The Illusion of Competence: Evaluating the Effect of Explanations on Users' Mental Models of Visual Question Answering Systems

Judith Sieker[†], **Simeon Junker**[†], **Ronja Utescher**[†], **Nazia Attari**[†],
Heiko Wersing[‡], **Hendrik Buschmeier**[‖], **Sina Zarrieß**[†]
[†]Computational Linguistics, Department of Linguistics, Bielefeld University
[‡]Honda Research Institute Europe
[‖]Digital Linguistics Lab, Department of Linguistics, Bielefeld University

Abstract

We examine how users perceive the limitations of an AI system when it encounters a task that it cannot perform perfectly and whether providing explanations alongside its answers aids users in constructing an appropriate mental model of the system's capabilities and limitations. We employ a visual question answer and explanation task where we control the AI system's limitations by manipulating the visual inputs: during inference, the system either processes full-color or grayscale images. Our goal is to determine whether participants can perceive the limitations of the system. We hypothesize that explanations will make limited AI capabilities more transparent to users. However, our results show that explanations do not have this effect. Instead of allowing users to more accurately assess the limitations of the AI system, explanations generally increase users' perceptions of the system's competence – regardless of its actual performance.

1 Introduction

Machine learning-based technologies (often called 'artificial intelligence', AI) are now commonly being deployed and used in real-world applications, influencing human decision-making (or automating decision-making altogether) with implications for societies, organizations, and individuals. Despite continuous advances and impressive performance on many tasks, these technologies are not always accurate and will likely never be. Machine learning models depend on curation of the data they are trained on, they are optimized according to criteria that may not do justice to the complexity of reality, and the context in which they are used cannot be fully modeled, to name a few reasons for their limitations. In addition, the underlying algorithms themselves have inherent weaknesses. Large language models (LLMs), e.g., are well known to hallucinate, i.e., to make predictions that are inconsistent with facts or themselves (Ji et al., 2023), or to be highly sensitive to spurious variations in their inputs/prompts (Sclar et al., 2023).

Many machine learning models also suffer from their own complexity: consisting of millions, billions, or even trillions of parameters, they are black-boxes, opaque to human understanding. However, in order to reliably use machine learning models and AI systems based on such models, human users must be able to assess their limitations and deficiencies, and to understand the decisions that such systems make and why (codified, for example, as the right "to obtain an explanation of the decision reached" in the legal framework of the General Data Protection Regulation of the European Union; GDPR, 2016, Recital 71). Research in Explainable AI (XAI) addresses this need, and recent years have seen an explosion of explainability methods that aim to make the internal knowledge and reasoning of AI systems transparent and explicit, and thus interpretable and accessible to users. Explainability of model predictions is thus seen as a solution, and it is assumed that they enable users to construct functional 'mental models' (Norman, 1983) of AI systems, i.e., models that closely correspond to the actual capabilities of the systems.

Whether this is the case is an active research question and there is evidence that explainability comes with new challenges. Important questions in XAI are what actually makes a good explanation, which criteria it needs to satisfy, and how the quality of explanations can be measured (Alshomary et al., 2024). Furthermore, recent perspectives emphasize that explanations should be social (Miller, 2019) and constructed interactively, taking into account the user's explanation needs (Rohlfing et al., 2021). Jacovi and Goldberg (2020) argue that evaluations of explanations should carefully distinguish plausibility (does it seem plausible to users) and faithfulness

Code and data of study are available at https://doi.org/10.17605/OSF.IO/4KDB5 or https://github.com/clause-bielefeld/IllusionOfCompetence-VQA-Explanations.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19459–19475
November 12-16, 2024

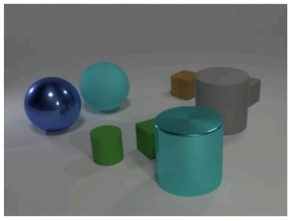
(does it reflect the model's internal reasoning) and that non-faithful, but plausible, explanations can be dangerous in that they let users construct faulty, and eventually dysfunctional, mental models that can lead to unwarranted trust (Jacovi et al., 2021).

In this paper, we investigate the effects of providing natural language explanations on users' mental models of an AI system in terms of its capabilities, and whether these explanations allow them to diagnose system limitations. We present the results of a study in the visual question answering and explanation (VQA/X) domain, artificially inducing a simple limitation by providing two VQA/X systems with images stripped of color information, i.e., in grayscale (see Figure 1). Participants, unaware of the manipulation, see the unmanipulated full color image, the question, the system's answer, and its explanation for the answer, and have to judge various system capabilities (including its ability to recognize colors) and its competence. This visual domain does not require participants to understand the internal processes of the system but should still enable them to estimate what it can and cannot do. The comparison of judgments to responses to non-manipulated system input and judgments of responses without explanations sheds light on participants' difficulties in using (natural language) XAI explanations to build accurate mental models, even for such a simple case. This raises the question of how effective explanations can be in real-world applications of XAI technology that involve more complex reasoning and problems.

Question: What season is it?

NLX-GPT (color):
ANSWER: summer EXPL: because the grass is green and the trees are casting shadows

NLX-GPT (greyscale):
ANSWER fall EXPL: because the trees are green and the grass is dry



Question: Are there more large yellow shiny objects than metal things?

NLX-GPT (color):
ANSWER: no EXPL: because there are no large yellow shiny things and there is a metal cylinder

NLX-GPT (greyscale):
ANSWER yes EXPL: because there is a large yellow shiny cylinder and there are two metal spheres and a metal cube

Figure 1: Items from our study: Answers and explanations generated with NLX-GPT for color/grayscale images in VQA-X (top) and CLEVR-X (bottom). Explanations in the grayscale condition refer to colors that were not available in the system inputs (*green, yellow*).

2 Background

Our work is related to previous studies that have examined whether explanations enhance users' trust in AI systems. Kunkel et al. (2019), for example, compared trust in personal (human) versus impersonal (recommender system) recommendation sources and examined the impact of explanation quality on trust. Their results showed that users rated human explanations higher than system-generated ones and that the quality of explanations significantly influenced trust in the recommendation source. Bansal et al. (2021) investigated whether explanations help humans anticipate when an AI system is potentially incorrect. They used scenarios where an AI system helps participants to solve a task (text classification or question answering), providing visual explanations (highlighted words) under certain conditions. Their findings revealed that explanations increased the likelihood of the participants to accept the AI system's recommendations, irrespective of their accuracy. Thus, rather than fostering appropriate reliance on AI systems, explanations tended to foster blind trust. Similarly, (Kim et al., 2022) conducted a large-scale user study for visual explanations, showing that these do not allow users to distinguish correct from incorrect predictions. Dhuliawala et al. (2023) investigated how users develop and regain trust in AI systems in human–AI collaborations. They found that NLP systems that confidently make incorrect predictions harm user trust, and that even a few incorrect instances can damage trust, with slow recovery. While these studies evaluate the influence of system explanations on users' trust in the system's output (a proxy for its perceived competence), they do not investigate users' understanding of the systems' reasoning processes and capabilities. In our study, we specifically address this issue and investigate the users' mental model of the systems' capabilities and limitations.

While the studies above found that nonverbal explanations can be misleading to users, natural language explanations are assumed to be more transparent or less difficult to interpret (Park et al., 2018; Salewski et al., 2022). Verbal explanations also offer the advantage that they can be collected from humans, which has led to the development of explanation benchmarks, particularly in multimodal domains (Kayser et al., 2021; Salewski et al., 2022). Thus, the dominant approach to verbal explanation generation currently is to leverage human explana-

tions during model training (Park et al., 2018; Wu and Mooney, 2019; Kayser et al., 2021; Plüster et al., 2023; Sammani and Deligiannis, 2023). While Lyu et al. (2024) discuss potential faithfulness issues related to supervising explanation generation with human explanations, we are not aware of work that explicitly tests these supervised models in a user-centered setting similar to ours.

3 Approach

We conduct a study to investigate how users of an AI system perceive its limitations when it encounters tasks that it cannot perform perfectly. We aim to investigate whether providing explanations alongside model responses helps users build an appropriate mental model of the AI system's capabilities and limitations. In the following, we detail the rationale of our approach (Section 3.1–3.3) and the hypotheses of our experiments (Section 3.4).

3.1 How Can We Test Users' Mental Models?

A key challenge for our study is that many modern AI systems are used in complex tasks that involve many interdependent capabilities simultaneously. This makes it difficult to isolate specific systems' capabilities and to establish or control which limitations they have, even for the developers of a system. Indeed, common evaluation protocols in NLP mostly report overall system performance according to holistic metrics (e.g., accuracies) and rarely involve a detailed assessment of specific errors or capabilities (cf. van Miltenburg et al., 2021). However, to assess whether users' perception of a system is accurate, we need to have as much control as possible over its capabilities and limitations.

To address this challenge, our study adopts an experimental setting that simplifies some aspects of testing XAI in complex tasks. First, we focus on two well-studied VQA tasks, including a synthetic VQA task in which system capabilities are relatively easy to distinguish (Section 3.2). Second, to gain at least some control over the VQA systems' limitations, we systematically manipulate their inputs in the dimension of color (Section 3.3). Third, we design a questionnaire for users to judge specific aspects of the system's capabilities. This allows us to measure whether users can diagnose which capabilities of the system have been perturbed through our explicit input manipulations (Section 3.4). The design of our study is summarized in Figure 2 and will be explained in detail below.

3.2 VQA Task and Abilities

We employ a visual question answering and explanation task: the input to the AI system is an image and a question in natural language, and its task is to generate an answer and a natural language explanation that justifies the answer. We select a visual question-answering setting as it is a rather simple task for humans and, at the same time, a task that involves distinguishable semantic-visual reasoning capabilities. This is important for our setting since we want to test whether users can differentiate specific system capabilities, based on generated explanations. Thus, inspired by Salewski et al.'s (2022) CLEVR-X benchmark for explainable VQA, we assume that these capabilities involve the abilities to process objects' (i) **color**, (ii) **shape**, (iii) **material**, and (iv) **scene** composition (e.g., spatial relations, relative size). In our study participants are asked to rate the AI system's capabilities along these four dimensions, next to other, more general criteria for competence and fluency (see Figures 8 and 9 in Appendix A.4). In the CLEVR-X benchmark, these dimensions are given by construction: the visual scenes are synthetically generated and composed of objects defined by attributes for color, material, and shape. The corresponding questions explicitly relate to one or multiple of these dimensions. In real-world image benchmarks, such as VQA-X (Park et al., 2018), these abilities are often more implicit, but still highly relevant (see examples in Figure 1). We run our study on items from both benchmarks.

3.3 Color vs. Grayscale Input

Our goal is to investigate whether explanations help users in diagnosing system limitations. To introduce these limitations in a controlled way, we manipulate the input of the VQA systems. Out of the four VQA capabilities explained above (color, shape, material, and scene), the color dimension lends itself to straightforward manipulation. During inference, systems either receive the image (i) in full color or (ii) in grayscale. This induced limitation resembles a situation where a multimodal AI model was trained on colored images but, at run-time, a camera/visual sensor is broken such that model inputs are perturbed. To ensure that this manipulation induces an incorrect model response, we only include items correctly answered with the full color image input *but* incorrectly answered with the grayscale image input. This item selection accounts for the fact that

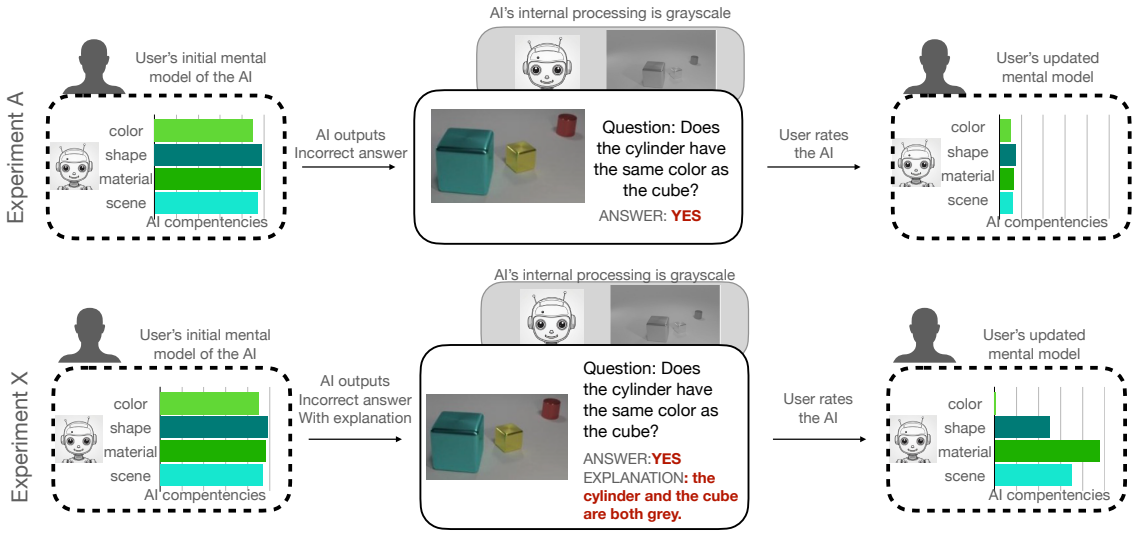


Figure 2: Illustration of our experimental design and hypotheses. In Exp.A, we do not expect users to spot the system defect (no color recognition due to grayscale input) since only answers are provided. In Exp.X, the system provides explanations which should help users in building a better mental model.

VQA models can be assumed to have further limitations that we cannot explicitly control for and exclude items (i) where the VQA does not generate the correct ground-truth answer for the colored image, and (ii) where the VQA generates the correct answer for the grayscale image. This gives us a clean set of items where the limitations of the AI system can be attributed to a particular error source. The participants in our study were unaware of the underlying color–grayscale manipulation: they saw images in color, along with the models' answers and explanations. Our goal was to determine whether participants were able perceive the limitations of the model, i.e., whether they could identify the system's lack of color recognition ability. See Figure 2 for an illustration of this set-up.

3.4 Experiments A and X

To investigate the effect of providing generated explanations alongside the system answers, we conduct two separate studies: In Experiment X, participants were shown both the answer and its explanation, whereas in Experiment A participants were shown only the answer without an explanation. In both studies, we ask participants to rate each item for the system's capabilities (color, shape, material, scene), the overall system competence, answer correctness, the consistency of answer/explanation, the consistency of explanation/image, and the explanation's fluency.

Importantly, participants in both Experiments A and X received mixed sets of items from all systems, data sets, and color conditions, and we collected judgments for each item. In this way, we wanted to prevent them from becoming "conditioned" to a particular setting, i.e., getting used to certain ways of answering or explaining and becoming overly sensitive to changes in patterns.

If explanations lead users to build more appropriate mental models, participants should, generally speaking, be able to differentiate items where systems processed grayscale vs. full color images. We approached this broad expectation with five hypotheses specific to our set-up (see Table 2 for a brief summary). First, hypotheses $H1_A$ and $H1_X$ relate to the differences in competence scores between color and grayscale conditions. Here, we expect that explanations help participants to differentiate between different system capabilities.

$H1_A$ In Exp.A, competence and all capability scores are lower in the grayscale condition than in the color condition.

$H1_X$ In Exp.X, competence and color capability scores are lower in the grayscale condition than in the color condition, but other capability scores are more stable.

Hypotheses $H2_A$ and $H2_X$ are concerned with the comparison between individual competence scores in the grayscale condition. Again, explanations

should help users to identify system deficiencies.

H2$_{A}$ In the grayscale condition of Exp.A, participants give similar scores for all capabilities.

H2$_{X}$ In the grayscale condition of Exp.X, participants rate the color capability lower relative to the other capabilities.

Hypothesis H3$_{A/X}$ pertains to the comparison of competence scores between Exp.A and X. If explanations make defects in color processing transparent, grayscale inputs should specifically affect scores for this dimension.

H3$_{A/X}$ In Exp.X the overall competence is rated higher than in Exp.A. In Exp.X, color competence is rated lower or the same as in Exp.A.

4 Experimental Setup

4.1 Data

We use two datasets in our study: VQA-X (Park et al., 2018) and CLEVR-X (Salewski et al., 2022). VQA-X is extensively utilized in Visual Question Answering (VQA) tasks, as an extension of the well-established Visual Question Answering v1 (Antol et al., 2015) and v2 (Goyal et al., 2017) datasets. The images within VQA-X originate from MSCOCO (Lin et al., 2015), and the questions are open-ended (see Figure 1, top). The style of the ground-truth explanations in VQA-X varies widely, ranging from simple image descriptions to detailed reasoning (Salewski et al., 2022).

CLEVR-X expands the synthetic dataset CLEVR (Johnson et al., 2017), incorporating synthetic natural language explanations. Each image in the CLEVR dataset depicts three to ten objects, each possessing distinct properties including size, color, material, and shape (see Figure 1, bottom). For each image–question pair in the CLEVR dataset, CLEVR-X contains multiple structured textual explanations. These explanations are constructed from the underlying scene graph, ensuring their accuracy without necessitating additional prior knowledge.

4.2 Models

For each dataset, we used two vision and language models: (i) NLX-GPT (Sammani et al., 2022) and PJ-X (Park et al., 2018) for VQA-X, and (ii) NLX-GPT and Uni-NLX (Sammani and Deligiannis, 2023) for CLEVR-X[1]. We did not use vanilla generative AI systems (such as ChatGPT) in this study, as we wanted to investigate models that were specifically constructed to provide explanations alongside their outputs.

NLX-GPT is an encoder–decoder model, which combines CLIP (Radford et al., 2021) as the visual encoder with a distilled GPT-2 model (Radford et al., 2019). Importantly, this model jointly predicts answers and explanations, i.e., it generates a single response string of the form "the answer is <answer> because <explanation>", given a question and image. For VQA-X, we use the model from Sammani et al. (2022), which is pre-trained on image-caption pairs and fine-tuned on the VQA-X data. For CLEVR-X, we use the published pre-trained weights and fine-tune the model on this dataset. Uni-NLX relies on the same architecture as NLX-GPT, but the model is trained on various datasets for natural language explanations (including VQA-X), to leverage shared information across diverse tasks and increase flexibility in both answers and explanations. We take the trained model from Sammani and Deligiannis (2023) and fine-tune it on CLEVR-X. While NLX-GPT and Uni-NLX generate answers and explanations simultaneously, the PJ-X model takes a two-step approach. It first predicts the answer with an answering model and, subsequently, generates visual and textual explanations based on the question, image, and answer[2].

For each model, we utilize the recommended model weights and fine-tune them on the two datasets. During fine-tuning, we supply each model with the original, i.e., full color images along with the questions, answers, and explanations for both datasets. During inference, images are presented in color alongside the question, or in grayscale.

4.3 User Study

We conducted the study online, using Prolific, and obtained ratings from 160 participants (80 each in Exp.A and X) who were native English speakers with normal color vision (selected using Prolific's filters). In both experiments, we utilized identical experimental items, differing only in the presence or absence of explanations. All items consisted of instances where the model provided correct answers for colored images and incorrect answers

[1]We tried to obtain model outputs from other explainable VQA-X models such as, e.g., OFA-X (Plüster et al., 2023), FME (Wu and Mooney, 2019), or e-UG (Kayser et al., 2021), but encountered significant reproducibility issues: code was unavailable or not running, authors were unavailable to provide model outputs, etc.

[2]We could not replicate Salewski et al.'s (2022) PJ-X results on CLEVR-X, and the authors could not provide model outputs. Therefore, we only report PJ-X on VQA-X.

for grayscale images. We selected a total of 128 items, evenly distributed across the datasets and models, comprising 64 for each dataset and 32 for each model, equally split between 16 colored and 16 grayscale items (for NLX-GPT, a total of 64 items were selected, with 32 items from CLEVR-X and 32 items from VQA-X). The items were distributed over four experimental lists, with each participant evaluating 32 individual items. We gathered 2560 judgments per experiment and 5120 overall.

We designed the evaluation as a rating task. We informed participants that we are assessing an AI system's ability to answer questions about images (and, for Exp.X, to generate explanations). The image, question, and answer for each item were presented at the top of the page, and, in Exp.X, the generated explanation was displayed below the answer. Each item had several questions and statements for the participants to assess. First, they were asked to evaluate the correctness of the answer. In Exp.X, participants were further asked to assess whether the explanation was (i) consistent with the answer, (ii) consistent with the picture, and (iii) overall fluent. Additionally, participants in both experiments were asked to judge whether they believed that the AI system correctly identifies (iv) shapes, (v) colors, and (vi) materials, as well as whether it (vii) understands the general scene in the image. Finally, (viii) participants judged the overall competence of the system. Participants indicated their agreement on five-point Likert scales, ranging from 1 ('strongly disagree') to 5 ('strongly agree'). For each criterion, we also offered the option of selecting "I don't know". Before providing ratings, participants received instructions and viewed an example item illustrating the evaluation criteria. They were paid at a rate of £9.00 per hour. See Appendix A.3 for example trials of the experiment.

5 Results

We organize the discussion of results based on the hypotheses outlined in Section 3. Since we ask whether explanations help participants determine that the systems could not recognize color, the following discussion concentrates on the grayscale condition and the differences between the grayscale and color conditions (see Appendix A.3 for detailed results of the color condition).

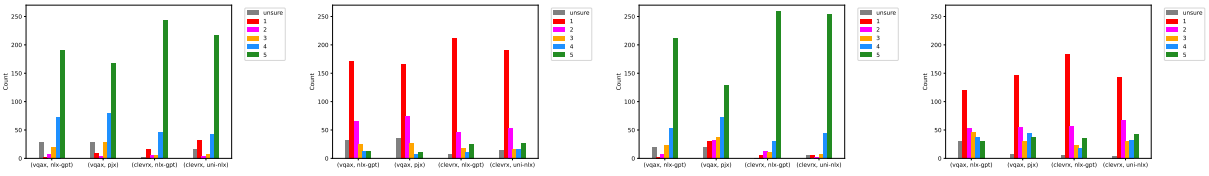
All systems received high ratings in all competency and capability dimensions when tested in the color condition of Exp.A and X, on both datasets (see Table 9 in Appendix A.3). These ratings decreased in very similar ways in the grayscale condition. Therefore, we were able to use all items from all systems to test our hypotheses, generalizing over minor system differences. We discuss differences between datasets and models in Appendix A.3, since these were not essential for testing our hypotheses. See Table 2 for summaries of hypotheses and results.

5.1 Hypotheses $H1_A$ and $H1_X$

Hypotheses $H1_A$ and $H1_X$ state our expectations on distinctions between the grayscale and color conditions in Exp.A and X, respectively. Figure 3 shows the distribution of participant ratings for the AI system's ability to recognize colors, for the grayscale and color conditions in both experiments (see Figures 4, 5, 6, and 7 in Appendix A.3 for results on the other capabilities). In Exp.A and X, there is a consistent trend of better assessments when systems have been seen the color images compared to grayscale images, across different systems, datasets, and all capabilities. Most users rate the color capability with the highest rating in the color condition (Figure 3a/c) and with the lowest rating in the grayscale condition (Figure 3b/d). The same holds for all other capabilities and competency (Figures 4, 5, 6, and 7). This confirms hypothesis $H1_A$, i.e., ratings for all capabilities decrease when the system does not see color. However, this does not support $H1_X$, as we expected that only overall competence and capability to recognize colors would be rated lower in the grayscale condition when explanations were given, and not all capabilities. This suggests that the AI's explanations did not help users diagnose the system's limitation in the grayscale condition, as all capability dimensions are similarly affected in Exp.X.

5.2 Hypotheses $H2_A$ and $H2_X$

Hypotheses $H2_A$ and $H2_X$ state our expectations for the grayscale condition. Table 1 presents the human evaluation results in Exp.A and X. Starting with Exp.A, Table 1 shows that all evaluation criteria in the grayscale condition receive relatively low scores. Interestingly, the manipulated capability, i.e., to recognize colors, does have slightly worse ratings than the other criteria (for most models and datasets). This outcome does not align with our expectation ($H2_A$) as participants in Exp.A solely viewed the answers without access to explanations, making it difficult to discern which specific abil-



(a) Exp.A – colored images (b) Exp.A – grayscale images (c) Exp.X – colored images (d) Exp.X – grayscale images

Figure 3: Human ratings on the evaluation criterion "Ability of the AI to **recognize colors**". Participants indicated their judgment on a scale from 1 (strongly disagree; here in red) to 5 (strongly agree; here in green).

			Colors		Shapes		Materials		General Scene		Competency	
Experiment	Dataset	Model	med	mean	med	mean	med	mean	med	mean	med	mean
Exp.A	CLEVR-X	NLX-GPT	1.0	1.69	1.0	2.08	1.0	1.94	1.5	1.97	1.0	1.68
		Uni-NLX	1.0	1.84	2.0	2.31	1.0	2.11	2.0	2.16	1.0	1.91
	VQA-X	NLX-GPT	1.0	1.73	2.0	2.23	1.0	1.71	1.0	1.87	1.0	1.64
		PJ-X	1.0	1.71	2.0	2.08	1.0	1.74	1.0	1.83	1.0	1.60
Exp.X	CLEVR-X	NLX-GPT	1.0	1.93	3.0	2.95	2.0	2.62	2.5	2.61	2.0	2.13
		Uni-NLX	2.0	2.27	3.0	2.89	3.0	2.82	2.0	2.61	2.0	2.21
	VQA-X	NLX-GPT	2.0	2.36	3.0	2.70	2.0	2.32	2.0	2.29	2.0	1.96
		PJ-X	2.0	2.25	2.0	2.53	2.0	2.32	2.0	2.23	2.0	1.88

Table 1: Human ratings on system capabilities for the **grayscale condition** of Exp.A (no explanations) and Exp.X (with explanations), as median and mean scores across raters.

ity or (limitation) influenced the model's answer. Results from Mann-Whitney U tests (see Table 4 in Appendix A.2) show significant differences between the ability to recognize colors and the ability to recognize other criteria for Exp.A (except for the models' overall competence), contradicting hypothesis ($H2_A$). This suggests that users in Exp.A were able to interpret incorrect system answers more than we expected. For Exp.X, the results in Table 1 suggest a very similar trend to Exp.A: the ability to recognize colors is rated slightly lower than the other capabilities. The Mann-Whitney U tests for Exp.X (reported in the lower part of Table 4 in Appendix A.2), again confirms significant differences between the perceived ability to recognize colors and the other abilities (except the systems' overall competence). Looking at Exp.X in isolation, these results seem to speak in favor of our hypothesis $H2_X$: users were indeed able to diagnose the system defect, at least to some extent. However, in light of our findings on $H2_A$, these results have to be interpreted with care: even without model explanations, users rated the color capability lower than others. This trend is a bit stronger in Exp.X but, overall, the differences between perceived capabilities are still rather small. The strongest expected trend in favor of $H2_X$ can be found for NLX-GPT on the CLEVR-X data: here, the median if the color rating is 1.0 and 3.0 or 2.0 for the other capabilities. For the other combinations of models and datasets in Exp.X, there is no clear difference in the median ratings for the perceived capabilities. We conclude that there is weak evidence in favor of $H2_X$, as explanations do not substantially improve users' assessments of system capabilities.

5.3 Hypothesis $H3_{A/X}$

Hypothesis $H3_{A/X}$ states our expectations regarding the differences between Exp.A and X for overall competency and color recognition ability.

Once again, consider Table 1. As expected, in Exp.A, i.e., without explanations, the overall competency of the models was rated low (with median values of 1.0 only). In Exp.X, although the values remain low at 2.0, there is a noticeable improvement relative to Exp.A. Thus, despite the answers being incorrect, the addition of the models' explanations enhances the perception of the models' overall competency. This could suggest that the explanations reveal other capabilities of the models, consistent with our hypothesis $H3_{A/X}$. However, contrary to $H3_{A/X}$, we also see a general increase in the ratings for the systems' color recognition ability in Exp.X compared to Exp.A. We expected that the explanations would make the color limitation explicit, which would result in color ability being rated worse or at least as poorly as in Exp.A. This also holds for *all* other model capabilities: all capability ratings

$H1_A$	competence and all capabilities rated lower in grayscale cond. than in color cond. in Exp.A	✓
$H1_X$	competence and color capability rated lower in grayscale cond. than in color cond. in Exp.X	✗
$H2_A$	similar ratings for color compared to other capabilities, in grayscale cond. in Exp.A	✗
$H2_X$	lower ratings for color compared to other capabilities, in grayscale cond. in Exp.X	(✓)
$H3_{A/X}$	competence rated higher for grayscale cond. in Exp.X than in Exp.A, color rated lower	(✓/✗)

Table 2: Overview of the validity of the hypotheses formulated in Section 3.

are comparatively higher in Exp.X than in Exp.A (even if lower than in the color condition). This observation is supported by the Mann-Whitney U tests (see the upper part of Table 4 in Appendix A.2), which show significant differences between Exp.A and X for all evaluation criteria. This suggests that users rate all system capabilities significantly higher when explanations are provided. From this we conclude that, instead of making systems' limitations more transparent, the explanations contribute to an overall more positive perception of the system, regardless of its capabilities. In other words, the AI system's explanations seem to create an illusion of the system's competence that does not correspond to its actual performance.

5.4 Automatic Evaluation

In the VQA-X domain, automatic measures for evaluating similarity or overlap with human ground-truth explanations are commonly used (cf. Salewski et al., 2022; Sammani and Deligiannis, 2023). To assess the construct validity of a representative automatic evaluation method, we compute BERTScores, measuring the similarity of ground truth explanations from both datasets to human evaluation scores. Table 3 reports the results of the BERTscore metric, showing that they do not exhibit any notable differences between the grayscale and color conditions, which clearly contradicts the results of our human investigation. Thus, while user ratings between the grayscale and color condition are located on opposite ends on the Likert scale, BERTscores show marginal differences across the board. Yet, when comparing the two datasets, the BERTScores for the CLEVR-X dataset show improved values (in both the grayscale and color conditions), aligning with the human results from Exp.X (see Table 1 and 9 in Appendix A.3).

5.5 Summary

Table 2 provides an overview of the validity of our hypotheses. Generally, our results show that explanations do not have a desirable effect on users' assessment of the system's competency and capabilities. They do not help users construct a more accurate mental model of the system and its capabilities and limitations, but simply lead to more positive user assessment overall. Our results are strikingly consistent across models and datasets. Even systems fine-tuned on the CLEVR-X benchmark, where explanations were designed to systematically mention the capabilities we assessed in our study (including color), do not address these limitations. Figure 1 shows representative examples of why this might be the case: rather than avoiding color words or using incorrect colors, systems seem to be able to guess the correct color from the question or the general context (e.g., *green* in the context of *tree*). This behavior is well-known in multimodal language models but should be avoided in explanation tasks since it counteracts transparency and appropriate user assessment.

		BERTScore	
Dataset	Model	color	grayscale
CLEVR-X	NLX-GPT	0.76	0.74
	Uni-NLX	0.75	0.74
VQA-X	NLX-GPT	0.72	0.72
	PJ-X	0.71	0.70

Table 3: BERTScores for explanations by condition.

6 Discussion of Implications

It is still not well understood how XAI can bridge the gap between highly complex black-box models with largely opaque internal reasoning processes and users' intuitive understanding of these. Generally, our study provides evidence that explanations generated by state-of-the-art systems do not always lead to the expected effects of XAI and that explanations may even further obstruct AIs' reasoning processes and trick users into believing that the AI is more competent than it actually is. This result is particularly noteworthy in light of the fact that the manipulation employed in our study introduced an obvious error that should be easy to spot for users (defects in systems' color recognition).

XAI Models Our study underlines the great importance of prioritizing faithfulness over plausibility in explanation methods (Jacovi and Goldberg, 2020). With today's AI systems and LLMs, users face the challenging situation that these systems present fluent outputs projecting confidence and competence. Yet, this confidence may not be grounded in actual system capabilities and reliability (Guo et al., 2017). Our findings suggest that this also holds, to some extent, for state-of-the-art approaches to natural language explanation generation. Looking at the architecture of these models, this is by no means surprising. At least within the domain of VQA-X, which we focused on in this paper, explanation generation approaches largely follow common language modeling architectures and prioritize generating fluent, human-like outputs. Despite the fact that the importance of faithfulness in XAI has been recognized for some time and it continues to be a challenge (Lyu et al., 2024).

Evaluation of XAI Our study also highlights the importance of evaluating explanation methods in thorough, detailed, and user-centered ways (cf. Lopes et al., 2022). In the domain of VQA-X, automatic, benchmark-based evaluations still seem to be in focus and widely accepted in the community. All systems we tested in our study have been assessed mainly in automatic evaluations (cf. Park et al., 2018; Kayser et al., 2021; Sammani et al., 2022; Sammani and Deligiannis, 2023). This stands in stark contrast to research showing that XAI evaluations often have little construct validity, i.e., do not assess the intended properties of explanations (Doshi-Velez and Kim, 2017; van der Waa et al., 2021). Our BERTscore-results lend further support to this argument.

7 Conclusion

This paper investigates the effects of providing natural language explanations on users' ability to construct accurate mental models of AI systems' capabilities, and whether these explanations allow them to diagnose system limitations. Results from two experiments show that natural language explanations generated by state-of-the-art VQA-X systems may actually hinder users from accurately reflecting capabilities and limitations of AI systems. Participants who received natural language explanations projected more competence onto the system and rated its limited capabilities higher than those who did not receive explanations.

Limitations

We identify the following limitations in our work:

The addition of further models and data sets might have provided additional insights into our experiments. Unfortunately, recently research on generating natural language explanations has not been very active. The best known approaches are models like PJ-X (Park et al., 2018) or e-UG (Kayser et al., 2021), which have older code bases with reproducibility issues. We have tried to include other models (see Section 4, footnotes 1 and 2).

For the grayscale condition, we remove color information at the inference level for models trained on colored input. An alternative approach would be altering inputs during model training, possibly leading to deficiencies that are harder to identify for participants. Similarly, other kinds of perturbations such as altering relative object sizes or scene layouts might affect different dimensions of perceived system capabilities than color recognition. Here, we focused on color, as this property is easier to control and less intertwined with other properties than, e.g., object size (which might also change how relative positions are described).

Ethics Statement

Our study focuses on user-centered evaluation of XAI systems and on understanding whether these systems fulfill the promise of making black-box AI systems more transparent for users. Therefore, we believe that our study contributes to understanding and improving the social and ethical implications of recent work in NLP, and Language & Vision. In our study, we collect ratings from Prolific users but, other than that, did not record any personal information on these users.

Acknowledgments

The first author acknowledges financial support by the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), an initiative of the Ministry of Culture and Science of the State of North Rhine-Westphalia. We also acknowledge funding and support by the Deutsche Forschungsgemeinschaft (DFG) (TRR 318/1 2021 – 438445824) and Honda Research Institute Europe.

References

Milad Alshomary, Felix Lange, Meisam Booshehri, Meghdut Sengupta, Philipp Cimiano, and Henning Wachsmuth. 2024. Modeling the quality of dialogical explanations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 11523–11536, Torino, Italy.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A diachronic perspective on user trust in AI under uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5567–5580, Singapore.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arxiv:1702.08608.

GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union*, L 119:1–88.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, Honolulu, HI, USA.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330, Syndney, Australia.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635, Virtual, Canada.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:248:1–248:38.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Honolulu, HI.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-ViL: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1224–1234, Montreal, Canada.

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. HIVE: Evaluating the human interpretability of visual explanations. In *Proceedings of the 17th European Conference Computer Vision*, pages 280–298, Tel Aviv, Israel.

Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, UK.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common objects in context. *Preprint*, arxiv:1405.0312.

Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. 2022. XAI systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12:9423.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50:1–67.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Donald A. Norman. 1983. Some observations on mental models. In Dedre Gentner and Albert L. Stevens, editors, *Mental Models*, pages 7–14. Psychology Press, New York, NY, USA.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8779–8788, Salt Lake City, UT, USA.

Björn Plüster, Jakob Ambsdorf, Lukas Braach, Jae Hee Lee, and Stefan Wermter. 2023. Harnessing the power of multi-task pretraining for ground-truth level natural language explanations. *Preprint*, arxiv:2212.04231.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, Virtual.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Katharina Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike Buhl, Hendrik Buschmeier, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13:717–728.

Leonard Salewski, A Sophia Koepke, Hendrik P A Lensch, and Zeynep Akata. 2022. CLEVR-X: A visual reasoning dataset for natural language explanations. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI – Beyond Explainable AI*, pages 69–88. Springer, Cham, Switzerland.

Fawaz Sammani and Nikos Deligiannis. 2023. Uni-NLX: Unifying textual explanations for vision and vision-language tasks. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4636–4641, Paris, France.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. NLX-GPT: A model for natural language explanations in vision and vision-language tasks. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8312–8322, New Orleans, LA, USA.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria.

Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. ACL.

Jialin Wu and Raymond Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, Florence, Italy.

A Appendix

A.1 Materials Availability Statement

We used the following public resources in our work:

- Source code for NLX-GPT is available from GitHub at https://github.com/fawazsammani/nlxgpt
- Source code for Uni-NLX is available from GitHub at https://github.com/fawazsammani/uni-nlx/
- Source code for PJ-X and VQA-X data is available from GitHub at https://github.com/Seth-Park/MultimodalExplanations
- COCO Images for VQA-X are available here: https://cocodataset.org/
- CLEVR-X data is available from GitHub at https://github.com/ExplainableML/CLEVR-X
- CLEVR images for CLEVR-X are available here: https://cs.stanford.edu/people/jcjohns/clevr/

The source code and data from our human evaluation study can be found at either of the following locations:

- https://doi.org/10.17605/OSF.IO/4KDB5
- https://github.com/clause-bielefeld/IllusionOfCompetence-VQA-Explanations.

A.2 Statistical Tests

Table 4 shows the results of Mann-Whitney U tests in the grayscale condition. The upper half of the table reports the differences in user ratings of system capabilities (color, shape, material, scene) and overall competence between Exp.A and X, all differences are highly statistically significant. The lower half of the Table reports the differences in ratings with Exp.A and X. Table 5 reports the same tests for the color condition. Here, only the difference between overall competence is statistically significant between Exp.A and X while all system capabilities are rated similarly with or without explanations. This further supports our finding that explanations enhance user's perception of system competence, regardless of the correctness of system answers.

Criterion	U-statistic	p-value
Colors	488421.0	$\mathbf{4.09 \times 10^{-15}}$
Shapes	460501.0	$\mathbf{5.81 \times 10^{-21}}$
Materials	428263.0	$\underline{\mathbf{3.06 \times 10^{-32}}}$
General Scene	457629.0	$\mathbf{3.38 \times 10^{-22}}$
Competency	464419.5	$\mathbf{3.01 \times 10^{-21}}$
Color / Shape (Exp.A)	452212.0	$\mathbf{1.64 \times 10^{-15}}$
Color / Shape (Exp.X)	506384.0	$\mathbf{4.70 \times 10^{-21}}$
Color / Material (Exp.A)	510967.5	$\mathbf{6 \times 10^{-04}}$
Color / Material (Exp.X)	548762.5	$\mathbf{3.43 \times 10^{-11}}$
Color / Gen. Scene (Exp.A)	486718.0	$\mathbf{1.70 \times 10^{-06}}$
Color / Gen. Scene (Exp.X)	557231.0	$\mathbf{4.54 \times 10^{-09}}$
Color / Comp. (Exp.A)	538178.0	0.52
Color / Comp. (Exp.X)	640143.5	0.73

Table 4: Mann-Whitney U test results for the **grayscale conditions** of Experiments A and X. In the upper part of the table, we measure whether the ratings of one evaluation criterion (e.g., the ability to recognize *colors*) of Exp.A differs significantly from the ratings of the same evaluation criterion from Exp.X. In the lower part of the table, we measure whether the ratings of the color criterion differ significantly from the ratings of the other evaluation criteria. p-values in bold indicate statistical significance ($p < 0.001$), the smallest p-value is underlined.

A.3 Additional Results

Answer Correctness First, recall that we only included cases where the models generated incorrect answers for grayscale images and correct answers for full-color images, according to ground-truth answers in the datasets. Table 6 displays frequency distributions of correctness ratings in our user study: 'no' ratings predominated in the grayscale condition, whereas 'yes' ratings were more prevalent in the color condition across both datasets. We also conducted a chi-squared test of independence on this evaluation criterion ($\chi^2 = 2.3617, df = 2, p = 0.67$), finding no statistically significant difference between Exp.A and X regarding the evaluation of the answers' correctness. These results replicate and confirm the correctness of ground-truth answers in VQA-X and CLEVR-X.

Differences between Datasets and Models If we first look at Exp.A (Table 1), only minimal distinctions are evident between datasets or models, particularly concerning the models' ability to recognize colors, materials, and their overall competency. While slight variations exist in the other evaluation criteria, none are notably remarkable. For instance, regarding their understanding of the general scene, the models exhibit slightly better performance with

Criterion	U-statistic	p-value
Colors	627628.0	0.77510
Shapes	632776.5	0.49522
Materials	606350.0	0.17573
General Scene	647675.0	0.06266
Competency	678234.5	**0.00003**
Colors / Shapes (Exp.A)	594055.5	0.23511
Colors / Shapes (Exp.X)	706324.0	0.14946
Colors / Materials (Exp.A)	626865.0	**0.00012**
Colors / Materials (Exp.X)	717614.5	**0.02390**
Colors / Gen. Scene (Exp.A)	569399.0	0.84294
Colors / Gen. Scene (Exp.X)	710226.5	0.08423
Colors / Competency (Exp.A)	572890.5	0.61815
Colors / Competency (Exp.X)	746006.5	**0.00002**

Table 5: Mann-Whitney U test results for the **color conditions** of Experiments A and X. In the upper part of the table, we measure whether the ratings of one evaluation criterion (e.g. the ability to recognize *colors*) of Exp.A differs significantly from the ratings of the same evaluation criterion from Exp.X. In the lower part of the table, we measure whether the ratings of the color criterion differs significantly from the ratings of the other evaluation criteria. p-Values in bold indicate significance ($p < 0.05$), the smallest p-values are underlined.

the CLEVR-X dataset. In Exp.X (Table 1), on the other hand, the results exhibit some more variation between models and datasets. For example, only for the models' overall competency, do we find the same (median) value across models and datasets. Overall, it also appears that the items based on CLEVR-X data perform slightly better in Exp.X, specifically in terms of the models' ability to recognize shapes and materials, as well as their general scene understanding and overall competence.

Table 7 shows the frequency of questions in the human evaluation study that contain the word "color[s]" or specific color terms like "red" or "blue" etc., categorized by dataset. It is evident that almost all questions in the CLEVR-X dataset contain color terms, with about half explicitly mentioning the word "color". Conversely, in the VQA-X dataset, only three out of 64 questions include the word "color[s]". Hence, the observed distinctions between the datasets may be attributed to this contrast.

Analysis of the Color Condition Table 9 shows the human evaluation results for the color condition in Exp. A and X. In contrast to the results of the grayscale condition (Table 1), with respect to all the evaluation criteria, the evaluation for both Exp.A and Exp.X is very good. This corresponds to our expectation because only items with correct model answers were included in the color condition.

Furthermore, we can see that in both Exp.A and Exp.X, there are no remarkable differences between the ability to recognize colors and the other tested abilities. This is also evident from the Mann-Whitney U Test results in Table 5, especially when compared to the Mann-Whitney U results for the grayscale condition in Table 4.

However, it is notable that, with respect to all evaluation criteria, the PJ-X model receives lower ratings in Exp.X compared to Exp.A. In other words, including explanations in Exp.X results in a decline in performance for the PJ-X model. For the other models, we do not observe this difference between the two Experiments; instead, their evaluation remains fairly consistent in the color condition across both experiments. Consequently, the explanations produced by the PJ-X model seem inferior to those of the other models. This discrepancy may be due to the unique architecture of the PJ-X model, which, unlike the other models, generates answers and explanations in two separate steps rather than one.

Correlations between BERTscore and human judgments Table 10 shows Pearson's correlation coefficients (ρ) between the automatic and human evaluation metrics for the CLEVR-X and VQA-X datasets. Interestingly, we find large differences between the datasets. While all human metrics show statistically significant correlations with BERTScore for the VQA-X dataset, we find no statistically significant correlations for the CLEVR-X dataset. However, one commonality between the two datasets is the lack of differentiation between various criteria. The fact that all skills either correlate or show no correlation suggests that the automatic BERTScore metric is not able to capture the nuanced distinctions that human evaluation can discern.

A.4 Online Experiment

Figures 8 and 9 show screenshots of the study, example items and evaluation criteria.

	Exp.A			Exp.X		
Condition	no	unsure	yes	no	unsure	yes
grayscale	1129	51	99	1157	36	86
color	82	67	1131	59	48	1172

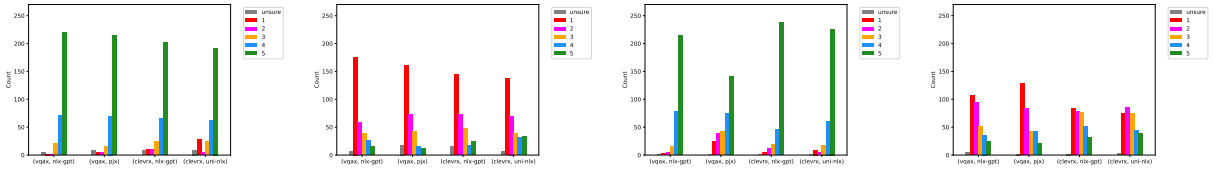
Table 6: Frequency distributions of ratings regarding correctness of system answers for Exp.A and X.

	"Color[s]" in question		Color term in question	
Dataset	yes	no	yes	no
CLEVR-X	34	30	59	5
VQA-X	3	61	3	61

Table 7: Occurrence of questions in the human evaluation study containing the word "color[s]" or specific color terms like "red" or "blue", differentiated by dataset (color terms include any instance of "color", a specific color term, or both).

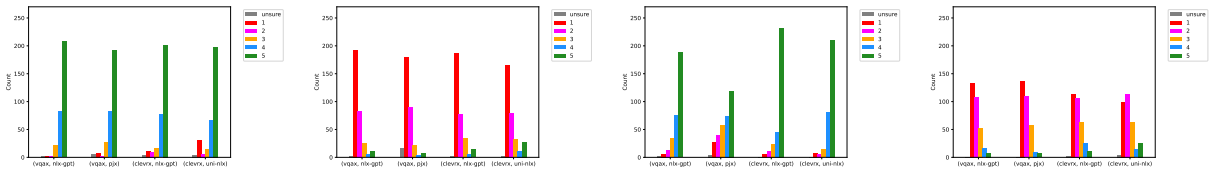
			Consist. of Expl. & Answ.		Consist. of Expl. & Img.		Fluency of Expl.	
Condition	Dataset	Model	median	mean	median	mean	median	mean
grayscale	CLEVR-X	NLX-GPT	4.0	**3.26**	1.0	1.53	4.0	3.27
		Uni-NLX	4.0	3.17	1.0	1.74	4.0	**3.46**
	VQA-X	NLX-GPT	2.0	2.67	1.0	1.85	4.0	3.42
		PJ-X	1.0	2.20	1.0	**2.02**	4.0	3.35
color	CLEVR-X	NLX-GPT	5.0	4.58	5.0	4.53	5.0	4.52
		Uni-NLX	5.0	**4.61**	5.0	**4.59**	5.0	**4.54**
	VQA-X	NLX-GPT	5.0	4.42	5.0	4.53	5.0	4.34
		PJ-X	4.0	3.56	4.0	3.63	5.0	3.85

Table 8: Human ratings for the additional evaluation criteria of **Exp.X**. We asked the participants to rate the *consistency of the explanation with the answer*, the *consistency of the explanation with the image*, and the *fluency of the explanation*. We report the median and mean scores across raters as the final scores, with bold values indicating conditions with the best (mean) values for that evaluation criteria.



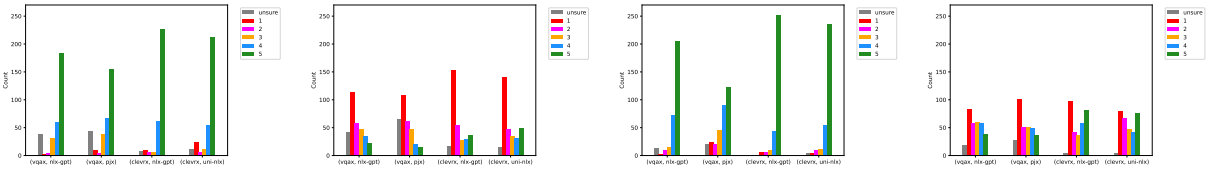
(a) Exp.A – colored images (b) Exp.A – grayscale images (c) Exp.X – colored images (d) Exp.X – grayscale images

Figure 4: Human ratings on the evaluation criterion "Ability of the AI system to **understand the general scene**". Participants indicated their judgment on a scale from 1 (strongly disagree; here in red) to 5 (strongly agree; here in green).



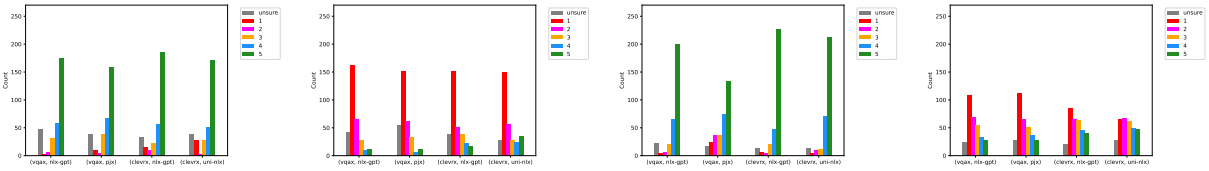
(a) Exp.A – colored images (b) Exp.A – grayscale images (c) Exp.X – colored images (d) Exp.X – grayscale images

Figure 5: Human ratings on the evaluation criterion "**Overall competency** of the AI system". Participants indicated their judgment on a scale from 1 (strongly disagree; here in red) to 5 (strongly agree; here in green).



(a) Exp.A – colored images (b) Exp.A – grayscale images (c) Exp.X – colored images (d) Exp.X – grayscale images

Figure 6: Human ratings on the evaluation criterion "Ability of the AI system to **recognize shapes**". Participants indicated their judgment on a scale from 1 (strongly disagree; here in red) to 5 (strongly agree; here in green).



(a) Exp.A – colored images (b) Exp.A – grayscale images (c) Exp.X – colored images (d) Exp.X – grayscale images

Figure 7: Human ratings on the evaluation criterion "Ability of the AI system to **recognize materials**". Participants indicated their judgment on a scale from 1 (strongly disagree; here in red) to 5 (strongly agree; here in green).

			Colors		Shapes		Materials		General Scene		Competency	
Experiment	Dataset	Model	med	mean	med	mean	med	mean	med	mean	med	mean
Exp. A.	CLEVR-X	NLX-GPT	5.0	**4.55**	5.0	**4.57**	5.0	4.34	5.0	4.43	5.0	4.47
		Uni-NLX	5.0	4.33	5.0	4.38	5.0	4.20	5.0	4.23	5.0	4.28
	VQA-X	NLX-GPT	5.0	**4.55**	5.0	4.50	5.0	**4.45**	5.0	**4.67**	5.0	**4.66**
		PJ-X	5.0	4.38	5.0	4.30	5.0	4.30	5.0	4.57	5.0	4.50
Exp.X	CLEVR-X	NLX-GPT	5.0	4.65	5.0	**4.66**	5.0	**4.58**	5.0	4.57	5.0	4.52
		Uni-NLX	5.0	**4.74**	5.0	4.61	5.0	4.56	5.0	**4.58**	5.0	**4.56**
	VQA-X	NLX-GPT	5.0	4.54	5.0	4.54	5.0	4.54	5.0	**4.58**	5.0	4.38
		PJ-X	4.0	3.80	4.0	3.86	4.0	3.84	4.0	3.86	4.0	3.71

Table 9: Human ratings on the different evaluation criteria for the **color condition** of Exp.A (i.e., no model explanations were shown to the participants) and Exp.B (i.e., model explanations were shown to the participants). For *Colors*, *Shapes* and *Materials*, we asked the participants to rate the AI system's ability to recognize the respective capability. Further, we asked the participants to rate the AI system's understanding of the *General Scene* as well as it's overall *Competency*. We report the median and mean scores across raters as the final scores. Bold values indicate conditions with the best (mean) values for that evaluation criteria.

		CLEVR-X		VQA-X	
Automatic metric	Human metric	ρ	p-value	ρ	p-value
BERTScore	Consist. of Expl. & Answ.	-0.090	0.31	0.251	**0.008**
	Consist. of Expl. & Img.	-0.020	0.82	0.278	**0.003**
	Fluency of Expl.	-0.033	0.71	0.304	**0.001**
	Shapes	-0.068	0.44	0.231	**0.02**
	Colors	-0.023	0.80	0.201	**0.04**
	Materials	-0.056	0.53	0.248	**0.009**
	General Scene	-0.051	0.57	0.251	**0.008**
	Competency	-0.051	0.57	0.252	**0.008**

Table 10: Pearson's correlation coefficient (ρ) between BERTScore results and human evaluation metrics for CLEVR-X and VQA-X data. p-values in bold indicate statistical significance ($p < 0.05$).

QUESTION:
What is the girl doing?

AI ANSWER:
flying kite

The answer is correct.

○ No ○ Yes ○ I don't know

Based on the answer, I think that the AI...

...correctly recognizes **shapes**.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

...correctly recognizes **colors**.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

...correctly recognizes **materials**.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

...understands the **general scene** in the image.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

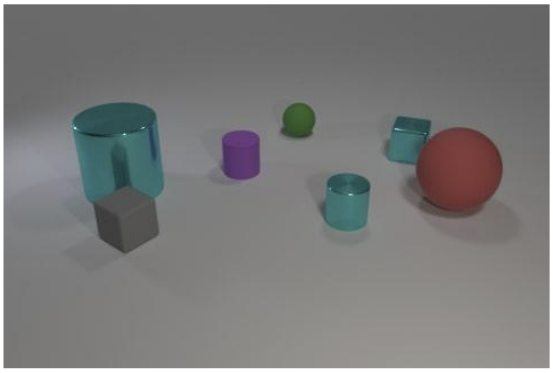
... overall is **competent**.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

Next

Progress:

Figure 8: A training item used in the online experiment to familiarize participants with the task and rating scales. This item comes from the **VQA-X** dataset and from **Exp.A**, i.e., the study without explanations.



QUESTION:
What number of gray objects are the same material as the big red ball?

AI ANSWER:
1

AI EXPLANATION:
because there is the gray rubber cube that has the same material as a big red ball

---

The answer is correct.

○ No ○ Yes ○ I don't know

The explanation is consistent with the answer.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

The explanation is consistent with the image.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

The explanation is fluent.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

Based on the answer and explanation, I think that the AI...

...correctly recognizes **shapes**.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

...correctly recognizes **colors**.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

...correctly recognizes **materials**.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

...understands the **general scene** in the image.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

... overall is **competent**.

strongly disagree ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 strongly agree ○ I don't know

Next

Progress:

Figure 9: An experimental item used in the online experiment. This item comes from the **CLEVR-X** dataset and from **Exp.X**, i.e., the experiment with explanations.