Additional Captions Generated by GPT-4 for Furniture Assembly Manuals

Ryuki Maeda[1], Maria Larsson[2], and Hironori Yoshida[1(✉)]

[1] Future University Hakodate, 116-2 Hakodate, Japan
hyoshida@hy-ma.com
[2] The University of Tokyo, Hongo7-3-1, Bunkyo, Tokyo, Japan
http://www.ma-la.com,https://www.hy-ma.com

Abstract. Furniture assembly is a creative task, yet varying styles in assembly instructions could make them difficult to follow. In this study, we explore the potential of increasing comprehension by automatically generating text descriptions for illustration-only furniture assembly manuals using a multi-modal large-language model such as a generative pre-trained transformer (GPT). Specifically, we inputted assembly illustrations into GPT-4 and generated captions for each assembly step. These captions were added to the instructions to accommodate users who prefer textual information. We conducted two user studies to evaluate the preferences for styles of manuals and the effectiveness of the additional captions generated by GPT-4. The first study revealed that the preferable length of added captions were those less than 150 words. We also learned that captions with some errors were negligible for users. Consequently, we conducted a second study in form of a comprehension test to determine whether captions with errors are still comprehensible. The results suggest that there is no significant difference in comprehension between correct and error-containing captions generated by the model.

Keywords: Furniture assembly support · LLM-support · Augmention of manuals

1 Introduction

There are many self-assembly furniture products on the market. Customers (or users) read assembly manuals that comes with manufacture-specific styles of assembly instructions. If a user is not used to an instruction style or has less experience with assembling furniture, the user might be puzzled by the way the instructions are expressed, worded, or laid out [1]. Even though some users skip reading furniture assembly manuals, and it would be ideal if it is possible users intuitively knows what to do without manuals, the role of instructions is still fundamental as it is often required that users follow a specific order of assembly procedures for a successful outcome. Shao et al. in their study [9] conducted a survey on furniture assembly instructions, confirming the growing importance

H. Degen and S. Ntoa (Eds.): HCII 2025, LNAI 15821, pp. 168–177, 2025.
https://doi.org/10.1007/978-3-031-93418-6_12

of research in this field. With this background, we formed the hypothesis that adding text to illustration-only instructions would enhance understanding, especially for those less skilled at furniture assembly. To confirm this assumption, we conducted a preliminary survey that showed that it is better to have some extent of textual information in the instruction manual compared to illustration-only manuals.

Recently, large language models (LLMs) have become multi-modal, being capable of taking image input and outputting a summary in text. We also conducted a feasibility study of GPT-4 by giving a page of a furniture assembly manual. The result showed that LLM remarkably understood the given page even though the page was detached from the whole manual and lost its context. The model was even able to list the necessary parts in the page. While it showed a degree of capability of understanding the assembly process, it tended to generate errors in the generated captions. Therefore, we decided to conduct a user study to see how much error users can accept in the additional captions of a furniture assembly manual. This study utilizes GPT-4 of ChatGPT as one of the functions of customization, and aims to generate instruction manuals with textual information from those with only diagrams, and to verify the effectiveness of this function. Furthermore, we aim to verify whether the generated captions can be tailored to the user's preferences and background.

In summary, our contributions are as follows.

- A comparative study of user preference of different styles of furniture assembly manuals in the market.
- A user study about preferences of furniture assembly instructions containing the captions generated by GPT-4.
- A user study measuring the level of understanding of the generated instructions.

2 Related Work

2.1 Augmented Furniture Assembly Manual

There are several works about augmented instruction manuals. BILT is a smartphone application that shows step by step furniture assembly instructions. Each screen corresponds to an assembly process with animation using 3D models, and users can freely swipe the steps [2]. It requires 3D models and preparing the animation for each step. Shao et al., proposed a system to take images of assembly manuals, make 3D models out of it, and animate them according to the extracted notations such as arrows [9]. AssembleAR [5] proposes an idea of an augmented reality (AR) interface for furniture assembly. It can show animations for assembly by recognizing parts. These projects require to prepare 3D models (except for Shao et al.) and a smartphone to visualize the animations.

Fig. 1. Workflow overview: An image of illustration-only furniture assembly instruction is inputted to GPT-4 and generate captions for each step. In this case, we prompted in Chat-GPT. Including the generated captions, we also prompted layouts and place text boxes accordingly.

2.2 Style Transfer for Instruction Manuals

Minotani et al. (2017) proposed a manga creation support system for instruction manuals [3]. The system allows users to manually select text and images and generate manga by choosing font size and style according to personal preferences. Manga is used in diverse fields, including education and entertainment, and it is recognized that familiarity and attachment to manga is particularly strong in Japan. Minotani et al.'s research targets instruction manuals for consumer electronics products. They constructed a system that can be easily operated by users without expertise in manga production. To evaluate their system, an experiment was conducted to compare the original instruction manuals and the generated manga to verify the usability of the system. Subjects responded to a questionnaire about their satisfaction with the generated cartoons and the usability of the system. This study is closely related to our work in that it facilitates understanding of the instruction manual. Unlike their approach which is adding illustration, our approach adds texts.

2.3 Applications of Multi-modal Learning Model

A multi-modal learning model is a type of deep learning model that integrates and relates different media data such as text and images. Soon after contrastive language-image pretraining (CLIP) [6] was released and showcased its zero-shot capabilities, generative models such as DALL-E [7], Imagen [8], and other models became available for creating images from textural input. These models are expected to capture the relationship and co-occurrence between images and text and extract information. Applications include natural language processing, image recognition, information retrieval, and multi-modal tasks. In our work, we use GPT-4 as it provides numerical representations of images and texts. GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks [4].

3 Experiments and Evaluations

We first conducted a feasibility study of GPT-4 for generating instruction texts with illustration-only inputs. We found that not all of the captions currently generated by GPT-4 are correct. And then, we conducted a study to see whether the inaccurate sentences generated by GPT-4 are useful or not. Our hypothesis is that the instruction captions generated by GPT-4 are useful as additional textual information. If this hypothesis is correct, an assembly instruction manual with instruction captions generated by GPT-4 will be more effective than the ones with only diagrams.

3.1 Feasibility Study with GPT-4

We fed an image of the assembly instructions and verified whether GPT-4 understands the image correctly or not by prompting questions. And then we prompted GPT-4 to generate assembly instructions in texts and checked how accurate they are As a result, GPT-4 successfully read the assembly instructions (the left of Fig. 1). Next, GPT-4 was prompted to explain the assembly instructions from the image, which generated the captions shown in the following steps staring from step three. The description was confirmed to be correct to some extent. However, the sentences were not completely correct, as some sentences showed GPT-4 is not certain about its understanding, such as *"probably indicate instructions for assembling furniture, etc.,"*. Also some names of parts and tools were incorrect.

Step 3 Insert the eight dowels into the indicated holes. The dowels are small wooden or plastic pins that help hold the part in place. Push or lightly tap each dowel into the hole with a hammer.
Step 4 Eight screws are used to secure the parts. A Phillips screwdriver is needed here to tighten the screws. Tightening the screws will ensure the stability and strength of the shelf.
Step 4 When assembling, cross-tighten the screws to ensure stability. When tightening the screws, be careful not to apply excessive force to avoid damaging the parts. By following these steps precisely, the assembly of the shelf unit should go smoothly. If you need more detailed instructions, please let us know.

While the feasibility study, we found that GPT-4 has some tendencies of errors in the text. We summarized these errors as summarized in the following list from F1 to F6. Despite of these errors, we found that the generated texts are still useful for additional captions, thus moved on the user studies in the next section.

F1 Name of parts tend to be incorrect.
F2 Tools used in steps tend to be incorrect.
F3 There are sentences that are completely unrelated.
F4 The beginning of sentences tend to be random.
F5 Sentence endings tend to be random.
F6 Sentence tones tend to be random.

3.2 User Study 1: Preferences on Generated Texts

The objective of the study was to understand which kinds of texts are preferred for furniture assembly manuals, and whether texts generated by GPT-4 is acceptable or not for users. The number of participants was 19: eight male and ten female. They were all university students whose ages ranged from 18 to 21 years old.

Before the study, to know the backgrounds of the participants, they were asked about their experiences in assembly, preferences of styles of manuals such as how much texts and diagrams as well as how many pages, size of texts and diagrams, orientation of manuals, and so forth. Next, we asked them to create a customized furniture manual in Miro[1] using generated captions by GPT-4 from furniture assembly diagrams. The duration of experiment was approximately 30 min.

Participants were informed that these descriptions were generated by GPT-4. There were seven steps in the instruction and we prepared four different lengths of captions for each step. We also explained what they can and cannot modify for creating manuals. What they can adjust are position, size, and width of text and diagram sections, the number of pages, and orientation (vertical or horizontal) of the instructions. Prohibited actions include the following: changing the text itself as well as its thickness and font, separating text into multiple text boxes, modifying diagrams such as trimming or adding other images.

Figure 2 shows examples of the customizable graphical elements used in the survey. The texts from assembly from step one to three were accurate. In assembly steps four and six, some descriptions contained incorrect names of parts and tools. Specifically, a Phillips screwdriver was described as a flathead screwdriver, and a screw as a dowel. In assembly steps five and seven, there were incorrect instructions. For example, there was an instruction text "fix screws" even though there was no such task in the diagram, and "fix to the wall" although it was not necessary. In particular, the description for assembly step.5 was mostly inaccurate. In addition, as the length of descriptions increases, some of them were redundant. For example, descriptions such as "by doing this, the parts will be aligned correctly, making it easier to assemble to the next step" and "hen all dowels are set correctly, please move on the next step" were not necessary to complete the assembly task.

Results and Discussion. Figure 3 shows a layout by a participant in the study. 19 participants incorporated at least one of the generated captions into their instruction layouts. Their reasons were as follows. "With a diagram, the captions with errors are acceptable" (P15); "I can tolerate some mistakes because I can see them with the diagram, even if wrong sentences are used" (P12); "I thought that an instruction with sentences is better than an instruction without sentences" (P7). Some participants also stated that they did not notice the errors in the texts. On the other hand, two participants did not put captions and

[1] Miro (https://miro.com/) is an online collaborative sketch interface.

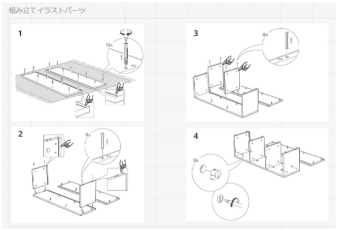

Fig. 2. An example of customizable graphical elements.

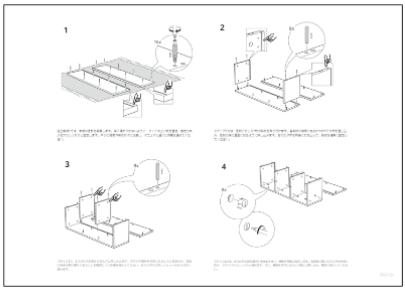

Fig. 3. An actual instructions made by a participant.

made diagram-only instructions. They described the reasons as follows: "adding captions that differed from the diagram would make it difficult to know whether to follow the diagram or the captions" (P7).

Based on the above, we considered that the captions produced by GPT-4 can be used as a supplement to the diagrams. We found that the preferable length of captions which is less than 150 characters according to the questionnaire. However, no correlation was found between length of text and level of expertise; all participants prefer about the same length of text. Also, there was no correlation between the length of the text and the number of pages or the number of diagrams in a page.

3.3 User Study 2: Validation of Customized Manuals with Errors

We conducted another study to see how the captions generated by the GPT-4 affect the comprehension of the assembly manuals. The approach was to compare manuals with completely accurate captions to those containing inaccurate captions by GPT-4. Our hypothesis was that there is no difference in comprehension. If this hypothesis were correct, the use of GPT-4 instructions in furniture manuals would be acceptable.

The participants were all university students whose ages ranged from 18 to 21 years old. The number of participants was 32, 20 were male and 12 were female. After a questionnaire to learn about their backgrounds, we asked participants to read a prepared instruction that contains errors in the captions. Within nine steps in the instruction, odd-numbered steps were instructions with captions generated by GPT-4, and even-numbered steps were the correct instructions. By exchanging the odd and even numbered steps, we prepared two types of instructions to avoid the order effect in the user study. The instructions alternate between the one with texts generated by GPT-4 which contains some errors, and the one with correct instructions for each step. An example of the descriptions are shown in the following.

Correct description "Attach the magnetic parts to the board with screws; insert the two screws into the holes indicated and tighten them firmly with a screwdriver. Note the orientation of the board."

Incorrect description by GPT-4 "Insert the screws into the board. Insert the screws, head down, into only one of the holes indicated and tighten firmly with a screwdriver. Be careful not to insert them into the other holes."

In order to examine the influence of the text, the participants were asked how much texts they read for each step. The response time was measured as follows. Participants were supposed to press the start button on the stopwatch at the beginning of solving the questions. At the end of solving the questions, the subject presses the stop button on the stopwatch. The author performed as an experiment supervisor and recorded duration. This procedure is performed for each question, and the time is measured. Figure 4 shows how the user study was conducted.

Fig. 4. A scene of the user study 2.

The comprehension quiz consists of ten questions, including multiple-choice, true-false, and questions (Q2 and Q5) asking participants to pick actual parts (physical parts questions). In order to do so, we prepared a real piece of furniture in this study based on the lessons learned from the previous experiment. We recorded the results of quiz and response time.

Results. Figure 5 shows the percentage of test questions answered correctly (correct answer rate) and Fig. 6 shows the duration taken to solve the test questions. The correct answer rate with correct captions was over 90% except for Q6, where the percentage of correct answers for Q6 was as high as over 80%. On the other hand with captions generated by GPT-4, it was over 70% for Q1, Q2, Q5, and Q10, while the other questions were less than 50%. There was no significant difference in response duration in Q2, Q4, and Q5, however, the other questions with captions by GPT-4 took longer.

Discussions. As a result, the captions by GPT-4 did not provide the expected results, as there were differences both in correct answer rate and response duration compared to the correct captions. However, since Q2 and Q5 are physical

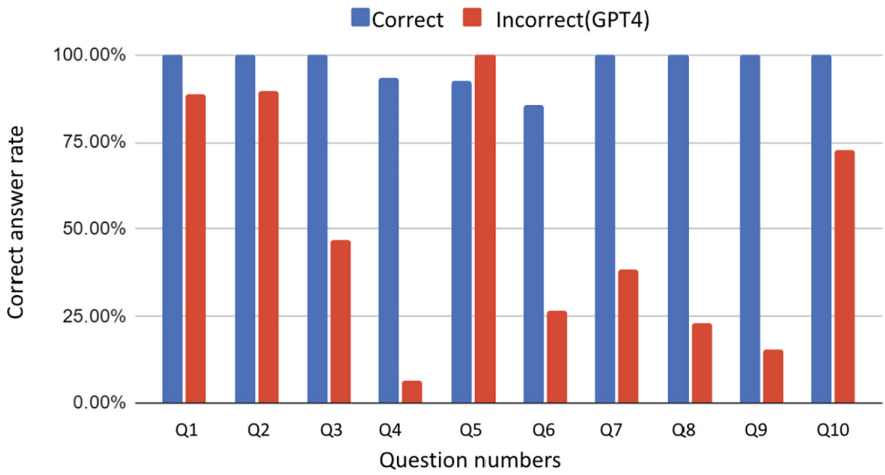

Fig. 5. Correct answer rates between correct captions and captions generated by GPT-4 with some errors.

parts questions that the participants had to actually pick the parts, no significant difference was observed between the two questions. Observing how participants solve the questions, they all double-checked by touching the actual furniture parts, which might help their understanding about the question even though the captions contained errors. On the other hand, Q7 was a similar question, but there was a large difference in both correct answer rate and response duration. The reason could be that the diagrams of the parts used in Q7 were somewhat difficult to understand. Therefore, it can be said that the description of GPT-4 is not a problem if the following two conditions are in consideration. Based on these results, a new hypothesis is that users largely depends on information they get through interacting with physical parts which are clearly visible in instructions.

In this experiment, we did not inform users that the captions might be incorrect, so those who relied only on the captions would have made a mistake. In particular, the correct answer rates for Q6, Q8, and Q9 were low because the participants saw the same sentences in the choice questions and the captions. Therefore, it can be considered that if the participants were informed in advance that there might be errors, they would correct the captions by double-checking the diagram.

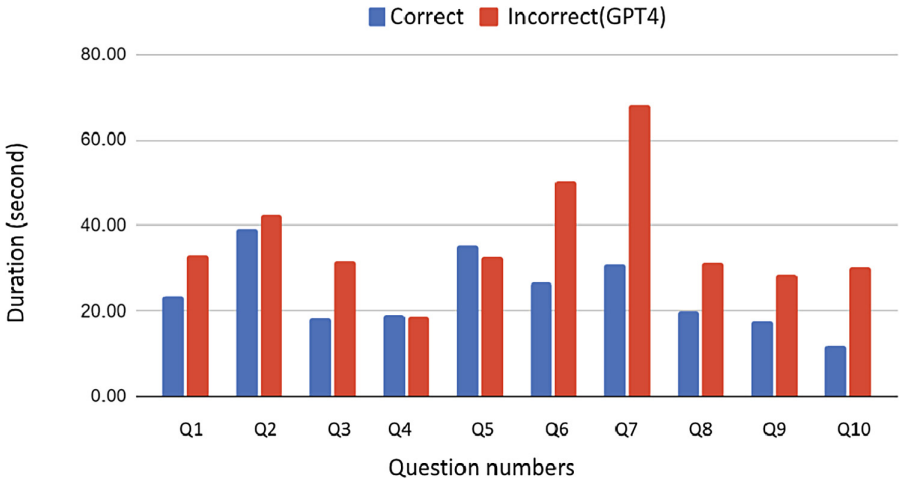

Fig. 6. Response duration between correct captions and captions generated by GPT-4 with some errors.

4 Conclusion and Future Work

In this study, GPT-4 was utilized as one of the functions to customize manuals, and its effectiveness was verified by generating manuals containing textual information from manuals containing only diagrams. We also verified the customization of manuals based on user attributes. As a result, we found that GPT-4 can generate explanatory text, but we could not confirm whether it is suitable for user attributes; since we could not confirm the effectiveness of GPT-4, it is difficult to incorporate it into actual instruction manuals at present.

In the future, we plan to conduct experiments in which users assemble furniture from scratch using the explanatory text generated by GPT-4. It is also possible to conduct experiments to see if the degree of influence of the explanatory text generated by GPT-4 changes depending on the difficulty level of the furniture to be assembled. From this, we believe that it is necessary to examine the indicators of layout as well. We believe that there is a possibility to customize user-preferred instruction manuals by combining an automatic layout generation system for graphic design magazines [10] and the surveyed indicators. Based on the above, we will continue to examine various perspectives in promoting understanding of assembly manuals.

Acknowledgments. This research was supported by Future University's Special Research Grant as well as Northtec Foundation Research Grant. We also thank to all the participants who joined the user studies.

References

1. Consumer Affairs Department, B.o.L., Culture, T.M.G.: Investigation report on the safety of assembled furniture (2015). https://www.shouhiseikatu.metro.tokyo.lg.jp/anzen/test/documents/houkokusyo.pdf. Accessed 10 Feb 2024
2. Incorporated, B.: Bilt app. https://biltapp.com. Accessed 10 Apr 2024
3. Minotani, A., Hagiwara, M.: Manga generation support system from instruction manuals. Jpn. Soc. Kansei Eng. J. **16**(1), 121–130 (2017)
4. OpenAI: Gpt-4: Generative pre-trained transformer 4 (2023). https://openai.com/research/gpt-4. Accessed 10 Apr 2024
5. Pickard, A.: Assemblear (2018). http://adampickard.com/AssembleAR. Accessed 10 Apr 2024
6. Radford, Aet al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
7. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
8. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
9. Shao, T., Li, D., Rong, Y., Zheng, C., Zhou, K.: Dynamic furniture modeling through assembly instructions. ACM Trans. Graph. **35**(6) (2016). https://doi.org/10.1145/2980179.2982416
10. Tabata, S.: Automatic layout generation system for graphic design magazines using ai. Unisys Techn. Rev. = Unisys Technol. Rev. **40**(1), 71–81 (2020)