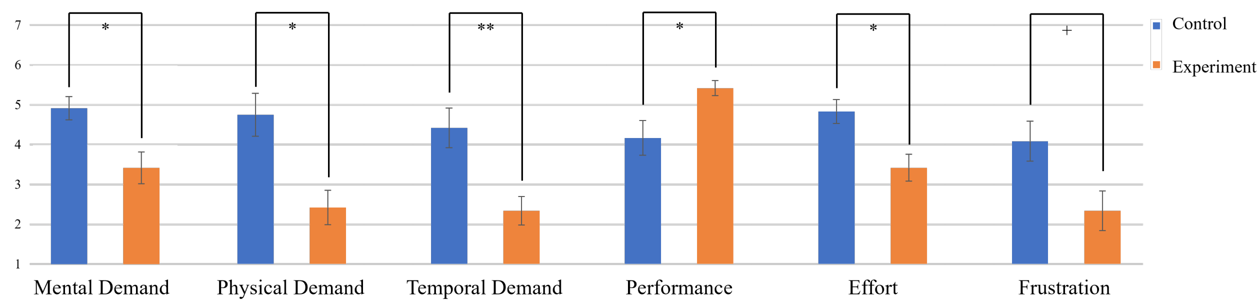


Fig. 7: Means and standard errors of the participants' cognitive load in the color palette refinement process on a 7-point Likert scale (+: $.05 < p < .1$, *: $p < .05$, **: $p < .01$).

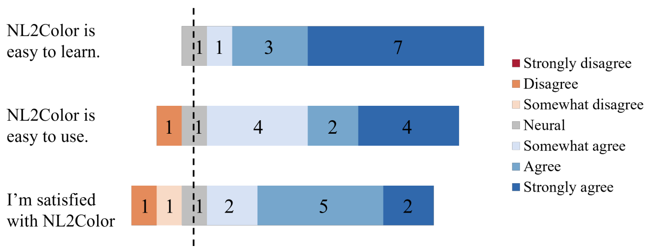


Fig. 8: User perception towards NL2Color.

and manual visualization editing in NLIs for visualizations to take the advantages of both the convenience of automation and the accurateness of manual operations.

7.2.2 Improve System Transparency to Promote NLI Discoverability and Debugging

In our user study, we observed that users had problems figuring out how to communicate effectively with NL2Color and how to evaluate its recommendations. Due to the black box nature of LLMs, users can only speculate about the reasons for the outcomes of the system [56]. Therefore, when receiving unexpected results, some users tried different expressions based on their speculation until obtaining satisfactory color palettes from NL2Color. Even if the returned results of NL2Color are basically satisfied, the unfamiliarity with the logic of LLMs makes users doubt whether there are other ways of expressing refinement requests that may result in better outcomes and users thus keep trying other expressions. Such issues cost a considerable amount of time during the color palette refinement process and negatively affect the effectiveness of our system. Therefore, when designing NLIs, designers can enhance model transparency and interpretability to prompt NLI discoverability (i.e., users' awareness of system-supported commands [16]) and debugging. For instance, in addition to the results of visualization manipulation, we could require LLMs also provide explanations of their recommendations.

7.2.3 Learn from Users to Mitigate Ambiguity Issues

One pain point we found from the user study is the expression ambiguity of vague and abstract revision requests. Consider a statement: *"I want the color palette to have the impression of a dark night"*. It is not clear whether the user expects a black or a dark blue palette. This issue is hard to resolve using current LLMs. A model that incorporates an understanding of user intent is thus required in NLI design. For example, NLIs could proactively ask users for further deliberation or give users multi-level choices to decode their thoughts behind their ambiguous commands. Moreover, NLIs could learn users' preferences from their historical decisions [21] so that they could infer user intents in ambiguous requests and provide personalized recommendations.

7.3 Failure Cases and Limitations

During the development of NL2Color, we observed some failure cases where our tool generates wrong or bad results. These failure cases mainly appear when fine-tuning the color palettes of charts. On the one hand, in some cases, such as users desire a brighter color palette while the colors in the original palette are relatively not dark, the difference between the fine-tuned color palettes and the original ones may not be noticeable. To resolve this issue, our system can allow users to explicitly request in prompts that new color palettes should showcase greater differences from the original ones. On the other hand, the difference between the multiple options of fine-tuned color palettes may also not be noticeable. In practical use, we can set a difference threshold based on the theory of just-noticeable difference [44] to constrain our system to return diverse color palettes.

In addition, for some complicated and long statements, such as *"Please revise the color palette to embody the essence of a summer sunset with hues that blend seamlessly from warm oranges to cool blues"*, NL2Color may not segment correctly and provide satisfactory color palettes. We believe this issue can be mitigated with more training data. Moreover, our system can decompose the complex revision requests into a series of sub-requests, each mapped to a distinct step that can be processed by LLMs. In this way, complicated palette refinement can be achieved by chaining and aggregating the results of each step [51].

Our design also has some limitations in terms of capability. NL2Color currently only supports modifying color palettes for SVG-based charts. This may limit users' flexibility during the color palette refinement process. Our system could be extended to be applicable to other types of files in the future. For example, we can apply algorithms, such as K-means clustering [32], enabling the color palette extraction from PNG-based or JPG-based charts.

8 Conclusion and Future Work

In this paper, we presented NL2Color, a tool that enables novice users to refine the color palettes of charts using natural language requests. The tool uses a dataset of 131 triplets each of which includes an original color palette of a chart, a vague or abstract request, and a new color palette designed by human experts according to the request. Our tool leverages the GPT-3 model to automatically fine-tune or generate brand-new color palettes by utilizing the triplets in our dataset whose refinement requests are similar to the user's input as few-shot prompts. Through a crowd-sourcing study and a within-subjects user study, we demonstrated the effectiveness and usefulness of NL2Color in helping novices modify chart color palettes with natural language.

In the future, we would like to expand our color palette extraction approach to support more file types. Moreover, we will enhance our dataset by collecting more high-quality data. On the one hand, we will collect more color palette revision requests with diverse expression styles and patterns so that our system can understand user intents in different commands and the expression ambiguity issue can be mitigated. On the other hand, we will invite more designers to help build our training dataset and ask different designers to refine the color palette for the same pair of an input original chart and a palette refinement request. This way, our system could learn different design styles from different designers and provide refined color palettes of diversity and high quality for a single input revision request to satisfy various user preferences. Furthermore, more state-of-art LLMs (e.g., GPT-4) will be employed to improve our tool's performance and robustness.

Acknowledgments

Many thanks to Zeyu Huang and Yicheng Hu for their valuable inputs. This research is supported in part by HKUST 30 for 30 Grant No.: 3030_003.

References

[1] Adobe Color. https://color.adobe.com/, Accessed in 2023.
[2] Adobe Illustrator. https://www.adobe.com/products/illustrator/, Accessed in 2023.
[3] Adobe Photoshop. https://www.adobe.com/products/photoshop/, Accessed in 2023.
[4] Color Hunt. https://colorhunt.co/, Accessed in 2023.
[5] Coolors. https://coolors.co/, Accessed in 2023.
[6] Inkscape. https://inkscape.org/, Accessed in 2023.
[7] Material UI. https://materialui.co/, Accessed in 2023.
[8] Palettable. https://www.palettable.io/, Accessed in 2023.
[9] Vega-Lite. https://vega.github.io/vega-lite/examples/, Accessed in 2023.
[10] H. Bahng, S. Yoo, W. Cho, D. K. Park, Z. Wu, X. Ma, and J. Choo. Coloring with words: Guiding image colorization through text-based palette generation. In *Proceedings of the european conference on computer vision (eccv)*, pp. 431–447, 2018.
[11] L. Bartram, A. Patra, and M. Stone. Affective color in visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1364–1374, 2017.
[12] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE transactions on visualization and computer graphics*, 19(12):2306–2315, 2013.
[13] C. Brewer. Color use guidelines for mapping and visualization. visualization in modern cartography, 1994.
[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
[15] W. Cui, X. Zhang, Y. Wang, H. Huang, B. Chen, L. Fang, H. Zhang, J.-G. Lou, and D. Zhang. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE transactions on visualization and computer graphics*, 26(1):906–916, 2019.
[16] Y. Feng, X. Wang, B. Pan, K. K. Wong, Y. Ren, S. Liu, Z. Yan, Y. Ma, H. Qu, and W. Chen. Xnli: Explaining and diagnosing nli-based visual data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2023.
[17] M. Harrower and C. A. Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.
[18] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.
[19] M. Hearst, M. Tory, and V. Setlur. Toward interface defaults for vague modifiers in natural language interfaces for visual analysis. In *2019 IEEE Visualization Conference (VIS)*, pp. 21–25. IEEE, 2019.
[20] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo. Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.
[21] H. Kaur, A. Williams, and W. Lasecki. Building shared mental models between humans and ai for effective collaboration. *CHI'19, May 2019, Glasgow, Scotland*, 2019.
[22] T. Kenter and M. De Rijke. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 1411–1420, 2015.
[23] T. S. Kim, D. Choi, Y. Choi, and J. Kim. Stylette: Styling the web with natural language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2022.
[24] M. Krause, T. Garncarz, J. Song, E. M. Gerber, B. P. Bailey, and S. P. Dow. Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 4627–4639, 2017.
[25] Y. Li, S. Sarcar, Y. Zheng, and X. Ren. Exploring text revision with backspace and caret in virtual reality. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–12, 2021.
[26] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum*, vol. 32, pp. 401–410. Wiley Online Library, 2013.
[27] C. Liu, Y. Han, R. Jiang, and X. Yuan. Advisor: Automatic visualization answer for natural-language question on tabular data. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*, pp. 11–20. IEEE, 2021.
[28] H. Lü and Y. Li. Gesture coder: a tool for programming multi-touch gestures by demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2875–2884, 2012.
[29] Y. Luo, N. Tang, G. Li, J. Tang, C. Chai, and X. Qin. Natural language to visualization by neural machine translation. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):217–226, 2021.
[30] B. J. Meier, A. M. Spalter, and D. B. Karelitz. Interactive color palette tools. *IEEE Computer Graphics and Applications*, 24(3):64–72, 2004.
[31] S. Mishra, D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi. Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*, 2021.
[32] I. Pavan Kumar, V. Hara Gopal, S. Ramasubbareddy, S. Nalluri, and K. Govinda. Dominant color palette extraction by k-means clustering algorithm and reconstruction of image. In *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19*, pp. 921–929. Springer, 2020.
[33] Y.-F. Peng and T.-R. Chou. Automatic color palette design using color image and sentiment analysis. In *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 389–392. IEEE, 2019.
[34] Q. Qiu, M. Otani, and Y. Iwazaki. An intelligent color recommendation tool for landing page design. In *27th International Conference on Intelligent User Interfaces*, pp. 26–29, 2022.
[35] Z. Qu and J. Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE transactions on visualization and computer graphics*, 24(1):468–477, 2017.
[36] M. M. Rashid, H. K. Jahan, A. Huzzat, R. A. Rahul, T. B. Zakir, F. Meem, M. S. H. Mukta, and S. Shatabda. Text2chart: A multi-staged chart generator from natural language text. In *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part II*, pp. 3–16. Springer, 2022.
[37] L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.
[38] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang. Mapping color to meaning in colormap data visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):810–819, 2018.
[39] V. Setlur and A. Kumar. Sentifiers: Interpreting vague intent modifiers in visual analysis using word co-occurrence and sentiment analysis. In *2020 IEEE Visualization Conference (VIS)*, pp. 216–220, 2020. doi: 10.1109/VIS47514.2020.00050
[40] V. Setlur and M. C. Stone. A linguistic approach to categorical color assignment for data visualization. *IEEE transactions on visualization and computer graphics*, 22(1):698–707, 2015.
[41] M. Shugrina, W. Zhang, F. Chevalier, S. Fidler, and K. Singh. Color builder: A direct manipulation interface for versatile color theme authoring. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.
[42] S. Smart, K. Wu, and D. A. Szafir. Color crafting: Automating the construction of designer quality color ramps. *IEEE transactions on visualization and computer graphics*, 26(1):1215–1225, 2019.
[43] D. So, W. Mańke, H. Liu, Z. Dai, N. Shazeer, and Q. V. Le. Searching for efficient transformers for language modeling. *Advances in Neural Information Processing Systems*, 34:6010–6022, 2021.
[44] M. K. Stern and J. H. Johnson. Just noticeable difference. *The Corsini Encyclopedia of Psychology*, pp. 1–2, 2010.
[45] C. Tseng, G. J. Quadri, Z. Wang, and D. A. Szafir. Measuring categorical perception in color-coded scatterplots. In *proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2023.
[46] M. Tuli, A. Li, P. Vaezipoor, T. Klassen, S. Sanner, and S. McIlraith. Learning to follow instructions in text-based games. *Advances in Neural Information Processing Systems*, 35:19441–19455, 2022.
[47] H. Voigt, Ö. Alaçam, M. Meuschke, K. Lawonn, and S. Zarrieß. The why and the how: A survey on natural language interaction in visualization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, pp. 348–374, 2022.

[48] T. vor der Brück and M. Pouly. Text similarity estimation based on word embeddings and matrix norms for targeted marketing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1827–1836, 2019.

[49] Y. Wang, Z. Hou, L. Shen, T. Wu, J. Wang, H. Huang, H. Zhang, and D. Zhang. Towards natural language-based visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1222–1232, 2022.

[50] M. Wijffelaars, R. Vliegen, J. J. Van Wijk, and E.-J. Van Der Linden. Generating color palettes using intuitive parameters. In *Computer Graphics Forum*, vol. 27, pp. 743–750. Wiley Online Library, 2008.

[51] T. Wu, M. Terry, and C. J. Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–22, 2022.

[52] L. Yan, E. L. Glassman, and T. Zhang. Visualizing examples of deep neural networks at scale. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

[53] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 3081–3089, 2022.

[54] B. Yu and C. T. Silva. Flowsense: A natural language interface for visual data exploration within a dataflow system. *IEEE transactions on visualization and computer graphics*, 26(1):1–11, 2019.

[55] L.-P. Yuan, Z. Zhou, J. Zhao, Y. Guo, F. Du, and H. Qu. Infocolorizer: Interactive recommendation of color palettes for infographics. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4252–4266, 2021.

[56] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2023.

[57] A. Zeileis, J. C. Fisher, K. Hornik, R. Ihaka, C. D. McWhite, P. Murrell, R. Stauffer, and C. O. Wilke. colorspace: A toolbox for manipulating and assessing colors and palettes. *arXiv preprint arXiv:1903.06490*, 2019.