Multimodal Large Language Model (MLLM) Noise Resistance

Prasham Shah

Northern Burlington County Regional High School

Abstract

The “noise resistance” of an AI model determines its usability under non-ideal real-world conditions—in technology products, scientific research, etc.—where a few characters in text or pixels in images are often inaccurate. Prior to this study, the noise resistance of classification models and text-based large language models (LLMs) had been investigated, but the noise resistance of multimodal LLMs (MLLMs) had not. Thus, I studied MLLMs’ noise resistance against both textual noise (misspellings) and image noise (Gaussian, salt-pepper, and speckle). I also employed two denoising algorithms, spell-check (“aspell”) for textual prompts and OpenCV’s “Fast NL Means” for image prompts, to see if such pre-processing improves MLLM accuracy. I developed 10 textual prompts and 30 image-based prompts, each then noised and then denoised. I tested two MLLMs (LLaVA and GPT-4o), alongside a traditional LLM (GPT-3.5) given the textual prompts for comparison. I hypothesized that MLLMs would have poor noise resistance (even worse than traditional LLMs) and be helped by denoising algorithms. The first hypothesis was supported by the data, but the second hypothesis was refuted—traditional denoising algorithms generally hurt model performance. I also predicted, though not central to my study, that lower-parameter models would fare worse, which the data supported; however, as it was not a factor I set out to measure, future controlled studies should confirm this. Future studies should employ larger sample sizes to reduce variability and experiment with using smaller AI models as denoisers. MLLM users should put effort into crafting clean prompts and avoid traditional algorithmic denoisers.

Keywords: LLM, multimodal, MLLM, noise resistance, experiment

Table of Contents

Introduction

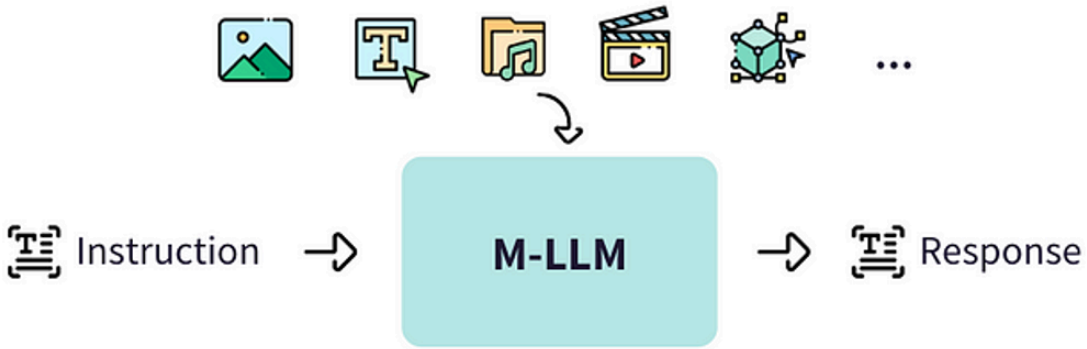
Artificial intelligence (AI) has entered a period of rapid development, innovation, and application in scientific research, "smart home" devices, and everyday use (Balas et al., 2020; Yan et al., 2023). Early AI models were primarily classification models, which simply took in input images or text and categorized them with predefined labels (e.g., "cat" and "dog" or "positive sentiment" and "negative sentiment"). These models used differentiation characteristics that they "learned" from training data (the data the model is "taught" from) to categorize new testing data (the data the model is used to analyze), giving the field its name: "machine learning" (Baskin et al., 2017). Initially, AI models were largely restricted to niche fields in the sciences and in automated sorting technologies (e.g., spam filters), but, catalyzed by the emergence of new architectural paradigms embodied by the launch of ChatGPT, an AI-powered chatbot, in late 2022, the application of AI has skyrocketed (Sain et al., 2024; Orgeldinger, 2024).

Specifically, the main new type of AI model that has been iterated upon over the past couple of years is, generically, the "large language model." A large language model (LLM) is trained on huge textual datasets gathered from various sources, including the Internet. Simplifying significantly (as the details are not particularly relevant to this paper), it can be said that the LLM's training algorithms adjust the "parameters" of the model as it "learns" from the training data. These tuned parameters ultimately produce a final predictive system capable of taking some sequence of input "tokens" (which are generally words or parts of words) and predicting the next token, with slight randomness ("temperature") often introduced to produce some response variability between subsequent runs (Kocoń et al., 2023). When trained on large enough volumes of data, the predictions of LLMs generally become coherent and often quite accurate. For that reason, huge LLMs with hundreds of billions of parameters are used to power

a plethora of applications, like GitHub Copilot (a popular code-completion tool for software developers) and, most famously, ChatGPT (which originally used a version of GPT-3.5, a purely text-based LLM).

However, more recently, a new type of LLM has emerged: the multimodal large language model. Multimodal LLMs (MLLMs) are large language models that can take more than just written language as input; they can interpret other media, including images. MLLMs still output textual responses, but they consider both input images and text as context for the response (Wu et al., 2023). This type of model could have tremendous potential in fields like scientific research, customer service relations, and more, and it is already integrated in platforms like ChatGPT when the user selects to use the more "advanced" models like GPT-4o, now OpenAI's flagship product.

Figure 1: MLLM Data Flow



Note. Like traditional text-based LLMs, MLLMs primarily take in a text prompt/instruction, but they also consider the contents of provided supplemental files—primarily images—in formulating their response.

If businesses, researchers, students, educators, and consumers begin relying increasingly on AI tools, as is happening right now, the models must be checked for robustness. While there are many types of robustness one might wish to measure, such as a model's overall reliability, a core type of robustness that has thus far been overlooked for MLLMs is noise resistance. By definition, a model with high noise resistance would not be significantly affected by slight random variations in the input (variations that would likely not change a human's response/analysis). Noise resistance is a good trait for a model as it means that it is predictable and that benchmarks showing its commendable performance in ideal cases can be extrapolated to more messy situations, like the analysis of satellite pictures tainted with sensor defects or the captioning of blurry photographs taken by shaking hands. Thus, the following broad question must be explored: "How does noise resistance vary across AI models in general, and MLLMs in particular, and what practical steps can be taken to improve it?" As will be discussed, much about MLLMs remains to be explored in the literature, which is what led me to my final question.

Literature Review

Much has already been discovered about the noise resistance of text classification models. For example, the insightful peer-reviewed study conducted by Agarwal et al. (2007) revealed that even under somewhat noisy conditions (including misspellings and other small errors), many text classification models can perform relatively well, although they break down more quickly when the input becomes excessively noisy. The authors find that overall, as the noise level increases, the classification accuracy ("fraction of correctly predicted document–class mappings") decreases. However, Agarwal et al. (2007) did find that when the text classification

models were trained on noisy data to begin with, they performed worse than models trained on clean data initially when the testing inputs had less noise, but those noise-trained models dropped less dramatically (more gracefully) in performance as the noise level increased, eventually surpassing the traditionally trained models. Overall, this means that noisy training data may result in better models if the testing data will also be noisy, for the model can learn to "overlook" the noise and pick up patterns (common typos, frequent abbreviations, etc.) within it. This led me to hypothesize that text-based LLMs, trained on noisy data from the Internet, should perform fine with some input noise; however, the two models follow completely different training and prediction systems, so generalization would be difficult without further research.

As further background, it is useful to consider image classification models, whose noise resistance may be indicative of what could be expected in MLLMs. In the peer-reviewed study conducted by Boonprong et al. (2018), the authors investigated the effects of three types of noise (zero-mean Gaussian, salt-pepper, and speckle noise) on three types of image classification systems (a back-propagation neural network, random forest model, and support vector machine—their implementation details are not relevant here). Gaussian noise is a type of additive noise that modifies each pixel's RGB values by noise sampled from a Normal distribution. Salt-pepper noise is when a few randomly selected pixels are made completely white or completely black. Speckle noise is a type of random multiplicative noise. Boonprong et al. used remotely sensed Landsat 8 images of a Thailand province and trained the three models to classify the environment's structure (empty, agricultural, construction, etc.). They found that while the models performed very well with little to no noise, as soon as the signal-to-noise ratio dropped below about 20 dB (i.e., as the noise increased), the accuracy rates began to plummet. Although salt-pepper noise was handled relatively well across all models, both of the other types

of input noise were devastating despite only slightly tweaking the image. Thus, while Agarwal et al. (2007) found that text classification algorithms are generally robust to noise, this study found that image classification algorithms are generally not. I thus hypothesized that MLLMs may also have poor image noise resistance, excepting salt-pepper noise.

Additionally, Boonprong et al. (2018) noted that image classification can be improved using "extended" algorithms and filters to clean the data before it is sent to the main AI model. With that approach, it seems that even image classification algorithms can be made highly resistant to noise. Overall, image-based algorithms generally seem worse at noise resistance than text-based algorithms, but classification algorithms as a whole are decent at noise resistance. These observations help set up the framework from which to hypothesize about and investigate LLM and MLLM noise resistance: It is now clear that (a) different types of image noise should be investigated, as each can have different effects, (b) denoising algorithms should be investigated to see if they rectify the effect of noise on models, and (c) image models can behave differently from text models in terms of noise resistance.

With classification thoroughly studied, LLM noise resistance can now finally be discussed. Levy et al. (2024) analyzed the effects of lengthening prompts and adding unnecessary additional information or noise on text-based LLM performance. Based on their results, it is evident that writing longer prompts with random information/gibberish embedded (a form of noise) hurts the quality and next-word accuracy (percentage of times the model correctly predicts the next few tokens) of LLM outputs. Salinas and Morstatter (2024) similarly analyzed the effect of small prompt changes (such as beginning a prompt with the phrase "Hello") on model performance. Their results corroborate Levy et al. (2024): It is clear that making small changes to the prompt (i.e., adding noise) frequently and adversely affects the quality and

accuracy of LLM outputs across different LLM models, especially smaller models with fewer parameters (an observation made by both studies). Thus, while textual classifiers are generally robust when faced with input noise, text-based LLMs are very vulnerable to input noise, highlighting that LLMs do behave differently from classification models in this area of investigation—there is no general answer for the noise resistance of all AI. Because of their shared architecture, one would expect MLLMs to share similar traits to traditional LLMs, and because of the media they deal with, one would also expect MLLMs to have some characteristics in common with image classification models. Fortunately, these two theories do not contradict one another, for both image-based classification models and text-based LLMs have poor noise resistance, leading to my hypothesis that MLLMs would also have poor noise resistance against textual noise and all image-noise types but salt-pepper noise (although, as Boonprong et al. [2018] hinted, there may be pre-processing algorithms to improve that performance).

Diving into MLLM baseline (no-noise) performance, Qi et al. (2023) interestingly find that MLLMs do not seem to understand supplementary images as thoroughly as one may expect. These researchers found that MLLMs do not reliably comprehend the essence of uploaded images, which makes it seem likely that random bit changes (i.e., superficial noise) could throw off the MLLM entirely. However, noise resistance was not actually tested, so more research was still necessary to confirm this hypothesis about the impact of noise on MLLMs experimentally.

Ultimately, after reading these papers, I began to suspect that while noise resistance has been tested for all types of classification models and text-based LLMs, only baseline performance has been measured for MLLMs, perhaps due to their novelty. Indeed, after a lengthy review of papers across many journals, databases, and search engines, including reviewing all Google Scholar results for “MLLM noise” and similar variations, it became

apparent that while general experiments on MLLMs have been conducted, their noise resistance in particular has not been thoroughly investigated. However, as described in the Introduction, it is vital to understand the robustness of such a prominent technology. Thus, the following research question is posed: "How significantly and in what ways does noise within text and within supplementary images affect the accuracy of an MLLM's textual reasoning abilities and image reasoning abilities (including comparisons to traditional LLMs)? If MLLM noise resistance is generally poor, do standard textual and image denoising algorithms help?"

Regarding pre-processing algorithms to reduce the noisiness of image inputs before they are passed onto the AI model, Ghofrani and Markarian (2016) highlight specific examples of such algorithms, although they do not apply them to AI, let alone MLLMs, in particular. This article, like the one by Boonprong et al. (2018), highlights three types of noise: Gaussian, salt-pepper, and speckle. The article references several noise-reduction algorithms that I can employ to see if they improve MLLM responses. Specifically, the article spends the most time on a modern noise-reduction algorithm called High-TV, which, from their human perspective, worked well on salt-pepper and speckle noise (although not Gaussian noise). However, High-TV has not been implemented in most standard libraries. Nevertheless, a somewhat similar "Fast NL Means" algorithm is implemented in standard image libraries, like OpenCV for Python, so I can use that algorithm to see if standard denoising algorithms improve, do not affect, or degrade MLLM responses to noisy images (Liu et al., 2008).

Based on the findings of my literature review, as stated earlier, I hypothesize that MLLMs will have especially poor noise resistance for text and images—worse than text-based LLMs. Moreover, lower-parameter MLLMs may have worse resistance (although that is not something I explicitly set out to test; the goal was simply to observe the differences between

MLLMs and LLMs overall and between MLLMs given clean, noisy, and denoised prompts). Although classification AI models are relatively noise-resistant, traditional text-based LLMs, the direct ancestors of MLLMs, have very poor noise resistance. Moreover, image processing always seems to be more noise-prone than text processing, even across classification models. Thus, combining the two least noise-resistant types of AI models, an image model and an LLM model, an MLLM is very likely to suffer in accuracy from noise. However, I suspect that denoising algorithms that smooth over or fill in odd pixels might improve MLLM image accuracy, as others have indicated is the case for classification models, and running spell-check on a noisy textual prompt might improve LLM textual accuracy alongside MLLM accuracy by providing more familiar tokens to the model. These hypotheses will be tested in the remainder of this paper.

Methods

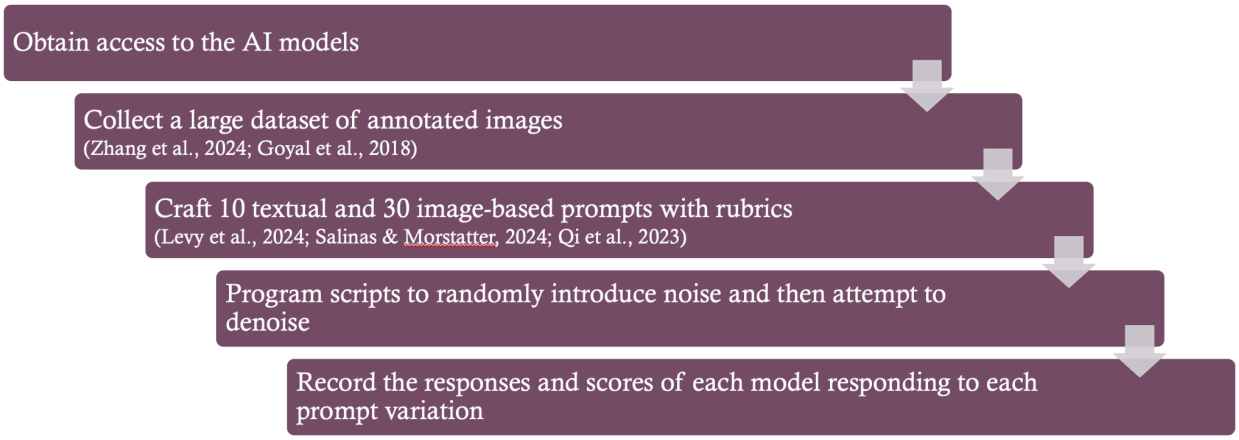
I used an experimental approach in order to establish a causal relationship between noisiness in the input images/textual prompts and model performance. My method, which involved providing prompt variations to the model with control (no noise) and noisy prompts, was grounded in standard research practices for the field of AI noise testing (Agarwal et al., 2007; Boonprong et al., 2018; Levy et al., 2024; Salinas & Morstatter, 2024). However, I used specialized image datasets developed for MLLMs, as will be discussed later, to cater to the more unique aspects of my question. I based my method in techniques widespread in the literature to provide validity to my study and enable my results to be widely accepted.

I believed an experimental design was appropriate as in order to test noise resistance, I needed to run the models on clean prompts, noisy prompts, and algorithmically denoised

prompts. Those three conditions can naturally be labeled as different treatments, and, with that framing, the AI models can be considered the experimental units. Following best practices, 40 different base prompts (10 textual, 30 image-based) were tested across 2 MLLMs and 1 text-based LLM. For each prompt, each model served as its own control under the control treatment of no noise, and each AI model also received both non-control treatments: a noisy version of the prompt and the denoised noisy prompt. This experimental design enabled me to determine a causal link between noise treatments (hereon interchangeably referred to as "noise levels") and model performance as all other factors were controlled (such as prompt difficulty, noise level, etc.) and each model served as its own control group in a matched-pairs-style setup (controlling for baseline model knowledge).

Meanwhile, if a content-analysis method were employed on existing user chats, perhaps gathered from MLLM platforms/providers, I would not have been able to control for those confounding variables. That method could also constitute an invasion of privacy. Interviews with MLLM users would be even more ineffective, for the users would not only need to remember their prompts but also describe the quality of the responses, which is inherently subjective without a per-prompt standardized rubric. Finally, it would also not have been prudent to conduct the experiment with different models receiving different prompts, as some prompts could be "easier" than others. Thus, my experimental design was optimal, for it controlled for every variable except each model's individual noise resistance, allowing for a comparison of noise resistance between LLMs and MLLMs and across noise levels/types.

Figure 2: Method Summary



Note. The key steps of my method, described and defended in more detail below, are summarized in this diagram.

To execute my method, I first obtained access to the models. For my MLLMs, I used LLaVA (the "34b" version, executed locally through Ollama) and GPT-4o (through OpenAI's API) to see if generalizations could be made about MLLMs as a whole class. I chose those two models specifically because they are very common in the literature. I also used GPT-3.5-turbo (hereon simply referred to as "GPT-3.5"), through OpenAI's API, as my text-based LLM for comparison, as it shares its core architecture (except for image support) with and is of similar size to GPT-4o.

Next, I collected a large dataset of annotated images. It was important to use professionally annotated images as those images would have been collected and documented by human professionals as a good test for AI models. As Zhang et al. (2024) point out, most available image datasets at the time of writing fell into one of two categories: they were large but annotated by AI models themselves and thus unreliable, or they were too small in quantity and scope. One relatively comprehensive dataset created by Zhang et al., which was initially

promising, was the MME-RealWorld dataset, which has many manually annotated images designed to challenge MLLMs. However, most of the images in the dataset already contained significant noise from the real world and were thus not valuable for the purpose of testing clean versus noisy images (they would render the "control" treatment meaningless). Thus, I decided to employ the annotated VQA (visual-question-answer) dataset highlighted in the paper by Goyal et al. (2018). This dataset had been credibly used in the field of MLLM research already, including by Qi et al. (2023), and the images were of generally good quality. Thus, I randomly selected 30 unique images from the dataset using the VQA website's "shuffle" feature and compiled their URLs in a text file. A table of the images used and their modified forms, as will be detailed later, can be found in Appendix A.

I then created a list of textual prompts and image-based prompts (see Appendix B). The textual prompts were to be given to all models, and the image-based prompts were to only be given to the two multimodal models. For the text prompts, I took inspiration from those used by Levy et al. (2024) and Salinas and Morstatter (2024). For the image-based prompts, I crafted questions based on the curated images from the previous step, inspired by the methods of Qi et al. (2023) and the annotations provided in the dataset from Goyal et al. (2018).

After finalizing the prompts, I programmed Python scripts to introduce noise (see Appendix C for the scripts and dependencies). I made one script that interactively takes text and randomly mutates characters with probabilities tuned after some general experimentation (to ensure that the script generated prompts that at least some models would be able to interpret). I also made another script that iterated through the list of image URLs from earlier, downloaded each image, applied a random noise type to each image (Gaussian, salt-pepper, or speckle), and attempted to denoise the image using OpenCV's flagship Fast-NL-Means algorithm (which was

not modified based on the noise type as, in a real-world setting, software would likely not be able to detect and adapt to different noise types). I chose to use those three noise types in particular because they were the most prevalent in the field's literature, as demonstrated by my Literature Review. Moreover, Luisier et al. (2011), Kakde (2024), and Choi and Jeong (2019) indicate that those three types of noise are suitable models for much real-world noise. The usage of Fast-NL-Means was already justified in the Literature Review.

For each of the textual prompts, I ran the first algorithm once per prompt and documented the new version of the prompt alongside the original. I then ran the standard GNU spell-check tool "aspell" on the noisy prompt, accepting the top suggestion for each flagged word (based on a common standard in the literature, as it is the best a real-world non-AI-based pre-processing text-denoising algorithm could do). For my image-based prompts, I just ran my second script one time to generate all of the images: 30 clean images and their corresponding noisy and denoised images (with the image type indicated in the filename).

Then, I created a rubric for what constitutes a "correct" (score 3), "mostly correct" (score 2), "partially correct" (score 1), or "incorrect" (score 0) response to each prompt (the rubrics are attached to each prompt in the appendices). This rubric was designed after some initial experimentation with prompts other than the prompts I would actually end up using to avoid any bias while still ensuring that I had enough experience to make the rubrics sensitive enough to the effects of noise to avoid vastly different quality answers collapsing into the same rubric score category. I decided to create this spectrum to allow me to do quantitative analyses and get a more nuanced understanding of model performance beyond a binary "completely right" and "wrong."

Then, I began data collection. I passed each of the three models all of the textual prompts and each of the two MLLMs all of the image reasoning prompts with noise, without noise, and

denoised. I recorded their responses and scores separately in a Google Sheets file (combined Appendix D for your convenience).

Conclusion: Results, Analysis, Discussion & Directions for Future Research

A total of 270 individual model responses were collected. 90 textual-prompt responses were collected (10 prompts × 3 varieties × 3 models) and 180 image-prompt responses were collected (30 prompts × 3 varieties × 2 models).

Based on these data, descriptive statistics were calculated and paired t-tests for mean differences in scores were conducted. The details for all of the relevant statistics and tests can be found in Appendix E. Findings significant at the α=.05 level are highlighted in red; at the α=.10 level, in orange; and at the α=.20 level, in yellow.

Textual Noise Impact

Ultimately, several conclusions can be drawn. First, models like LLaVA (that is, smaller models) seem to *not* be resistant to textual noise. With a one-tailed paired–t-test–for–mean–difference p-value of .048, there is convincing evidence LLaVA performs worse under textual noise conditions compared to clean conditions (at α=.05). That conclusion cannot be stated with nearly the same confidence in the case of GPT-4o, for which the one-tailed p-value was .172 (which would not be considered significant at most common levels). Note however that the p-value for the same test when combining both the LLaVA data and GPT-4o data into a "combined MLLM" category is very significant: .025. Thus, when considering MLLMs as a whole, there is convincing evidence that they are hurt by textual noise. Text-based GPT-3.5, in contrast, showed no mean difference in score between the prompts with and without textual noise. With these results, two comparisons can be made. First, as I

hypothesized and did actively try to test, MLLMs as a whole seem to be less adept at resisting textual noise when compared to text-based LLMs. This is a fair comparison as GPT-4o and GPT-3.5-turbo are almost identical in architecture except for multimodal capabilities (which is part of why I chose them, as explained in my Methods section). Second, it appears that smaller models (LLaVA) have much weaker noise resistance compared to larger models (GPT-4o and GPT-3.5) when it comes to textual noise, as I initially predicted based on my Literature Review (though did not aim to test), corroborating the studies conducted by Salinas and Morstatter (2024) and Levy et al. (2024). This is a reassuring conclusion that speaks to the validity of my research. Thus, practically speaking, when noise resistance is more important than cost for an AI project, it appears that larger models are preferable. That said, for future research, it would be a good idea to explicitly use variations of the same model with different parameter counts or "quantizations" to see whether parameter count is in fact the determining factor behind noise resistance or whether my results are better explained by some other confounding characteristic of the model (as, because parameter count was not a core part of my research question, I did not explicitly control all other factors except parameter count but rather simply selected three different models that were accessible, widespread in the industry, and allowed for comparison between text-based and multimodal models). Based on the results of my Literature Review and the data collected in my study, I would expect future studies to confirm that there is a strong negative correlation between MLLM parameter count and noise-induced performance reduction.

It is important to note that while GPT-3.5 did significantly outperform GPT-4o in terms of noise resistance, answering my question about how MLLM noise resistance compares to traditional LLMs, this should not be construed as implying that GPT-3.5 should be preferred in actual applications (except where it is more economical). Although GPT-3.5 resists noise better,

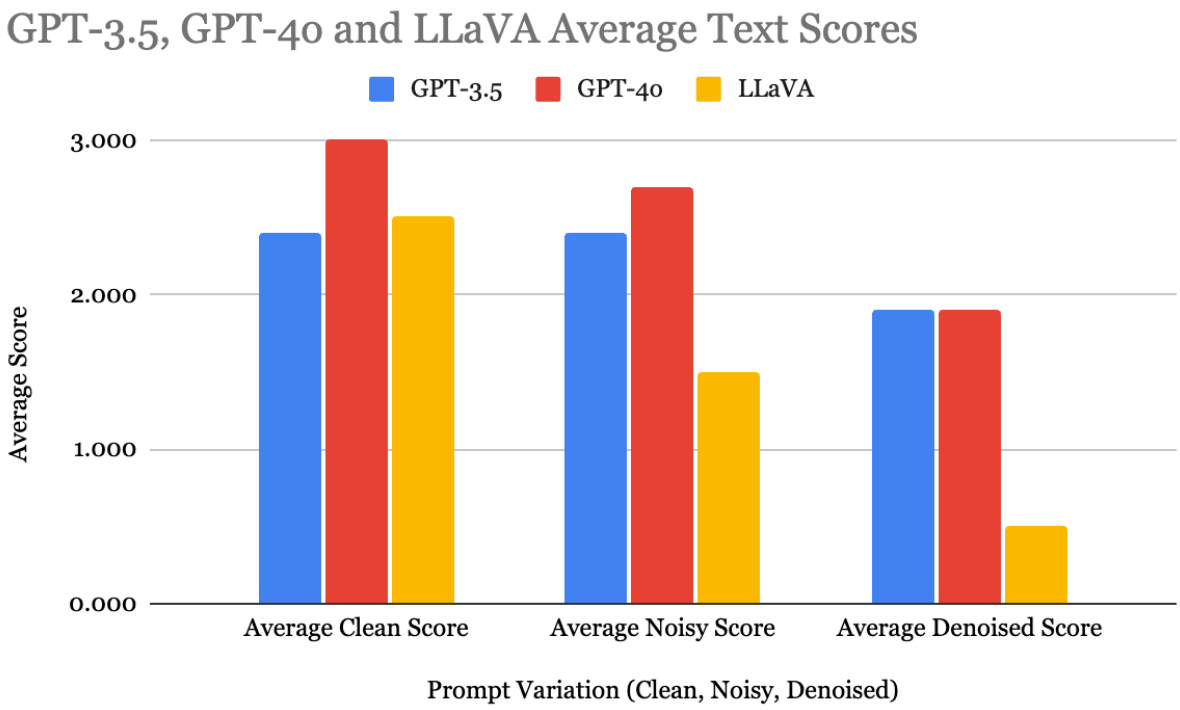
GPT-4o still outperforms or equals GPT-3.5 in average performance across all prompt variations. Thus, in cases where raw accuracy is more important than consistency, it seems GPT-4o is the better choice, as it is superior at all noise levels (no noise, noisy, and denoised) despite being less consistent between noise levels.

Textual Denoising Impact

Even more interestingly, contrary to my prediction and the recommendations of Boonprong et al. (2018) discussed in my Literature Review, it appears that a basic spell-check textual-denoising algorithm actually hurts LLM/MLLM accuracy rather than helping it. A two-tailed paired t-test for mean difference leads to p-values of .138, .087, .074, and .010 for GPT-3.5, GPT-4o, and LLaVA, and "combined MLLM," respectively. If this were transformed into a one-tailed test to see if there is convincing evidence that denoising hurts AI accuracy, the p-values would become .069, .044, and .036 for GPT-3.5, GPT-4o, and LLaVA respectively (the last two of which are statistically significant at α=.05). These results, although contradicting my initial hypotheses, do not directly contradict my Literature Review material, for Boonprong et al. (2018) only recommended denoising/pre-processing algorithms for classification models. As the comparisons made in my Literature Review between the text-classification study of Agarwal et al. (2007) and the LLM study of Levy et al. (2024) reveal, classification and large-language models are inherently different in how they deal with noise, so it makes sense that what helps one model could hurt the other. One plausible explanation for this is that LLMs, trained on vast quantities of language, can serve as extremely powerful (albeit expensive) spell-checkers on their own, so pre-processing the text with a more erroneous (less trained/nuanced) spell-checker is likely to confuse the model more than if the model is simply given the raw noisy text. Indeed,

GPT-3.5 had the same average clean score and noisy score of 2.4 but had a much lower average score when given the denoised prompts—1.9. Thus, in the broader context of LLM-based development, my recommendation would be to remove any intermediaries between the user input and the text that goes to an LLM for optimal accuracy. One other potential denoising algorithm that perhaps would have performed better would have been to use a (perhaps smaller) AI model as the pre-processor/denoiser. This would, of course, cost more in a practical application, which is why I felt testing it would not be very useful when considering MLLM noise resistance's real-world implications, but future research could be conducted to see if using a small LLM as a textual denoiser in a pre-processing step is effective for when that extra robustness is needed (by scrambling some text and measuring the percentage of the original text restored by a small AI model or by replicating my study, simply swapping "aspell" for an LLM).

Figure 3: Textual Prompt Scores Across Models and Noise Levels



Note. The general trend for textual prompts is clear: clean prompts perform the best, and denoised prompts perform the worst.

Image Noise Impact

Moving onto the image-based analysis, GPT-4o was very significantly hurt by noise overall, with a highly significant one-tailed paired–t-test–for–mean–difference p-value of .030 and mean score difference of -0.3. Meanwhile, LLaVA only had a mean score difference of -0.033, leading to a one-tailed p-value of 0.453 (essentially not significant at all). However, this seemingly greater noise resistance for LLaVA could partially be attributed to its far lower mean baseline score to begin with (2.300 for LLaVA compared to 2.833 for GPT-4o). Like in the previous comparison between GPT-4o and GPT-3.5 for textual noise, GPT-4o outperforms

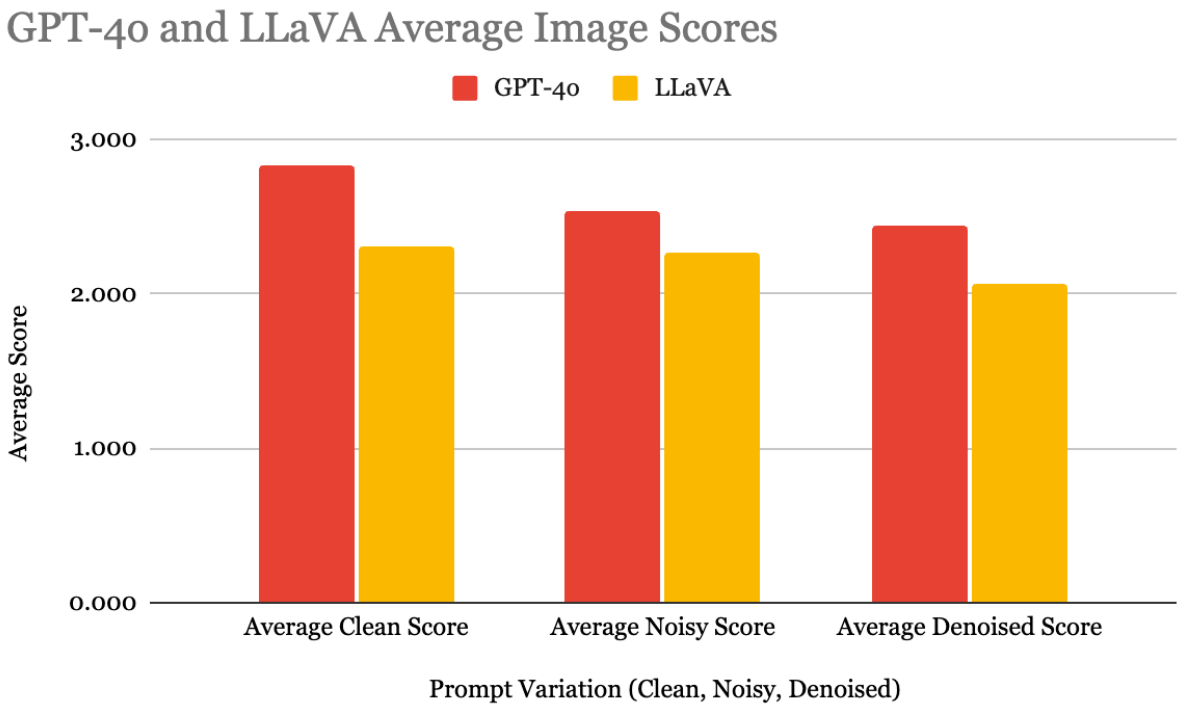
LLaVA at every image noise level, so if raw accuracy is vital above consistency and all other factors (such as cost and privacy, both of which benefit from running a local model like LLaVA), GPT-4o is the way to go. Similarly to the suggestion made in the discussion on the textual noise results, this is likely because of GPT-4o's higher parameter count. That said, as stated before, for future research, it would be a good idea to explicitly use variations of the same model with different parameter counts or "quantizations" to proactively isolate its effect. Regardless, in terms of my original research question, my hypothesis that MLLMs would be hurt by image noise seems to be true overall, although the observed variability limits my confidence. I encourage future researchers to replicate my study with a larger sample size to increase confidence and statistical power.

Image Denoising Impact

Even more interestingly, like with textual noise, contrary to my prediction and the recommendations of Boonprong et al. (2018) discussed in my Literature Review, it appears that a basic image denoising algorithm actually hurts AI accuracy rather than helping it. Compared to the noisy prompt, the denoised prompt caused an average drop in performance of 0.1 points for GPT-4o and 0.2 points for LLaVA. One plausible explanation for this is that MLLMs, trained on vast quantities of images, can serve as extremely powerful (albeit expensive) image denoisers on their own, so pre-processing the image with a more erroneous, less nuanced, and less trained OpenCV algorithm is likely to confuse the MLLM more than if the model is simply given the raw noisy image. However, it should be noted this comparison (denoised versus noisy) is not very statistically significant, with two-tailed paired–t-test–for–mean–difference p-values of .620, .489, and .389 for GPT-4o, LLaVA, and "combined MLLM," respectively. Thus, in the broader

context of MLLM-based software development, I recommend removing any intermediaries between the user input and the image that goes to an LLM for optimal accuracy. For future research, I recommend investigating the utility of smaller MLLMs as visual denoisers both independently and as a pre-processing step before an image is passed to the main MLLM, the latter of which could be tested simply by replicating my study with a small MLLM instead of OpenCV's algorithm.

Figure 4: Image Prompt Scores Across Models and Noise Levels

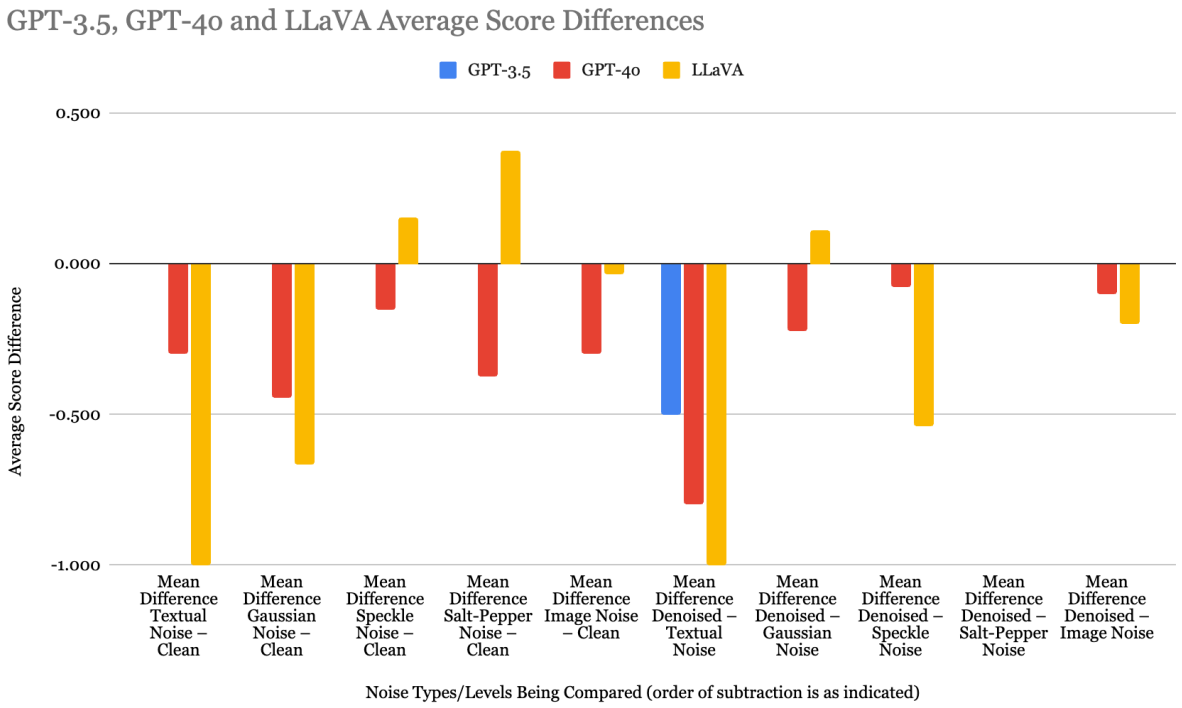


Note. The general trend for image prompts is also clear: clean prompts perform the best, and denoised prompts perform the worst.

Most and Least Problematic Image Noise Type

Finally, breaking the analysis down by individual type of image noise, Gaussian noise appeared to be most harmful, causing a mean score change of -0.444 and -0.667 for GPT-4o and LLaVA, respectively. This aligns with what Boonprong et al. (2018) found in their image classification study. Interestingly, the only time denoising caused an improvement in MLLM performance (which is *not* a measure of objective image restoration or human opinion, which is why the result seemingly contradicts the study of Ghofrani and Markarian [2016]) was with Gaussian noise for LLaVA (mean improvement of 0.111 points, albeit at a high two-tailed paired–t-test–for–mean–difference p-value of 0.880). Thus, I would tentatively suggest that in MLLM applications where Gaussian-like noise is expected, denoising algorithms could *potentially* be useful, but in other applications, they would very likely be detrimental. However, while just collecting the 180 image-based data points for the 30 image-based prompts took a significant amount of time, I hope others in the field can now pick up the torch and take on the challenge of generating, executing, and collecting data on many more prompts to reduce uncertainties and increase the statistical power of these tests. It would also be helpful for future researchers to investigate how increasing the severity of noise within each noise type (e.g., increasing/decreasing the spread of the Gaussian noise) affects MLLM responses: Is the relationship between noise severity and accuracy linear, curved, or something else entirely? Limited by time constraints, I focused my research question on investigating the effects of different *types* of noise and denoising algorithms on MLLMs versus LLMs, but noting the effects of noise severities could also be useful in informing real-world decisions on when MLLM use could be appropriate.

Figure 5: Score Differences by Noise Type/Level



Note. Gaussian noise clearly causes the worst performance impact, but it is also the only case in which denoising somewhat helped.

Summary

Overall, I found that both textual and image noise significantly hurt MLLM performance, even worse than for text-based LLMs. Among the various types of image noise, Gaussian noise appears to deal the worst damage. In general, denoising algorithms are not useful and are often actually detrimental to model performance, Gaussian noise denoising potentially excepted. Thus, as detailed earlier, my recommendation to the technology industry can be summarized as follows: do *not* pre-process inputs; encourage users to take their time when constructing their prompts, if at all possible; and generally stick with larger-parameter models (although that was

not a factor explicitly controlled in my study but rather an observation made and reported here to help guide MLLM users until further studies are published and to spark future research). My recommendation to future researchers, also explained in detail earlier, can be summarized as follows: try to replicate my study with larger sample sizes (as the variability in my results hindered the statistical power of some of my tests), attempt different denoising algorithms (including perhaps using smaller AI models as the denoisers), and explicitly control for parameter count.

References

Agarwal, S., Godbole, S., Punjani, D., & Roy, S. (2007). How much noise is too much: A study in automatic text classification. *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, 3-12. https://doi.org/10.1109/ICDM.2007.21

Balas, V. E., Kumar, R., & Srivastava, R. (2020). *Recent trends and advances in artificial intelligence and Internet of Things*. Springer. https://doi.org/10.1007/978-3-030-32644-9

Baskin, I. I., Marcou, G., & Horvath, D. (2017). Classification models. In A. Varnek (Ed.), *Tutorials in chemoinformatics* (pp. 175-192). Wiley. https://doi.org/10.1002/9781119161110.ch11

Boonprong, S., Cao, C., Chen, W., Ni, X., Xu, M., & Acharya, B. K. (2018). The classification of noise-afflicted remotely sensed data using three machine-learning techniques: Effect of different levels and types of noise on accuracy. *ISPRS International Journal of Geo-Information*, *7*(7), 274. https://doi.org/10.3390/ijgi7070274

Choi, H., & Jeong, J. (2019). Speckle noise reduction technique for SAR images using statistical characteristics of speckle noise and discrete wavelet transform. *Remote Sensing*, *11*(10), 1184. https://doi.org/10.3390/rs11101184

Ghofrani, S., & Markarian, H. (2016). Using high order total variation for denoising speckle, Gaussian, salt & pepper. *2016 International Conference on Micro-Electronics and Telecommunication Engineering*, 282-286. https://doi.org/10.1109/icmete.2016.37

Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., & Parikh, D. (2018). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, *127*(4), 398-414. https://doi.org/10.1007/s11263-018-1116-0

Kakde, B. (2024). Study of salt and pepper noise removal by different filtering techniques. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4989151

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *ACL Anthology: Vol. 1. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 15339-15353). Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.818 (Original work published 2024)

Liu, Y.-L., Wang, J., Chen, X., Guo, Y.-W., & Peng, Q.-S. (2008). A robust and fast non-local means algorithm for image denoising. *Journal of Computer Science and Technology*, *23*(2), 270-279. https://doi.org/10.1007/s11390-008-9129-8

Luisier, F., Blu, T., & Unser, M. (2011). Image denoising in mixed poisson–gaussian noise. *IEEE Transactions on Image Processing*, *20*(3), 696-708. https://doi.org/10.1109/tip.2010.2073477

Orgeldinger, J. (2024). AI driven large language models LLMs (ChatGPT, Copilot and Gemini) in finance – opportunities and challenges. *Zeitschrift Für Das Gesamte Bank- Und Börsenwesen*, *72*(5), 325-333. https://doi.org/10.47782/oeba202405032501

Qi, S., Cao, Z., Rao, J., Wang, L., Xiao, J., & Wang, X. (2023). What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing. *Information Processing & Management*, *60*(6), Article 103510. https://doi.org/10.1016/j.ipm.2023.103510

Sain, Z. H., Thelma, C. C., Baharun, H., & Pigesia, A. C. (2024). ChatGPT for positive impact? Examining the opportunities and challenges of large language models in education. *International Journal of Educational Development*, *1*(3), 87-100. https://doi.org/10.61132/ijed.v1i3.75

Salinas, A., & Morstatter, F. (2024, April 1). The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance [preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2401.03729

Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023). Multimodal large language models: A survey. *2023 IEEE International Conference on Big Data*. https://doi.org/10.1109/bigdata59044.2023.10386743

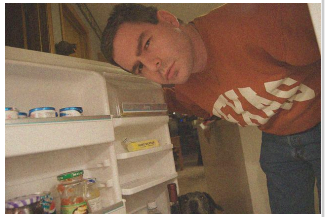
Wu, K., Jiang, B., Jiang, Z., He, Q., Luo, D., Wang, S., Liu, Q., & Wang, C. (2024, May 31). NoiseBoost: Alleviating hallucination with noise perturbation for multimodal large language models [preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2405.20081

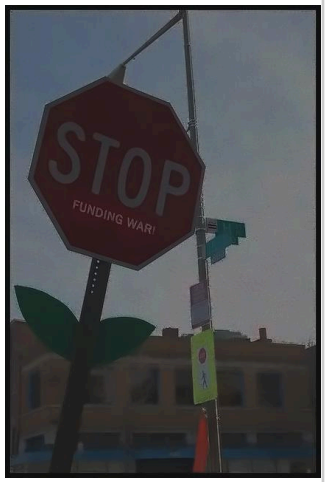
Yan, Y., Li, B., Feng, J., Du, Y., Lu, Z., Huang, M., & Li, Y. (2023). Research on the impact of trends related to ChatGPT. *Procedia Computer Science*, *221*, 1284-1291. https://doi.org/10.1016/j.procs.2023.08.117

Zhang, Y.-F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., Wang, L., Jin, R., & Tan, T. (2024, September 11). MME-RealWorld: Could your

multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? [preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2408.13257

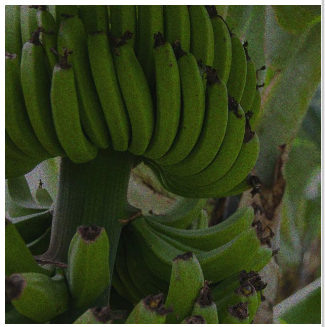
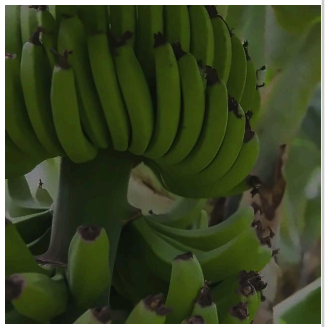
Appendix A: Images Used

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
1				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000397645.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/1-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/1-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/1-gaussian-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
2				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000049920.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/2-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/2-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/2-speckle-denoised.jpg?raw=true

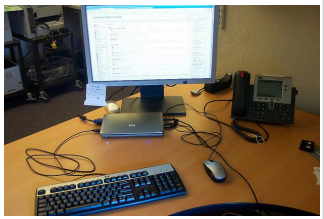
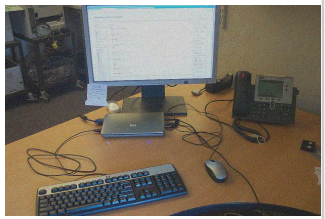
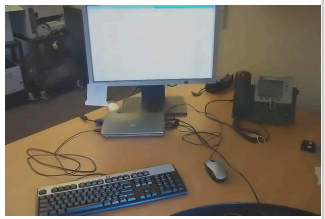
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
3				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000055474.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/3-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/3-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/3-speckle-denoised.jpg?raw=true

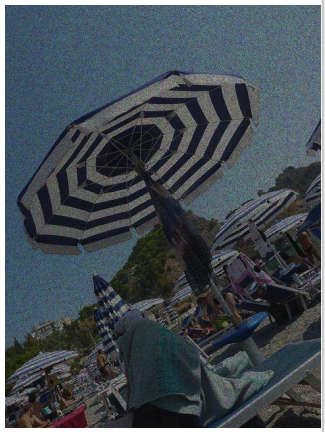
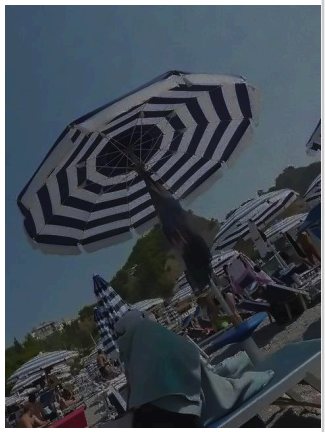
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
4				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000026017 5.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/4-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/4-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/4-gaussian-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
5				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000056840.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/5-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/5-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/5-speckle-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
6				Salt-Pepper Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000059610.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/6-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/6-salt-pepper.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/6-salt-pepper-denoised.jpg?raw=true

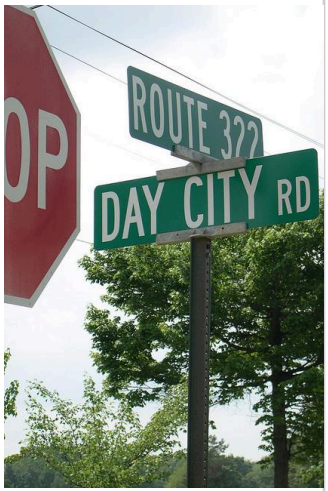
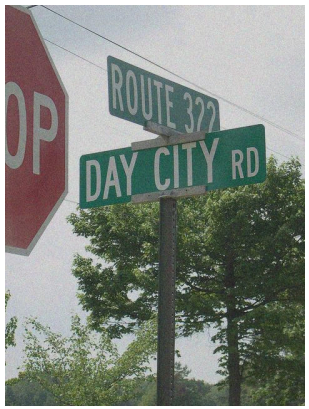
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
7				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000080105.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/7-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/7-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/7-speckle-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
8				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000150276.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/8-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/8-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/8-gaussian-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
9				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000376524.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/9-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/9-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/9-speckle-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
10				Salt-Pepper Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000021956.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/10-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/10-salt-pepper.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/10-salt-pepper-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
11				Salt-Pepper Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000027749.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/11-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/11-salt-pepper.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/11-salt-pepper-denoised.jpg?raw=true

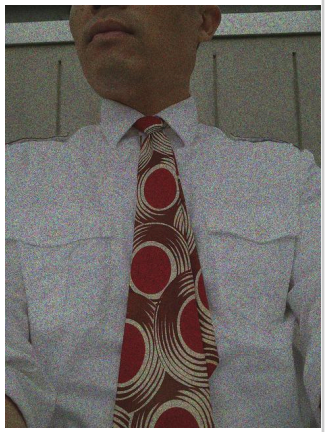
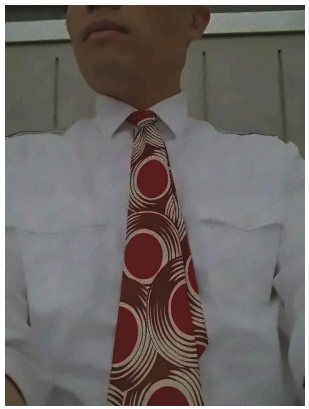
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
12				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000052182.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/12-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/12-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/12-gaussian-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
13				Salt-Pepper Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000340138.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/13-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/13-salt-pepper.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/13-salt-pepper-denoised.jpg?raw=true

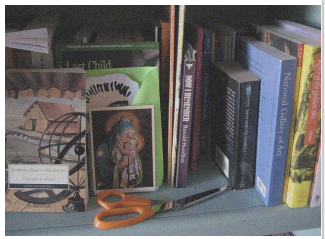
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
14				Salt-Pepper Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000315242.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/14-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/14-salt-pepper.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/14-salt-pepper-denoised.jpg?raw=true

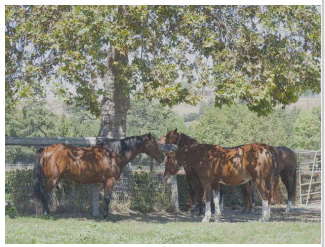
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
15				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000045954 3.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/15-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/15-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/15-speckle-denoised.jpg?raw=true

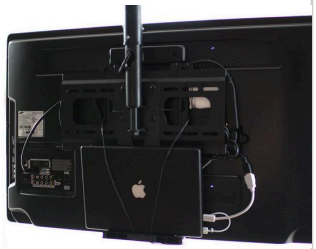
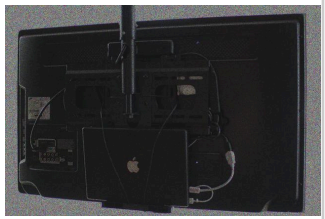
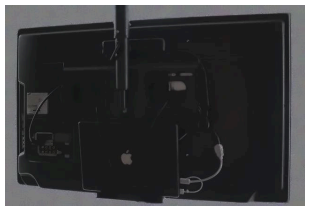
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
16				Salt-Pepper Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000042268.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/16-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/16-salt-pepper.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/16-salt-pepper-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
17				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000035324.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/17-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/17-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/17-speckle-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
18				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000091681.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/18-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/18-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/18-speckle-denoised.jpg?raw=true

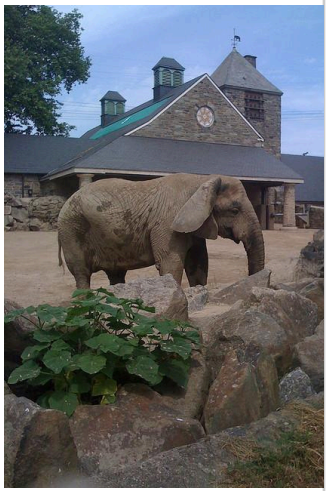
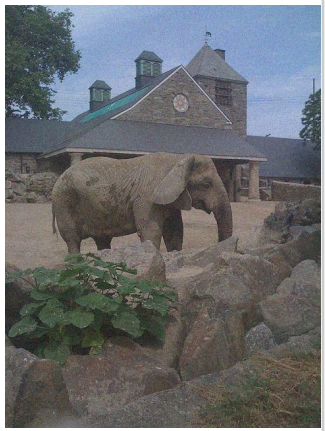
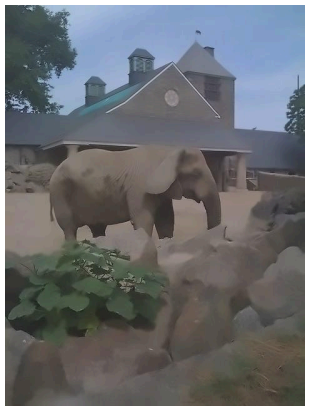
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
19				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000039376.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/19-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/19-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/19-gaussian-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
20				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000552276.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/20-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/20-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/20-gaussian-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
21				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000040645 2.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/21-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/21-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/21-speckle-denoised.jpg?raw=true

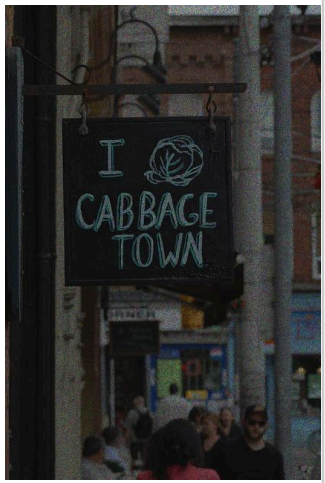
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
22				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000029875 9.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/22-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/22-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/22-speckle-denoised.jpg?raw=true

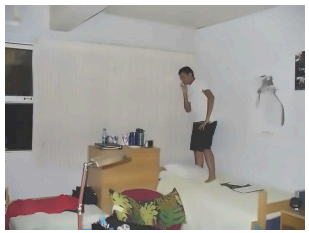
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
23				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000019759 1.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/23-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/23-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/23-gaussian-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
24				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000049449.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/24-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/24-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/24-gaussian-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
25				Salt-Pepper Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000051894.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/25-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/25-salt-pepper.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/25-salt-pepper-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
26				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000254290.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/26-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/26-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/26-speckle-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
27				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000368893.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/27-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/27-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/27-speckle-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
28				Gaussian Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000148502.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/28-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/28-gaussian.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/28-gaussian-denoised.jpg?raw=true

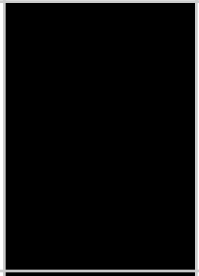
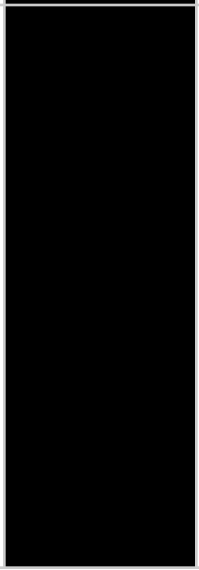
ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
29				Speckle Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000385341.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/29-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/29-speckle.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/29-speckle-denoised.jpg?raw=true

ID	Clean Image	Noisy Image	Denoised Image	Noise Type	Original Image Source (From VQA v2 Dataset)	Self-Hosted GitHub Links to All Image Versions (Anonymous Account)
30				Salt-Pepper Noise	https://vqa_mscoco_images.s3.amazonaws.com/train2014/COCO_train2014_000000053755.jpg	Clean: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/30-clean.jpg?raw=true Noisy: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/30-salt-pepper.jpg?raw=true Denoised: https://github.com/multimodal-llm-noise/code-and-images/blob/main/images/30-salt-pepper-denoised.jpg?raw=true

Appendix B: Prompts

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
1	Context: Jack's living room is marble-floored, a reality that is as intrinsic to the building as its very foundations. James's living room is not marble-floored, which is something he has been regretting for all eternity. Ethan Washington has now come to visit James and is sitting in James's living room, a fact that has become as much a part of the place as the walls and the ceiling. Question: Whose home is	Dontext: Jackos living goom is marble-flRored, aFre5liQK thEQ ws as in7NinWic to the buimdin8Has its very foundVAionszCDames's liming 7oom Hs nlt marKlW6flo(red, whicQ Fs symething he has beUn rehretting for all edernity. Et5anwUashizYton has now come to misit J"mes and is sitoing in Zames's >iving roWm, a fact that has becomecas)much a p*rm of theNplacb:as th2 walls and uhe ceilingK	"Context: Jacks living gloom is marble-floored, afire5liq thew SW as in7Nunki to the building8Has its very foundVAionszCDames's liming 7Odom HS BLT Mark6Flo(red, which F's something he has begun regretting for all eternity. ET5Anniston has now come to misfit J""mes and is siting in Ames's >Irving ROM, a fact that has becomes)much a p*rm of unplug:as Th2 walls and hue ceiling	[redacted]	0: All three elements incorrect 1: Exactly one element correct 2: Exactly two elements correct 3: All three elements correct Answer: James's room, living room, no (not	People in Rooms / Levy et al. (2024)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
	Ethan Washington in, what room is Ethan Washington in, and is it a marble-floored room?	Questi9n: 4hose home is Ethan lrshington in, wh9t roomfis Ethan Wvshington in, aad iS it a marblR-flooTed rYom?	Quest9n: 4hose home is Ethan Washington in, WHO9t roomfuls Ethan Washington in, Ada is it a marble-floored ROM?	[redacted]	marble-floored)	
2	What is the solution to the following problem? Write the reasoning and highlight the answer at the end starting with: "Answer:". Problem: "Alice has two red apples. Bob gives Alice one apple. But then John decides to take away three from her. Apples are quite delicious, are they not! Alice, Bob, and John all	What3is the solutiooTtoGtheOfollowin g problem5 W0ite the reas6nimg and1high0iZht the +nswer at the end star5ing cith: "DnsNer:". Proj0em: "Alile has two red axples. Bo9 givJs Alice one apple. But then John decid2s to take away three frSm her. Apfl?siar; quite deliSious, are the5 not!	"What3is the solutiooTtoGtheOfollowin g problem5 W0Ute the rears6MiG and1high0AZT the +answer at the end star5ING cit: ""Denser:"". PRO0em: "Allie has two red axles. BO9 gives Alice one apple. But then John decide2s to take away three from her. Apple?soar; quite Delicious, are the5 not!	[redacted]	0: Incorrect answer and completely incorrect reasoning 1: Incorrect answer but partially correct reasoning (at least 1 intermediate calculation	MathQA / Salinas and Morstatter (2024)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
	love apples! How many apples does Alice have now?"	AlicTZ Bob, and Johl allalove apples!3How many ap}leA does Alice have nowl"	Elects Bob, and Kohl allover apples!3How many AP}Lea does Alice have Noel""		correct) 3: Correct answer Answer: 0 apples	
3	Which one of the following attributes most likely describes the following text: "not sarcastic" or "sarcastic"? Text: "Alice has two red apples."	Which onu of thedfollowMng at1ributes most lioely describes mheNfotlswing (exK: "noE sarcastic" or "sarcUstic"? ueBt:K"Klice has twu red applos.y	Which onus of defilement at1tributes most lowly describes mheNfotlswing (ex: "Noe sarcastic" or "sarcastic"? debt:K"Lice has Tu red apples.y		0: "Sarcastic" with no reasonable justification 1: "Sarcastic" with somewhat reasonable justification 3: "Not sarcastic" Answer: "Not sarcastic"	iSarcasm / Salinas and Morstatter (2024)

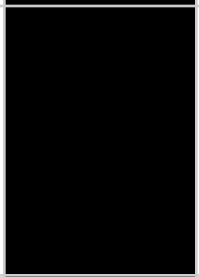
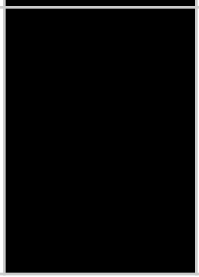
ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
4	Which one of the following attributes most likely describes the following text: "not sarcastic" or "sarcastic"? Text: "I'm sure Alice really loves sitting alone at home doing nothing!"	0eWt: qV3m sur0VAlice really loves sPttiLg al~ne at h4me doing notSinga7"Which one of the following attribuuesimost l9kelytdescribe] the following t7Lt: "not sarcastic" o} "0arcastxc"? Text: "I'm su=e Alice reall| loves 2itting alone at,hcmq ^oing nothing!"	"0wet: WV3m Sir0Alice really loves spottily AL~NE at h4me doing nosing7""Which one of the following attribuuesimost l9kelytdescribe] the following t7Lt: "not sarcastic" o} "0acoustics"? Text: "I'm Si=e Alice really| loves 2fitting alone at,Hickman ^OKing nothing!""	[redacted]	0: "Not sarcastic" with no reasonable justification 1: "Not sarcastic" with somewhat reasonable justification 3: "Sarcastic" Answer: "Sarcastic"	iSarcasm / Salinas and Morstatter (2024)

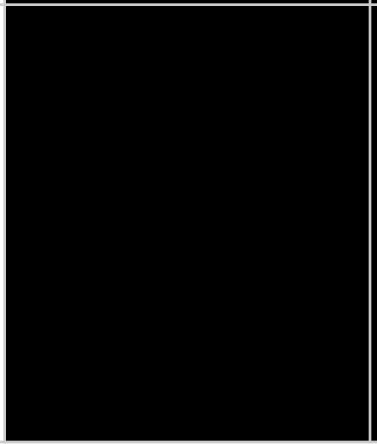
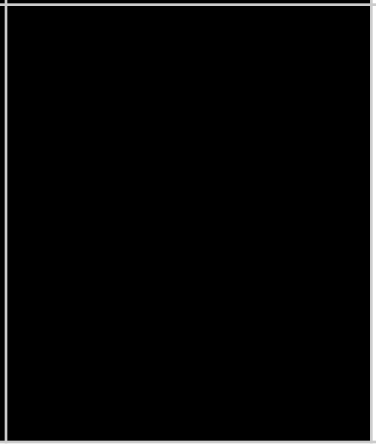
ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
5	Which one of the alternatives, "Alternative 1" or "Alternative 2," is most plausible given the premise? Premise: "Bob's plant died." Alternative 1: "The plant decided it didn't like living in Bob's home." Alternative 2: "Bob failed to water the plant enough."	Which 4ne of the alternatives, "Alt>rnative h" or "AlternMtive 2," is most7plausible gXven the premise? Nremiso: qBob's plan\ died." Aeternative 1: "The plu~t decided;iB didn't2likV liWing in BoB's home.5 AZte0native 2Du"Bob failed to7wBger 2he plant enob=h."	Which 4NE of the alternatives, "Alt>native h" or "Alternative 2," is most7plausible given the premise? Remiss: Bob's plan\ died." Alternative 1: "The ply~t decided;Ibo didn't2like lowing in Bob's home.5 Aztec0native 2Du"Bob failed to7wager 2he plant nob=h."	[redacted]	0: "Alternative 1" with no reasonable justification 1: "Alternative 1" with somewhat reasonable justification 3: "Alternative 2" Answer: "Alternative 2"	CoPA / Salinas and Morstatter (2024)

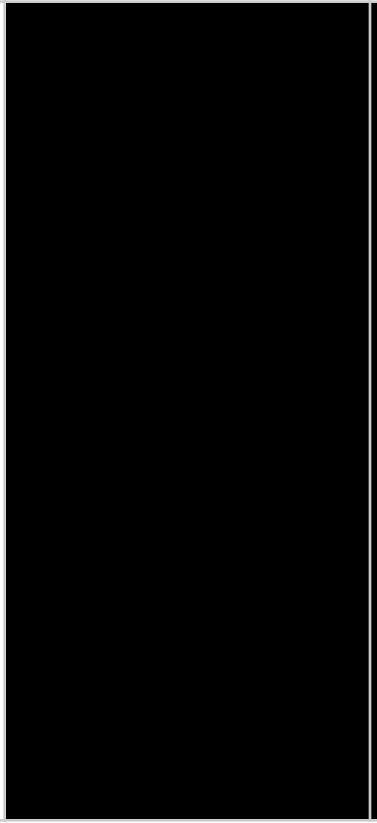
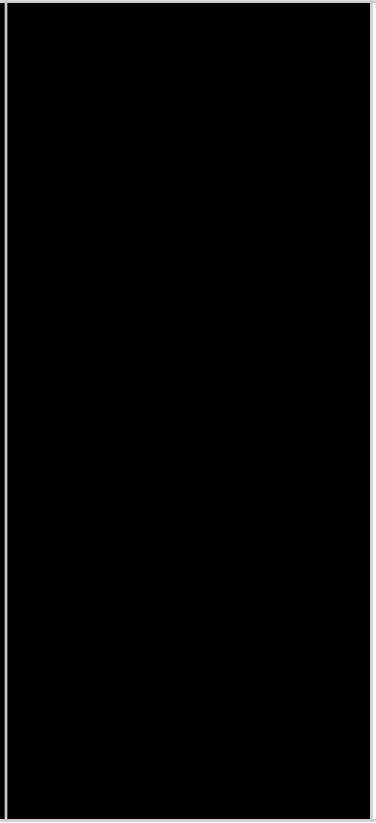
ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
6	Which one of the alternatives, "Alternative 1" or "Alternative 2," is most plausible given the premise? Premise: "I have a new home." Alternative 1: "I bought a house a month ago and it's finally ready." Alternative 2: "I woke up one morning and found a new house had magically appeared."	WhFch one SfNth4 a1terna6xvgs,r"Altzgnative 1" o: "Alte2na3iveO2," 0. mos6 plJVsibll given t<e preiise? Plpmise: "I have a newLhome." Alternative 1: "I b.ught a h.use#a`mont1kagF and it's finE9hu ready." Alternet/Me 2: "I loke up ono momning and foundga newdhouse hadamaCicallv appeared."	Which one Synth4 a1terns6xv gs,r"Alternative 1" o: "Alter2Na3I've2," 0. mos6 plausible given t<e precise? Plumes: "I have a newline." Alternative 1: "I b.ugh a h.use#a`Mont1jag and it's fine9HI ready." Alternate/Me 2: "I Locke up Ono moaning and founds madhouse headscarf appeared."	[redacted]	0: "Alternative 2" with no reasonable justification 1: "Alternative 2" with somewhat reasonable justification 3: "Alternative 1" Answer: "Alternative 1"	CoPA / Salinas and Morstatter (2024)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
7	Context: Bob's house is painted brown, like Jack's house. Pretty much everything inside Bob's house is brown too, except his dining room, which is bright yellow. John has come to visit Bob and is in his living room. Question: Whose home is John in, what room is John in, and what color is the room?	C2nteJt:yBob's h:use is pg[nt2= brown, 2ike Jacw's house. Pregby mNchIeverything insi9e Bob's hoZ4e is bro#n too, e`Bept his dining room, which is bright yellow. John 4Hs com0 to visit Bob and iB in his living room. 8uestion: WhVse 6ome is J4hn Xn, whay5room is John in(and wh9t colRr is the room?	C2nitwit:Bob's h:use is pg[NT2= brown, 2Ike Jaw's house. Rugby maneuvering INS9e Bob's Hz4e is bro#n too, e`Bet his dining room, which is bright yellow. John 4HS com0 to visit Bob and Ibo in his living room. 8question: Whose 6Moe is J4Han Zn, what5room is John in(and WHO9t corr is the room?	[redacted]	0: All three elements incorrect 1: Exactly one element correct 2: Exactly two elements correct 3: All three elements correct Answer: Bob's room, living room, brown	People in Rooms / Levy et al. (2024)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
8	What is the solution to the following problem? Write the reasoning and highlight the answer at the end starting with: "Answer:". Problem: "Jack has an allowance of $10.50. He buys four chocolate ice cream cones for his friends, as well as one for himself. Each ice cream cone costs $2.00. Now, he wants to buy bubble gum. One pack costs $0.75. Does he have enough money to buy the gum?"	,h2t is tie solution]toJthe fvllowing hroJl}m? Write thq rAasoning andmhighlight the answer at the end s22rting with: KAnswrr:G. ProblemG "Jack uaP an #7lowance on $10.50. He buys 4our c#ocolate ice croam cones 0or his fri39d$, as welS as one for hzmself. Each vce[cream cone costs $2.00. Now, hw want2 to buyMbubble guA. One pacg cqsts $0.75. Does he have en]ugh money !t buy t^zJgum)"	,h2t is tie solution]tithe fallowing Heroku}m? Write HQ reasoning annihilate the answer at the end s22rating with: Kans:G. Problem "Jack USP an #7Lance on $10.50. He buys 4our c#acolyte ice cram cones 0or his Fri39d$, as Wells as one for himself. Each vice[cream cone costs $2.00. Now, Haw want2 to bumble Guam. One PAC casts $0.75. Does he have en]ugh money !t buy t^scum)"	[redacted]	0: Incorrect answer and completely incorrect reasoning 1: Incorrect answer but partially correct reasoning (at least 1 intermediate calculation correct) 3: Correct answer Answer: No (he has only	MathQA / Salinas and Morstatter (2024)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
					$0.50 left, less than the cost of a pack of gum)	
9	Chase is younger than Emma. Emma is older than Bob. Is Bob younger than Chase? (Respond with yes, no, or cannot be determined.)	Chnseois=younger than Emma.&Jmma is oTyer than Bob+8Is Bob BoungeN than Chas3G (RwspoLd with ?ss, no, or cannot be d'termined.)	Chinese's=younger than Emma.&Jam is otter than Bob+8Is Bob Bounden than Chase3G (Resold with ?SS, no, or cannot be determined.)		0: Incorrect 3: Correct Answer: Cannot be determined	MonoRel / Levy et al. (2024)
10	Chase is younger than Emma. Emma is younger than Bob. Is Bob younger than Chase? (Respond with yes, no, or cannot be determined.)	Chase isx8ouncer th5n umJaS Emma is younge\ than BBbz Is Bob younger than Chyse? (Respond Cith y0s, 1o, oF cannot be Ketbrwined.)	Chase six8ounce Th5n umiaks Emma is younger\ than BBB Is Bob younger than Chase? (Respond Cit y0s, 1o, of cannot be Birdbrained.)		0: Incorrect 3: Correct Answer: No	MonoRel / Levy et al. (2024)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
11	What state is on the man's shirt?			1	0: Incorrect 3: Correct Answer: Texas	VQA / Qi et al. (2023)
12	What does the sign say?			2	0: Completely incorrect 1: "STOP" only 3: Completely correct Answer: STOP FUNDING WAR!	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
13	What is the man doing? Be specific.			3	0: Incorrect 1: Just "skateboarding" 3: Completely correct (detailed) Answer: Performing an ollie on a skateboard (appropriate description without using the word "ollie" also earns full credit)	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
14	What color is the train?			4	0: Incorrect 3: Correct Answer: Purple	VQA / Qi et al. (2023)
15	Are the bananas ripe?			5	0: Incorrect 3: Correct Answer: No	VQA / Qi et al. (2023)
16	How many horses are there?			6	0: Incorrect 3: Correct Answer: 9	VQA / Qi et al. (2023)
17	Is the grass taller, shorter, or the same size as the baby elephant?			7	0: Incorrect 3: Correct Answer: Taller	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
18	What color is the telephone handset?			8	0: Incorrect 3: Correct Answer: No	VQA / Qi et al. (2023)
19	What colors compose the umbrella?			9	0: Both colors incorrect 1: One of the two colors correct 3: Both colors correct Answer: Blue and white	VQA / Qi et al. (2023)
20	What is located directly above the animal's ears?			10	0: Incorrect 3: Correct Answer: Horns	VQA / Qi et al. (2023)

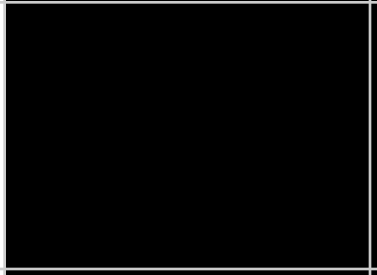
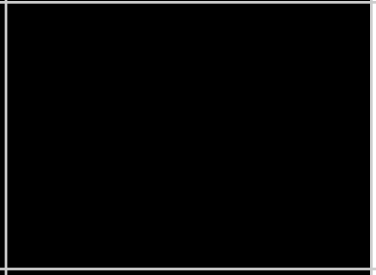
ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
21	What is the man doing? Be specific.			11	0: Incorrect 3: Correct Answer: Cutting cake	VQA / Qi et al. (2023)
22	What is the street name?			12	0: Neither Route 322 or Day City Rd 1: Route 322 (other piece of text in the image, but it's a highway, not a street) 2: Both Route 322 (not a street) and Day City Rd 3: Only the	VQA / Qi et al. (2023)

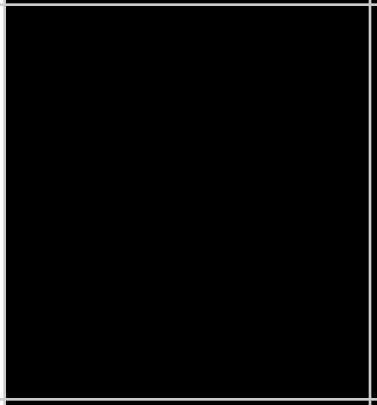
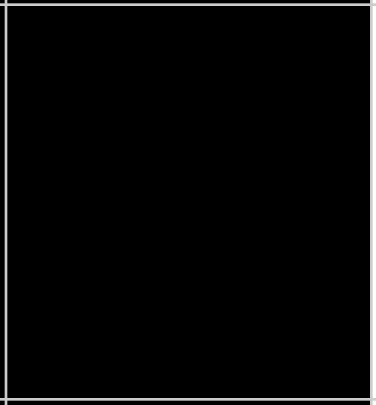
ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
		[redacted]	[redacted]		correct answer Answer: Day City Rd	
23	Is the pizza vegetarian? Why or why not?	[redacted]	[redacted]	13	0: Vegetarian 1: Non-vegetarian, but incorrect reason 3: Correct answer and reason Answer: Non-vegetarian because there's pepperoni	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
24	What does the skateboard say?			14	0: Incorrect 3: Correct Answer: Burton	VQA / Qi et al. (2023)
25	Is this a normal sight?			15	0: Incorrect 3: Correct Answer: No	VQA / Qi et al. (2023)
26	What time is it?			16	0: Incorrect 3: Correct Answer: 2:20	VQA / Qi et al. (2023)
27	How many distinct colors does the tie have?			17	0: Incorrect 3: Correct Answer: 3	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
28	What is the weather like?	[redacted]	[redacted]	18	0: Incorrect 3: Correct Answer: It's raining	VQA / Qi et al. (2023)
29	What religious figures, if any, are there in this image?	[redacted]	[redacted]	19	0: No religious figures 1: Incorrect religious figures 2: One correct religious figure 3: Both correct religious figures Answer: Virgin Mary and (baby) Jesus	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
30	How many horse heads are at least partially visible?			20	0: Incorrect 1: 4 horses (because there is an extra horse, but it's head isn't visible) 3: Correct Answer: 3	VQA / Qi et al. (2023)
31	What company made this?			21	0: Incorrect 3: Correct Answer: Apple	VQA / Qi et al. (2023)
32	What utensils are shown?			22	0: No correct utensils 1: One correct utensil	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
					3: Completely correct Answer: Knife and spoon	
33	Are there traffic lights?			23	0: Incorrect 3: Correct Answer: Yes (in background)	VQA / Qi et al. (2023)
34	What animal is there?			24	0: Incorrect 3: Correct Answer: Elephant	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
35	What material is the bowl made out of?			25	0: Incorrect 3: Correct Answer: Glass	VQA / Qi et al. (2023)
36	What pattern is displayed most prominently on the items in this image? Be specific.			26	0: Incorrect 3: Correct Answer: Union jack (or a detailed an appropriate description thereof)	VQA / Qi et al. (2023)
37	Is this China Town?			27	0: Incorrect 3: Correct Answer: No (it's Cabage	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
					Town)	
38	Where is the man standing?			28	0: Incorrect 3: Correct Answer: On a bed	VQA / Qi et al. (2023)
39	How many people are there, and what do they appear to be doing?			29	0: Incorrect number and activity 1: Correct number or activity 3: Correct number and activity Answer: 2 people playing	VQA / Qi et al. (2023)

ID	Original Prompt	With Textual Noise *(only if textual prompt)*	With Denoised Text *(only if textual prompt)*	Associated Image ID *(only if image prompt—see Appendix A for the image's noise variations)*	Rubric	Task Type / Reference Source
		[redacted]	[redacted]		a video game	
40	What color is the bird's head?	[redacted]	[redacted]	30	0: Incorrect 3: Correct Answer: White	VQA / Qi et al. (2023)

Appendix C: Python Scripts

Note. All of these scripts are compatible with Python 3. To run them, you must install the Python packages OpenCV, Pillow, and Numpy. The image noise script requires a file named “images.txt” with a newline-separated list of image URLs. The text noise script is simply interactive.

Textual Noise Script

GitHub Link

```
import random
import string

# Interactively add noise to text
if __name__ == "__main__":
   print("Type text to add noise to it. Type /exit to exit.")

   # Probabilities of each type of character noise
   randomLetterProb = 0.07
   randomDigitProb = 0.03
   randomPunctuationProb = 0.02

   while True:
       text = input("> ")
       if text == "/exit":
           break
```

```
    for i in range(len(text)):
        randomVariable = random.random()

        if randomVariable < randomLetterProb:
            text = text[:i] + random.choice(string.ascii_letters) + text[(i + 1) :]
        elif randomVariable < randomLetterProb + randomDigitProb:
            text = text[:i] + random.choice(string.digits) + text[(i + 1) :]
        elif (
            randomVariable
            < randomLetterProb + randomDigitProb + randomPunctuationProb
        ):
            text = text[:i] + random.choice(string.punctuation) + text[(i + 1) :]

    print(text + "\n")
```

Image Noise Script

GitHub Link

```
import random
import ssl
import urllib.request
from typing import Literal, Union
```

```
import cv2
import numpy as np
from numpy.typing import NDArray
from PIL import Image


# Read file as ndarray
def load_image(filename: str):
    image = Image.open(filename)
    return np.asarray(image)


# Save ndarray as file
def save_image(image: NDArray, filename: str):
    if image.dtype != np.uint8:
        image = (255 * (image - image.min()) / (image.max() - image.min())).astype(
            np.uint8
        )  # convert to uint8 if needeed (gaussian and speckle)

    image_data = Image.fromarray(image)
    image_data.save(filename)


# Add gaussian, salt & pepper, or speckle noise to an image
def add_noise(
    noise_type: Union[
        Literal["gaussian"], Literal["salt & pepper"], Literal["speckle"]
    ],
    image: np.ndarray,
```

```
):
    row, col, ch = image.shape

    if noise_type == "gaussian":
        # Gaussian parameters
        mean = 0
        sigma = 15

        # Math
        gauss = np.random.normal(mean, sigma, (row, col, ch))
        gauss = gauss.reshape(row, col, ch)
        noisy = image + gauss

        return noisy
    elif noise_type == "salt & pepper":
        # Parameters
        threshold = 0.005  # half the proportion of pixels affected
        high = 250  # salt
        low = 5  # pepper

        # Math
        noisy = np.copy(image)
        random_matrix = np.random.rand(row, col)
        noisy[random_matrix >= (1 - threshold)] = high
        noisy[random_matrix <= threshold] = low

        return noisy
    elif noise_type == "speckle":
        # Parameters
```

```
        variance = 0.25  # strength of the multiplicative noise

        # Math
        gauss = np.random.randn(row, col, ch)
        gauss = gauss.reshape(row, col, ch)
        noisy = image + image * gauss * variance

        return noisy


def standard_denoise(image: NDArray):
    if image.dtype != np.uint8:
        image = (255 * (image - image.min()) / (image.max() - image.min())).astype(
            np.uint8
        )  # convert to uint8 if needeed (gaussian and speckle)
    return cv2.fastNlMeansDenoisingColored(image, None, 10, 10, 7, 21)


# For each image URL in ./image.txt (separated by newlines), download it and apply a random noise type
# Files are saved in ./images/
if __name__ == "__main__":
    ssl._create_default_https_context = (
        ssl._create_unverified_context
    )  # needed to avoid SSL errors

    with open("images.txt", "r") as images:
        for index, image in enumerate(images):
            i = index + 1
```

```
        urllib.request.urlretrieve(image, f"images/{i}-clean.jpg")

        image = load_image(f"images/{i}-clean.jpg")

        rand = random.randint(0, 2)

        if rand == 0:
            noised = add_noise("gaussian", image)
            attempted_denoise = standard_denoise(noised)

            save_image(noised, f"images/{i}-gaussian.jpg")
            save_image(attempted_denoise, f"images/{i}-gaussian-denoised.jpg")
        elif rand == 1:
            noised = add_noise("speckle", image)
            attempted_denoise = standard_denoise(noised)

            save_image(noised, f"images/{i}-speckle.jpg")
            save_image(attempted_denoise, f"images/{i}-speckle-denoised.jpg")
        else:
            noised = add_noise("salt & pepper", image)
            attempted_denoise = standard_denoise(noised)

            save_image(noised, f"images/{i}-salt-pepper.jpg")
            save_image(attempted_denoise, f"images/{i}-salt-pepper-denoised.jpg")
```

Appendix D: Data—Responses & Scores

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response

https://doi.org/10.33774/coe-2025-6z9zx Content not peer-reviewed by Cambridge University Press.

1	Textual Noise	Ethan Washington is in James's home, specifically in the living room, which is not marble-floored. Score: 3	Ethan Washington is in James's home, in the living room, and it is not a marble-floored room. Score: 3	Ethan Washington is in James's home, specifically in James's living room, which does not have marble flooring. Score: 3	Ethan Washington is in James's living room, and it is not a marble-floored room. Score: 3	Ethan Washington is in James's home, specifically in James's living room, which is not marble-floored. Score: 3	The text you provided does not mention Ethan Irshington or his home. It appears to be describing the living room of a character named Jackos and how it is different from another character, James', living room which has marble floors. The context suggests that Ethan Wvshington	It appears the text provided is somewhat garbled and may contain typos or errors. However, based on the information given, it seems that Ethan Washington is in James' Irving room, which has a marble floor. Score: 2	Based on the context, Ethan Washington isn't explicitly mentioned as being in a specific home or roomful. The marble-floored living room described belongs to Jack. There is no clear indication of Ethan Washington's specific location in this context. Score: 0	It is not clear from the context provided whether Ethan Washington's room has marble flooring. However, we can infer that Ethan Washington is also living with Jack and Ames as they are referred to in the text. It is possible that Ethan Washington's room may have a similar design

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
							might be visiting James in his living room but there is no direct information about Ethan Irshington or his home. Score: 2			to Jack's gloom or room number 708 at the Irving ROM hotel mentioned in the context. Score: 0

2	Textual Noise	To solve this problem, we need to follow the actions described in the problem step by step. Alice starts with two red apples. Bob gives Alice one apple, so she now has 2 + 1 = 3 apples. John decides to take away three apples from Alice. Since she only had 3 apples and John took away 3, Alice is left	To solve this problem, we follow the sequence of events: Initial Apples: Alice starts with 2 red apples. Bob's Contribution: Bob gives Alice 1 apple. New total: (2 + 1 = 3) apples. John's Action: John takes away 3 apples from Alice. New total: (3 - 3 = 0) apples.	Answer: Alice has 2 apples after everything is said and done. Reasoning: Alice initially has two red apples. Bob gives her one more apple, so she now has three apples. John takes away three apples from Alice, but since he only took away two of them (leaving her with one), she	To solve this problem, we need to follow the steps given in the problem: Alice has two red apples. Bob gives Alice one apple, so now Alice has 2 + 1 = 3 apples. John takes away three apples from her, so Alice now has 3 - 3 = 0 apples left. Answer: Alice has 0 apples now.	To solve the problem, follow these steps: Initial Apples: Alice starts with two red apples. Bob Gives an Apple: Bob gives Alice one apple. Now, Alice has 2 + 1 = 3 apples. John Takes Away Apples: John takes away three apples from Alice. Calculate the remaining number of	DnsNer: Alice had two red apples to start with. Then she received one apple from Bob, giving her a total of three apples. However, John decided to take away three apples from her, which means that Alice ended up having no apples left. Score: 3	The solution to the problem is as follows: Allie starts with two red axles. BO9 gives Allie one apple, so she now has 1 + 2 = 3 apples. John takes away three apples from Allie, so she now has 3 - 3 = 0 apples. Bob and Kohl have all the apples. Therefore, Alice has 0 apples.	Let's break down the problem: Allie starts with two apples. Bob gives her one apple, so she now has three apples. John takes away three apples from her. Now, calculate the apples Alice has: Starting with 2 apples. Gain 1 apple from Bob: 2 + 1 = 3 apples.	To solve this problem, we first need to determine the number of apples that Alice starts with. The text states "Allie has two red axles." This statement is not relevant to the problem at hand and seems like a typo or error in the text. Next, the text says "BO9 gives Alice one apple." We can

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		with 3 - 3 = 0 apples. Therefore, Alice now has 0 apples. Answer: Alice now has 0 apples. Score: 3	Even though the narrative mentions that apples are delicious and that everyone loves them, this doesn't affect our calculations. Answer: Alice has 0 apples now. Score: 3	still has one apple left after John's actions. Therefore, Alice has one apple at the end. Score: 1	Score: 3	apples: 3 - 3 = 0. Answer: Alice has 0 apples. Score: 3		Answer: "Denser" Score: 3	John takes away 3 apples: 3 - 3 = 0 apples. Therefore, Alice has 0 apples. The final statement using "Denser:" seems to ask for a concise conclusion. So: Denser: 0 apples.	assume this means Bob (or Bo9) gives Alice one apple. The next sentence states "But then John decides to take away three from her." This means that John takes away three apples from Alice.

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
									Score: 3	The final sentence reads "Apple?soar; quite Delicious, are the5 not! Elects Bob, and Kohl allover apples!3How many AP}Lea does Alice have Noel" This is a nonsensical statement and does not provide any

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										information relevant to the problem at hand. The word "Delicious" in quotes could be interpreted as emphasizing the deliciousness of apples, but it does not provide any mathematical information needed to solve the problem.

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										Therefore, based on the given text, we cannot determine the number of apples that Alice has at the end. We would need additional relevant information or clarification to solve this problem.

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										Score: 0
3	Textual Noise	The text "Alice has two red apples" most likely describes the attribute "not sarcastic." Score: 3	The attribute that most likely describes the text "Alice has two red apples." is “not sarcastic.” Score: 3	Answer: "not sarcastic" is the attribute that most likely describes the given text. The text simply provides factual information without any implied tone or emotion, such as sarcasm.	It seems like you are trying to describe someone who has two red apples. The most likely attribute to describe this person based on your example would be "fruitful" or "fruit lover."	The statement "Alice has two red apples" is most likely described as "not sarcastic." Score: 3	It is not clear what you are asking, but it seems like you might be trying to describe the sentiment of a statement using an ontology or some other form of structured data. If you could	I'm sorry, it seems like there may have been a mistake or typo in your question. Could you please clarify or rephrase it so I can provide a helpful response? Score: 0	I'm sorry, but I couldn't quite understand your request. Could you please provide more context or clarify? Score: 0	It is not clear what you are asking for, as your question contains a lot of unrelated information and seems to be missing context. Could you please provide more information or clarify your

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				Score: 3	Score: 0		provide more context or clarify your question, I would be happy to help. Score: 0			question? I will do my best to help you if you can provide further details. Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
4	Textual Noise	The attribute "sarcastic" most likely describes the text, as it appears to be making a sarcastic comment about Alice's activities. Score: 3	The text is most likely "sarcastic." Score: 3	Based on the context provided in the text, it is more likely that the attribute describing it would be "sarcastic." The phrase "really loves sitting alone at home doing nothing" suggests a negative connotation	The following text is most likely to be interpreted as "sarcastic." Score: 3	The text is most likely "sarcastic." The statement suggests an unlikely enthusiasm for sitting alone doing nothing, which is typically not something people are genuinely excited about.	The attribute that is most likely to describe the text "not sarcastic" is the phrase "0arcastxc". This is because the text seems to be expressing a positive sentiment about Alice's love for sitting alone at	The attribute most likely to describe the text provided is "not sarcastic". Score: 0	The phrase “I’m Sir Alice really loves sitting alone at home doing nothing!” is more likely to describe “not sarcastic” rather than “acoustics.” The text appears to express a genuine affinity for solitude and inactivity,	It is most likely that the text describes a "not sarcastic" attribute. The text talks about Alice loving spending time alone and doing nothing in Hickman, which indicates a sincere or genuine sentiment rather

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				and implies that Alice may not enjoy being by herself or engaging in leisurely activities, which could be interpreted as sarcasm. Score: 3		Score: 3	home, which is not sarcastic in nature. Score: 0		lacking any elements related to sound properties. Score: 1	than sarcasm. Additionally, there is no mention of sound or acoustics in the text provided. Score: 1

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
5	Textual Noise	Based on the premise that "Bob's plant died," the most plausible alternative is Alternative 2: "Bob failed to water the plant enough." This is a common reason for plants dying, as insufficient watering can lead to	Given the premise "Bob's plant died," Alternative 2 is the most plausible explanation: "Bob failed to water the plant enough." Plants do not make decisions about where they live, so Alternative 1 is not plausible. Lack of proper	Based on the premise, Alternative 2 is more plausible than Alternative 1. The premise states that "Bob's plant died," but it does not provide any information about the plant's emotions or thoughts. Therefore,	Alternative 2: "Bob failed to water the plant enough." Score: 3	Based on the premise "Bob's plant died," the most plausible alternative is "Alternative 2: Bob failed to water the plant enough." This directly explains why a plant might die, unlike the first alternative, which suggests the plant had a	Based on the premise provided, "Alternative 2: Bob failed to care for the plant enough." is more plausible than "Alt>rnative h: The plan didn't like living in Bob's home." because it provides a reason why	Given the premise "Bob's plant died," the most plausible alternative would be "Alternative 2: Bob failed to water the plant enough." This alternative directly relates to the plant's death and is a common reason	Based on the premise "Bob's plan died," the most plausible alternative is: Alternative 2: "Bob failed to water the plant." This alternative suggests a reason why Bob's plan (or plant) failed,	To provide a coherent answer, I need to understand the context and make sure that all the words in your question are spelled correctly. Please correct any spelling errors or clarify the context so I can assist you better.

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		dehydration and ultimately the death of the plant. Alternatively, the notion that "the plant decided it didn't like living in Bob's home" is less believable and lacks a scientific basis for explaining the plant's death.	care, like insufficient watering, is a common reason for a plant's death. Score: 3	it would be reasonable to assume that a lack of care or proper maintenance, such as failing to water the plant enough, could have contributed to its death. On the other hand, Alternative 1 is an		preference for living conditions. Score: 3	Bob's plan might have died, which aligns with the given premise. Score: 3	for a plant dying. Score: 3	aligning with the premise. Score: 3	Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		Score: 3		anthropomorphism, attributing human-like emotions and intentions to the plant, which does not align with the scientific understanding of plants. Score: 3						

6	Textual Noise	Based on the premise that "I have a new home," Alternative 1: "I bought a house a month ago and it's finally ready" is the most plausible option. Alternative 2, which suggests that a new house magically appeared, is not likely to happen in reality. Score: 3	Based on the premise "I have a new home," Alternative 1 is the most plausible. It suggests a logical and common scenario where someone buys a house, and it is ready after some time. In contrast, Alternative 2 involves a magical occurrence, which is far less likely in reality.	Alternative 1 is the most plausible given the premise. It provides a reasonable explanation for having a new home based on common circumstances, such as purchasing a property or renting it from someone else. Alternative 2 is not plausible because it involves an unrealistic and fantastical	The most plausible alternative given the premise "I have a new home" is "Alternative 1: I bought a house with a mortgage and it's finally ready." This option aligns with the conventional way people acquire new homes through the process of purchasing a property with a mortgage and waiting for it to	Given the premise "I have a new home," the most plausible alternative is: Alternative 1: "I bought a house with a mortgage and it's finally ready." This alternative describes a realistic scenario, while Alternative 2 involves a magical occurrence,	Given the premise "I have a new home," Alternative 1: "I bought a house and it's finisheD is more plausible than Alternative 2: "I woke up one morning and found a new house had appeared." because buying a house is a common reason for acquiring a new home, while the second	Based on the context and the coherence of the text, the most plausible option is: Alternative 1: "I bought a house in Montenegro and it's really nice. I'm ready." This option seems to make the most sense given the mention of buying a house and the positive description of it.	It seems like the text has some garbled or unintelligible parts. Based on what is clear, here is a comparison of the two alternatives: Alternative 1: "I bought a house at Montauk and it's finally ready." This sentence is coherent and grammatically correct.	It is not entirely clear what the question is asking for, as it appears to be missing context or relevant information. Please provide more details so I can better understand and assist with your inquiry. Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
			Score: 3	scenario that defies physical laws and human experiences. Score: 3	be ready for move-in. Score: 3	which is less plausible. Score: 3	alternative suggests an unlikely and fantastical scenario. Score: 3	Score: 3	Alternative 2: "I woke up one morning and found a madhouse headscarf appeared." This sentence is less coherent, with an unusual phrase "madhouse headscarf." Given this comparison,	

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
									Alternative 1 is the more plausible option, as it is clear and understandable. Score: 3	

7	Textual Noise	John is in Bob's home, in the living room, which is brown. Score: 3	John is in Bob's home, in the living room, and the room is brown. Score: 3	John is in Bob's home, specifically in his living room, which is brown in color. The only exception to this is Bob's dining room, which is bright yellow. Score: 3	John is in Bob's house, specifically in the living room. The living room is brown, like the rest of Bob's house. Score: 3	John is in Bob's home, in the living room, and the room is brown. Score: 3	Based on the context provided, John's home is not described or identifiable. Therefore, we cannot determine from the given information which specific room John is in, nor its color. All we know is that John is visiting Bob and is currently sitting in Bob's living room. Score: 3	John's 6ame is John, he is i8 Bob's living room, and the color of the `or is not sp8cified. Score: 2	John is in Bob's house, specifically in the living room. The color of the room is brown. Score: 3	It seems like there's been a mix-up with some of the letters in your message. Let me try to decipher it as best I can: * Bob's house has brown walls. * Rugby maneuvering INS9e means that Bob's house is also on top of a hill, which makes it harder for someone to reach him by

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										running up the hill. * Bob's house has a bright yellow dining room. * John visited Bob and Ibo (who may be another person in Bob's household) in their living room. * John Zn is likely another

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										person who is not currently at Bob's house, as there is no mention of him being present. * The room John is in is the living room, which matches with what he was visiting. Score: 3

8	Textual Noise	To solve this problem, we need to calculate the total cost of the ice cream cones first. Jack buys four chocolate ice cream cones for his friends, and one for himself. Each ice cream cone costs $2.00. So, the total cost of all the ice cream cones can be calculated as 5 ice cream cones	First, we need to calculate the total cost of the ice cream cones. Jack buys four ice cream cones for his friends and one for himself, making it a total of 5 cones. Each ice cream cone costs $2.00, so: [5 \text{ cones} \times $2.00/\text{cone} = $10.00] Jack's original allowance is	To determine if Jack has enough money to buy the bubble gum, we need to first calculate the total cost of the ice cream cones and then subtract that amount from his allowance. We can do this by finding the sum of the number of ice cream cones bought and their individual prices:	To solve this problem, we need to calculate the total amount Jack spent on the ice cream cones and how much money he has left to determine if he can afford the bubble gum. Given: Cost of one chocolate ice cream cone = $2.00 Jack buys 4 ice cream cones for	To determine if Jack has enough money to buy the bubble gum, we need to calculate his total expenses after buying the ice cream cones and then see if he has enough left to purchase the gum. Initial Allowance: Jack starts with $10.50. Cost of Ice Cream Cones:	K Answerr: G Reasoning: Jack received an allowance of $10.50 and wants to buy four chocolate ice cream cones for his friends, as well as one for himself. Each cone costs $2.00, so the total cost for the five cones is 4 cones * $2/cone + 1 cone * $2/cone = $8. Since Jack has $10.50 and	To solve this problem, we need to determine if Jack has enough money to buy that gum after buying the ice cream cones. Total cost of ice cream cones: 4 cones for friends = 4 x $2.00 = $8.00 1 cone for himself = $2.00 Total cost of ice cream cones = $8.00 + $2.00 = $10.00	To determine if Jack can buy gum, we need to calculate his total expenses on ice cream and compare it to his initial amount of money. Initial Amount: Jack has $10.50. Cost of Ice Cream Cones: Jack buys 4 cones for his friends and 1 for himself; so	Let's break down the problem and write HQ reasoning for each step: 1. First, we need to determine how much Jack has spent on ice cream cones. He bought four for his friends and one for himself. The cost of each cone is $2.00. So, the total amount he spent

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		* $2.00 = $10.00. Next, we add the cost of the gum, which is $0.75. Therefore, the total cost of the ice cream cones and the gum is $10.00 + $0.75 = $10.75.	$10.50. After buying the ice cream cones, he spends $10.00, so he has: [$10.50 - $10.00 = $0.50] Now, we check if he has enough money to buy a pack of bubble gum, which costs $0.75. Comparing his	Number of ice cream cones = 4 (for friends) + 1 (for himself) = 5 Total cost of ice cream cones = Number of ice cream cones * Price per cone = 5 * $2.00 = $10.00 Now, we need to subtract the total cost of the	his friends and 1 for himself, so the total cost of ice cream cones = 5 cones * $2.00 = $10.00 Jack's initial allowance = $10.50 Now, let's calculate how much money Jack has left after buying the ice cream cones:	Jack buys 5 ice cream cones (4 for his friends, 1 for himself). Each cone costs $2.00. Total cost for the cones: (5 \times 2.00 = 10.00). Remaining Money after Ice Cream:	wants to spend $8 on ice cream cones, he will have $10.50 - $8 = $2.50 left over for other purchases. Jack also wants to buy a pack of bubble gum that costs $0.75. To determine if Jack has enough money to buy the bubble gum, we can subtract	Total cost of all ice cream cones and gum: Total cost of ice cream cones ($10.00) + cost of one pack of gum ($0.75) = $10.00 + $0.75 = $10.75 Jack's total expenditure: USP for Lance ($10.50) - Total cost of all ice	he buys a total of 5 cones. Each cone costs $2.00. Total cost for the cones: (5 \times 2.00 = 10.00) Remaining Money After Ice Cream: Subtract the cost of the cones from the initial amount:	on ice cream cones is 5 * 2 = <<5*2=10>>$ <<10_dollar|US dollars>>. 2. Next, we need to determine how much Jack has left after spending money on ice cream cones. He originally had $75.00 and he

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		Since Jack has an allowance of $10.50, he does not have enough money to buy the gum. Answer: No, Jack does not have enough money to buy the gum. Score: 3	remaining amount with the cost of the gum: [$0.50 < $0.75] So, Jack does not have enough money to buy the bubble gum. Answer: No, he does not have enough money to buy the gum.	ice cream cones from Jack's allowance to find out how much money he has left: Money left for bubble gum = Allowance - Total cost of ice cream cones = $10.50 - $10.00 = $0.50	$10.50 (initial allowance) - $10.00 (spent on ice cream cones) = $0.50 We can see that Jack has $0.50 remaining after buying the ice cream cones. The cost of one pack of bubble gum is $0.75. Since Jack only	He spends $10.00 on ice cream cones. Remaining money: (10.50 - 10.00 = 0.50). Cost of Bubble Gum: One pack of bubble gum costs $0.75. Check if Jack Can Buy the Gum:	the cost of the ice cream cones and bubble gum from his original allowance: $10.50 - $8 - $0.75 = $1.75. Since Jack's remaining allowance is only $1.75, he does not have enough money to buy the pack of bubble gum.	cream cones and gum ($10.75) = $10.50 - $10.75 = -$0.25 Since the above calculation results in a negative value, it means Jack does not have enough money to buy the gum after buying the ice cream cones	[10.50 - 10.00 = 0.50] Cost of Gum: The gum costs $0.75. Can Jack Buy the Gum? Jack needs $0.75 for gum, but he only has $0.50 left. Therefore, he does not have	spent $10.50 on ice cream cones. So, the amount of money he has left is 75 - 10.5 = <<75-10.5=64.5>>$<<64.5_dollar|US dollars>>. 3. Finally, we need to determine whether Jack has enough

https://doi.org/10.33774/coe-2025-6z9zx

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
			Score: 3	Since a pack of bubble gum costs $0.75, Jack does not have enough money to buy it. Therefore, the answer is: No, he does not have enough money to buy the bubble gum. Score: 3	has $0.50 left, he does not have enough money to buy the bubble gum. Answer: No, he does not have enough money to buy the bubble gum (he is $0.25 short). Score: 3	Jack needs $0.75 for the gum but only has $0.50 left. Answer: Jack does not have enough money to buy the bubble gum. Answer: G. No, Jack does not have enough money to buy the gum.	Therefore, the answer is G, which means that Jack does not have enough money to make all the purchases he wants to make. Score: 1	for his friends and himself. Therefore, the answer is: Kans: Jack does not have enough money to buy the gum. Score: 3	enough money to buy the gum. Answer: G: No, Jack does not have enough money to buy the gum. Score: 3	money to buy postcards. The cost of each postcard is $0.75. He wants to buy 2 postcards. So, the total amount he needs to spend on postcards is 2 * 0.75 = <<2*0.75=1.5>>$<<1.5_dollar|US dollars>>.

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
						Score: 3				4. We can now check whether Jack has enough money to buy postcards by comparing the total amount he has left after spending on ice cream cones with the total cost of the postcards. He has $64.50 left, and he

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										needs $1.50 for the postcards. So, Jack does have enough money to buy postcards because 64.5 > 1.5. 5. To summarize our HQ reasoning, we found that Jack initially had $75.00, spent $10.50 on ice cream

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										cones, and has left with $<<75-10.5=64.5>>$<<64.5_dollar|US dollars>>. He wants to buy 2 postcards that cost a total of <<2*0.75=1.5>>$<<1.5_dollar|US dollars>>. Since 64.5 > 1.5, Jack has enough money

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										to buy the postcards. Answer: Yes, Jack does have enough money to buy postcards because he has $<<75-10.5=64.5>>$<<64.5_dollar|US dollars>> left after spending on ice cream cones and needs only

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										<<2*0.75=1.5>>$<<1.5_dollar|US dollars>> to buy 2 postcards. Score: 1

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
9	Textual Noise	Yes Score: 0	Cannot be determined. Score: 3	Yes, Bob is younger than Chase. Score: 0	Cannot be determined Score: 3	To solve this, let's break down the information: Chnseois is younger than Emma. Emma is older than Bob + 8. We don't have any direct information about Chnseois and Bob's relative ages, so we can't	I'm sorry, but I am not sure what you are asking. Can you please provide more context or clarify your question? Additionally, the text appears to contain some errors which might make it difficult for me to	Chinese's=younger than Emma.&Jam is otter than Bob+8Is Bob Bounden than Chase3G (Resold with ?SS, no, or cannot be determined.) Score: 3	There seems to be some confusion in the text, but I'll try to interpret it for you: Chinese < Emma (meaning Chinese is younger than Emma). Jam > Bob + 8 (meaning Jam is	It seems like there is a mix of English language and some other words that might not make sense in the context provided. If you could please clarify what you mean by "Chinese's" and "Jam is otter than Bob+8,"

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
						determine if Bob is younger than Chnseois based on the given statements. Therefore, the answer is: cannot be determined. Score: 3	understand your question fully. If possible, could you please correct any mistakes in your question and resubmit it? Score: 0		older than Bob by 8 years). Is Bob older than Chase? From the given information, we cannot directly determine the relationship between Bob and Chase. You would need more information about Chase's age or their	I would be happy to help you answer your question about whether or not Bob is bounden than Chase 3G with a ?SS. Score: 0

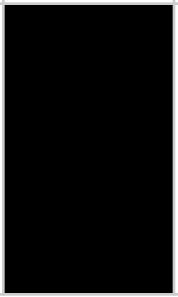
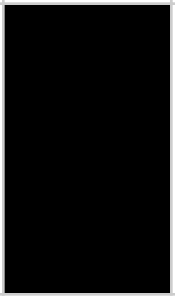
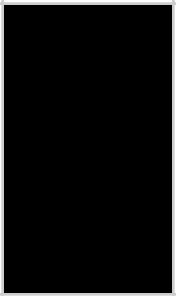
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
									relative ages to each other. So the answer would be: cannot be determined. Score: 3	

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
10	Textual Noise	Yes Score: 0	No, Bob is not younger than Chase. Chase is younger than Emma, and Emma is younger than Bob, so Bob is the oldest. Score: 3	No, Bob is not younger than Chase. The statement says that Emma is younger than Bob and Chase is younger than Emma, so Bob must be older than both of them. Score: 3	Cannot be determined. Score: 0	Cannot be determined. The information provided doesn't establish a clear comparative relationship between Bob and Chase. Score: 0	Without more context or information about the relationships and ages of the individuals mentioned (Chase, Emma, BBz, Bob, Chyse), it is impossible to provide an accurate answer to your	Yes Score: 0	Cannot be determined. Score: 0	Based on the information provided in your question, it is not possible to determine if Bob is younger than Chase. The only relationship mentioned between Emma and BBB is that Emma is younger than BBB, but

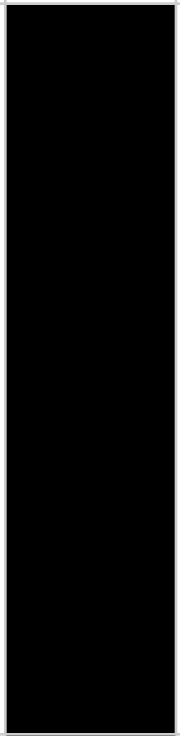
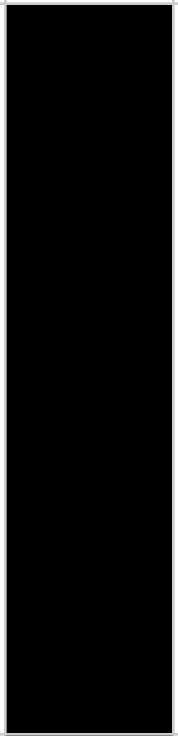
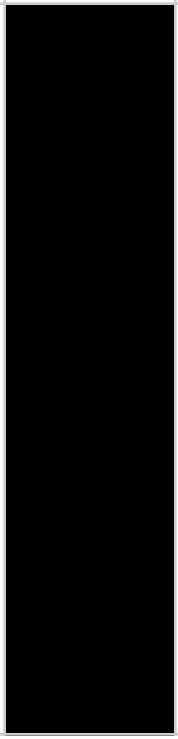
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
							question. Please provide additional details or clarify your query. Score: 0			there is no information about how old either Chase or Bob are in comparison to each other or Emma. Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
11	Gaussian Noise		The man's shirt has the word "TEXAS" on it. Score: 3	The image shows a man wearing a long-sleeve shirt with the letters "TX" printed on it. These letters are commonly associated with Texas, so it can be inferred that the state of Texas is represented on his		The man's shirt has "TEXAS" written on it. Score: 3	The image you provided does not show a state name on the man's shirt. Instead, it appears that he is wearing a sweatshirt with what seems to be a "T" logo, which might represent a university or sports team.		The man's shirt has the word "TEXAS" on it. Score: 3	The image shows a man with "Texas" written across his shirt, which indicates that he is wearing an item of clothing associated with the state of Texas. Score: 3

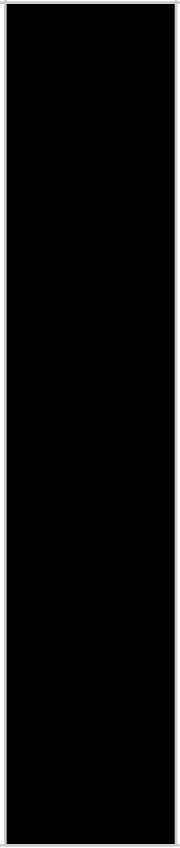
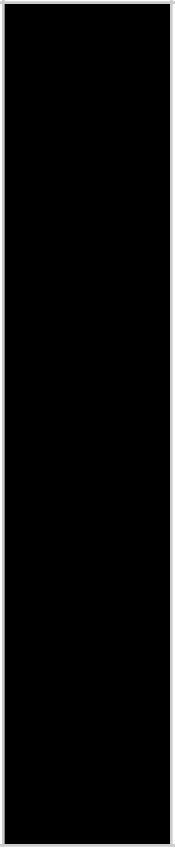
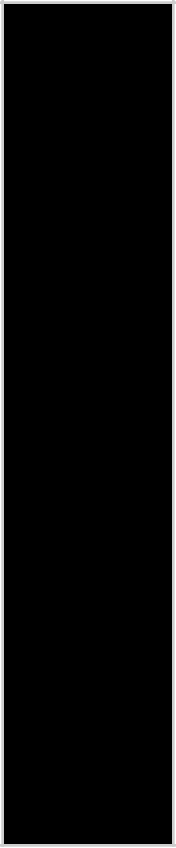
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		shirt. Score: 3	[redacted]		However, without more context or a clearer view of the design on his shirt, I am unable to determine which specific entity the "T" represents, nor can I identify the state it's associated with if that is	[redacted]		

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
							indeed what you are asking about. Score: 0			

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
12	Speckle Noise		The sign says "STOP FUNDING WAR!" Score: 3	The sign in the image says "STOP FUNDING WAR!" It has been added to a standard red stop sign, which is why you can also see the word "STOP" on it. The message appears to be advocating for		The sign says "STOP" at the top, and below that, it says "FUNDING WAR!" Score: 3	The image shows a stop sign with an additional message below it that reads "STOP FUNDING WAR." This suggests that the person or group responsible for adding this message are advocating		The sign says "STOP FUNDING WAR." Score: 3	The sign says "STOP" and there is additional text below it that reads "FUNDING WAR!" It appears to be a form of protest or activism, urging people not only to physically stop at the

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				an end to war funding. Score: 3			against war funding. The photo seems to be taken in an urban area with buildings in the background, and there is also a street sign visible above the stop sign. Score: 3			intersection but also advocating for the cessation of funding war. Score: 3

13	Speckle Noise	[redacted]	The man is performing a skateboarding trick where he is in mid-air, catching the skateboard with one hand while his body is lifted off the ground. His posture suggests he is in the middle of executing a trick, with one foot on the board and the other assisting in the maneuver.	The man in the image appears to be performing a skateboarding trick, specifically an ollie or a variation of it, which involves jumping with the skateboard without using hands to hold onto it. He's caught mid-air, suggesting he is executing the maneuver with skill and control. His	[redacted]	The person is performing a skateboarding trick called an ollie. This involves popping the skateboard into the air without using hands, likely against a vertical wall or surface in this context. Score: 3	The man in the image appears to be performing a skateboard trick. He is captured mid-air, with his body facing towards the ground and his arms extended outward for balance. The skateboard is below him, suggesting that he has launched off of an obstacle or ramp, and is executing a	[redacted]	The man is performing a skateboard trick in mid-air, lifting the skateboard off the ground while likely jumping and using his feet to manipulate the board. Score: 3	The man in the image appears to be performing a skateboarding trick. He is airborne with his skateboard, which suggests that he has either launched off a ramp or an inclined surface and is now executing a maneuver such as a flip or spin before landing back on the ground or another surface. The exact

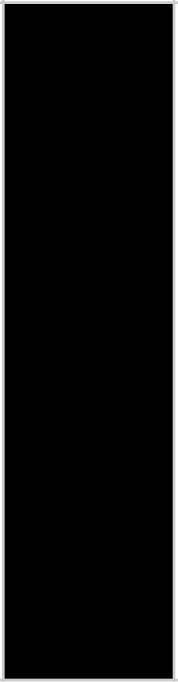
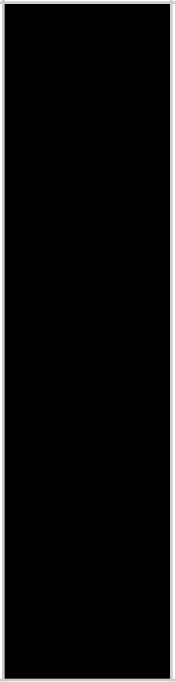
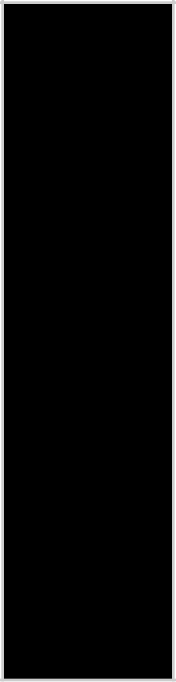
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
			Score: 3	body posture indicates that he has good balance as he maneuvers his board in the air. Score: 3			maneuver where he rotates his body and manipulates the board in some way before landing back on it or on another surface. The exact nature of the trick cannot be determined without seeing the full sequence of			nature of the trick isn't clear from this single frame, but it involves skillful control of the skateboard and good balance to maintain airtime and execute the intended move. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]			[redacted]		movements involved. Score: 3	[redacted]		
14	Gaussian Noise	[redacted]	The train is purple. Score: 3	The train in the image is purple. Score: 3	[redacted]	The train is purple. Score: 3	The train in the image is purple. Score: 3	[redacted]	The train is purple. Score: 3	The train in the image is purple. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
15	Speckle Noise		The bananas in the image are not ripe. They are green, indicating they need more time to ripen. Score: 3	The bananas in the image appear to be unripe. Ripe bananas typically have a yellow skin with small brown spots, while unripe bananas are green and firm. The ones in the photo show signs of being early in their		The bananas in the image are not ripe; they are still green. Score: 3	The bananas in the image are not yet ripe. They appear to be green, which indicates that they are still unripe and will need more time to mature before they turn yellow and reach their optimal sweetness and		The bananas are not ripe; they are still green. Score: 3	The bananas in the image appear to be unripe, as they are still green. Ripe bananas typically have a yellow color with small brown spots, which is not present here. The green color indicates that these bananas

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		ripening process. They will likely become sweeter and softer as they continue to ripen over time. Score: 3	[redacted]		texture for consumption. Score: 3	[redacted]		will need more time on the plant before they can be harvested and consumed, as their flavor and nutritional content would be at an optimal level when fully ripe. Score: 3

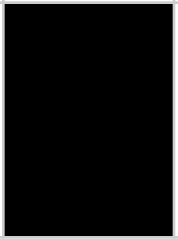
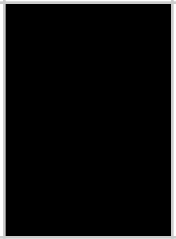
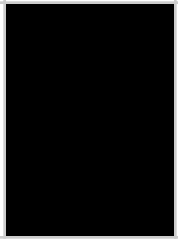
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
16	Salt-Pepper Noise		There are nine horses in the image. Score: 3	In the image provided, I can see several horses, but it's not possible for me to give you an exact number without seeing the entire scene or knowing the resolution of the image. There are at least four horses		There are eleven horses in the image. Score: 0	I can't count the horses in this image because it appears to be pixelated or distorted, making it difficult to discern individual details. If you have a clearer image, I would be able to help you better.		There are eleven horses in the image. Score: 0	There are five horses visible in the image. Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				clearly visible in the foreground, with their backs towards the camera, and there may be more horses behind them as well. Score: 0			Score: 0			

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
17	Speckle Noise		The grass appears to be taller than the baby elephant. Score: 3	The grass appears to be taller than the baby elephant. The baby is standing near the adult elephants in a field with vegetation that reaches above their legs. Score: 3		The grass is taller than the baby elephant. Score: 3	The grass is taller than the baby elephant. The adult elephant in front of the baby elephant is obscuring it slightly, but you can still see that the grass is higher than the height of the small elephant. Score: 3		The grass appears to be about the same height as the baby elephant. Score: 0	The grass in the image appears to be taller than the baby elephant. The baby elephant is standing behind a clump of tall grass, and its height does not reach above this grass. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
18	Gaussian Noise	[redacted]	The telephone handset is black. Score: 3	The telephone handset in the image is black. Score: 3	[redacted]	The telephone handset is black. Score: 3	The image provided does not show a clear view of the phone's handle. However, based on the context that it is an office environment with a computer and keyboard in front of you, one would typically find phones	[redacted]	The telephone handset is black. Score: 3	The telephone handset in the image is black. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]			[redacted]		with handles in shades of black or silver. Without more specific details from your image, I cannot provide the exact color of the telephone handset. Score: 0	[redacted]		
19	Speckle Noise	[redacted]	The umbrella is composed of	The umbrella in the image is composed of	[redacted]	The umbrella is composed of alternating	The umbrella in the image is blue and white.	[redacted]	The umbrella is composed of	The umbrella in the image is composed of

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
			blue and white stripes. Score: 3	white and blue stripes. Score: 3		black and white stripes. Score: 1	Score: 3		white and dark blue stripes. Score: 3	black and white stripes. Score: 1

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
20	Salt-Pepper Noise		The animal has horns located directly above its ears. Score: 3	The image shows a cow with large, curved horns located directly above its ears. Score: 3		Directly above the animal's ears are two horns. Score: 3	The image shows a cow with its ears. Directly above each ear, you can see the cow's forehead area and possibly part of its horns. The cow appears to be standing outside under what seems like an overcast or		Directly above the animal's ears, there are horns. Score: 3	The image shows a bull with horns located directly above its ears. The horns are typically used for identification, as they can serve different purposes depending on the species, such as fighting or defense.

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]			[redacted]		cloudy sky. If there are other objects or details that were supposed to be directly above the animal's ears that are not clearly visible in this image, please provide more context so I can assist you better. Score: 3	[redacted]		In this case, it appears to be a domestic cattle bull, and horns like these are often found in male members of the species. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
21	Salt-Pepper Noise		The man is cutting a cake on a table. Score: 3	The man appears to be at a celebration or event, possibly related to an anniversary given the context provided in the text description. He is standing behind a table that has multiple items on it, including		The man is cutting a cake at a table that is set with various items, including another cake, a watermelon, and greeting cards. Score: 3	The man in the image appears to be cutting a piece of cake from what looks like a celebration or event, possibly a religious one given his attire which resembles that of a priest. He's using a knife and seems		The man is cutting a cake on a table that is covered with a tablecloth. There are greeting cards and another decorated cake or possibly carved item nearby. Score: 3	The man in the image appears to be a clergy member, possibly a priest or a bishop, based on his attire which includes religious vestments and a pectoral cross. He is standing at a table with an assortment of

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				what looks like a sheet cake with writing on the top. The man seems to be either cutting the cake or about to cut into it, using a knife. There are also place cards and possibly some other decorative items on the table. It's not clear if he is			to be making sure to cut a neat slice from the cake. Score: 3			desserts, including what looks like a cake with candles that may indicate a celebration such as a birthday or anniversary. The man is using a knife to cut into one of the cakes. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		making a speech or simply engaging in the process of serving the cake. Score: 3	[redacted]			[redacted]		

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
22	Gaussian Noise		The street name is Day City Road. Score: 3	The street name in the image is "Route 321." It's written on a green rectangular sign above another street name, which is "Day City Rd." Both signs are attached to a stop sign post. The second sign indicates		The street name is Day City Road. Score: 3	The street signs in the image show "Day City Rd" and "Route 372." Score: 2		The street name is Day City Road. Score: 3	The street sign in the image indicates that it's "Day City Road." Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		that Day City Road intersects with Route 321. Score: 2	[redacted]			[redacted]		

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
23	Salt-Pepper Noise		The pizza is not vegetarian because it has visible slices of pepperoni, which is a meat product. Vegetarian options would not include meat toppings like pepperoni. Score: 3	Based on the image, it appears that the pizza may be considered vegetarian because there are visible vegetable toppings such as red onions and bell peppers. However, it is important to		This pizza is not vegetarian because it contains pepperoni, which is a type of meat. A vegetarian pizza would not include any meat products. Score: 3	The pizza in the image is not entirely vegetarian. It contains toppings that are typically considered meat, such as pepperoni, which is a type of processed meat that includes beef and sometimes pork.		The pizza is not vegetarian because it contains pepperoni, which is a type of meat. Score: 3	The image shows a pizza with several toppings. It appears that there are sliced vegetables on it, which would make it vegetarian if that is all it has. However, without seeing a complete view of all the

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		note that without knowing the specific ingredients of the sauce, cheese, and other possible toppings not clearly visible in the photo, one cannot definitively say if it's entirely vegetarian or	[redacted]		Additionally, it has vegetables like onions and green peppers, but these are mixed with the meat toppings. To be vegetarian, a pizza would need to include only plant-based ingredients and no meat or animal	[redacted]		ingredients or knowing the specific topping varieties, it's not possible to definitively confirm whether this pizza is vegetarian. If it contains cheese or any other animal-derived products, it would not be

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		not. If any non-vegetarian components are present, they would likely be hidden within the ingredients or mixed into the sauce, which is not apparent from this view. Score: 0	[redacted]		byproducts. Score: 3	[redacted]		considered vegetarian. Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
24	Salt-Pepper Noise		The snowboard says "Burton." Score: 3	The skateboard in the image says "MURPHY." Score: 0		The snowboard says "burton." Score: 3	The image you've provided is of a snowboard, not a skateboard. The snowboard has text on it, but due to the quality and angle of the photo, I'm unable to clearly read the text. If there's something specific		The skateboard says "Burton." Score: 3	The image shows a young child holding a snowboard, not a skateboard. The snowboard has "Burton" written on it. Burton is a well-known brand that specializes in snowboards and related gear. Score: 3

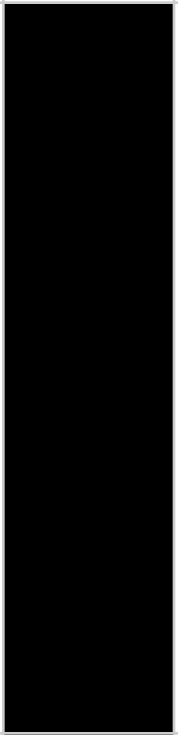
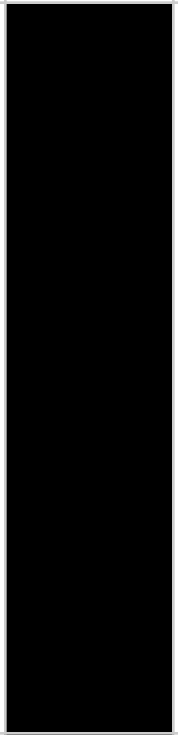
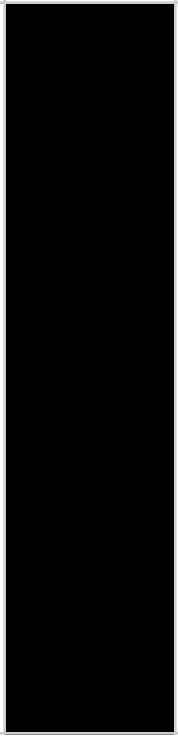
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]			[redacted]		you're interested in about the snowboard or if you have another question related to the image, feel free to ask! Score: 0	[redacted]		

25	Speckle Noise	[redacted]	No, this is not a normal sight. The image shows people performing a stunt on a motorcycle, with multiple individuals stacked on top of each other. Such stunts are unusual and generally performed by professionals under controlled conditions. It is important to prioritize safety and adhere to traffic laws	This image depicts an unusual and potentially dangerous situation. It shows several individuals piled on top of each other on what appears to be the back of a motorcycle or moped, with one person seemingly falling off. This is not a typical scene as it presents multiple safety risks:	[redacted]	This doesn't appear to be a normal sight. The image shows a person riding a motorcycle with several mannequins or figures attached, which is unusual and likely staged for artistic or humorous purposes. It's not something typically seen in everyday life. Score: 3	The image shows several people on what appears to be an overloaded motorcycle or scooter. This is not a typical scene and would likely be considered unsafe as it violates most traffic laws regarding the number of passengers on a single vehicle, especially without proper safety measures	[redacted]	This isn't a normal sight. It looks like a group of people is performing a stunt or a balancing act on a motorcycle. Such acts require skill and should be done with strict safety measures. Score: 3	This image appears to be digitally altered or staged for comedic effect, as it shows two individuals riding on a motorcycle in an unusual and exaggerated manner. The person in the back is not seated properly but rather positioned upright, which defies the laws of gravity and balance that would normally

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]	when riding a motorcycle. Score: 3	overloading the vehicle, unstable balance that could lead to accidents, and potential injury to those involved due to improper handling of both the vehicle and the passengers. It's important for everyone to	[redacted]		like helmets or appropriate seating arrangements. It's important for all road users to follow traffic regulations to ensure their own safety and that of others. Score: 3	[redacted]		govern such situations. In real life, this kind of sight would be highly irregular and unsafe due to the risk of injury during an accident or if the motorcycle were to make a sudden stop. It seems designed for humor or artistic

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		follow traffic rules and safety guidelines when using any mode of transportation. Score: 3	[redacted]			[redacted]		expression rather than depicting a realistic scenario. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
26	Salt-Pepper Noise		The time on the clock is approximately 10:08. Score: 0	The clock in the image shows that it's approximately 12:50. The minute hand is pointing towards ten minutes past twelve, and the hour hand is just touching twelve, suggesting that it's a few		The time on the clock is 10:10. Score: 0	It's just past 12:05, as indicated by the hands on the clock in the image. Score: 0		The time on the clock is approximately 10:09. Score: 0	The image shows a clock with Roman numerals, and the hands are pointing approximately to the number six. However, without an exact reference for what the hour hand should be pointing to, I

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				seconds after noon or midnight, depending on whether it's day or night at this location. Score: 0						cannot provide an accurate time based on this image alone. If you need the current time or have specific questions about the clock, feel free to ask! Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
27	Speckle Noise		The tie has three distinct colors: red, brown, and cream. Score: 3	The tie has two distinct colors: red and a darker brown or maroon. The red is used for the circles on the tie, while the background color is the darker shade. Score: 0		The tie has three distinct colors: red, dark brown, and cream. Score: 3	The tie in the image has at least three distinct colors: red, white (or possibly a very light color that might appear gray due to the image quality), and black. These colors are present in the design of the tie, which features circular		The tie has three distinct colors: red, white, and brown. Score: 3	The tie in the image has two distinct colors: red and white. The pattern consists of circles with a combination of these colors Score: 0

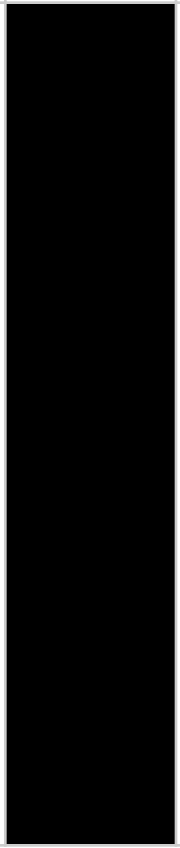
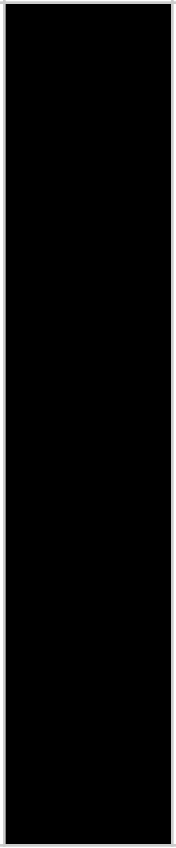
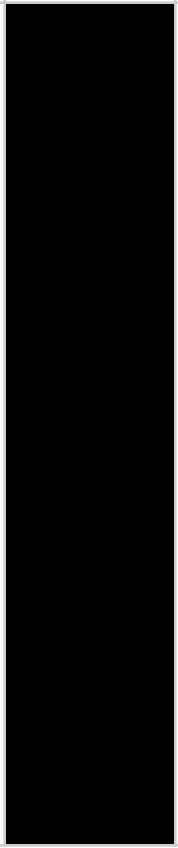
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]			[redacted]		patterns. The background also appears to be a neutral tone, but it's difficult to discern if there are any more colors due to the quality of the image. Score: 3	[redacted]		

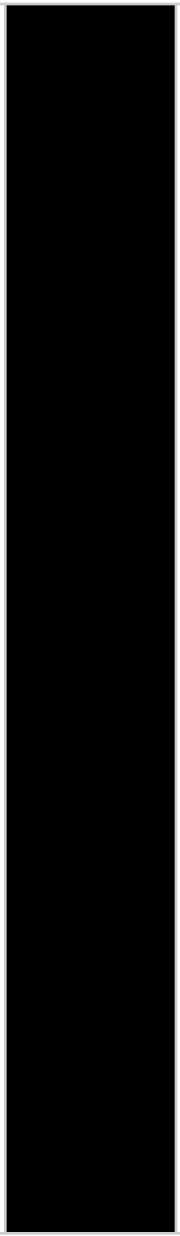
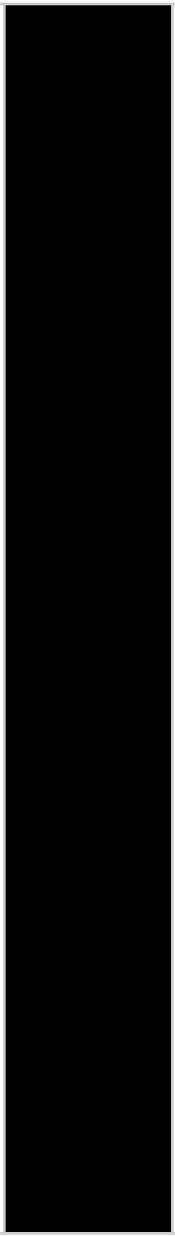
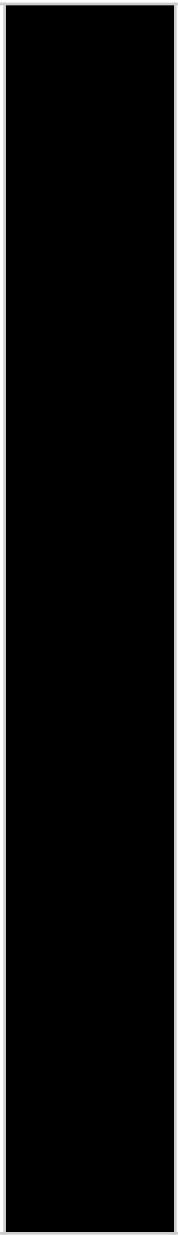
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
28	Speckle Noise		The weather appears to be rainy, as indicated by the presence of wet streets and people using umbrellas. Score: 3	The image shows a person holding up what appears to be a floral branch with small flowers, possibly magnolia blossoms. In the background, it's raining as evidenced by wet roads and people using umbrellas.		The weather appears to be rainy, as the ground looks wet, and someone is holding an umbrella. Score: 3	The image shows a person holding what appears to be a branch with white flowers, possibly cherry blossoms. The weather seems to be rainy as there's an umbrella visible in the background and the ground looks		The weather appears to be rainy, as evidenced by the presence of umbrellas and wet ground. Score: 3	The image shows rain, as indicated by a person holding an umbrella and the wet ground. The flowers are also drooping, which suggests that it might be raining or has recently rained. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		There are also trees visible with green leaves, suggesting that this photo was taken in spring or summer when such flowers typically bloom. The sky is overcast, indicating a gloomy or rainy day.	[redacted]		wet. Additionally, the visibility is not great due to raindrops on the camera lens or glass, which is common during rain. Score: 3	[redacted]		

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		Score: 3	[redacted]			[redacted]		
29	Gaussian Noise	[redacted]	There is an image of what appears to be the Virgin Mary holding baby Jesus in the center of the image. Score: 3	There is a religious figure visible in the image. It appears to be a card or book cover featuring an icon of Mary and Jesus. The presence of such images is common in various	[redacted]	The image contains a depiction of religious figures that appear to be the Virgin Mary holding baby Jesus. Score: 3	In the image you've provided, I don't see a figure of any religious significance. However, I can see various books on a shelf and a pair of scissors resting against one of the books. If	[redacted]	The image includes what appears to be a depiction of a religious figure with a halo, likely representing a saint. The specifics aren't clear enough to determine who it is.	The image is not clear enough for me to definitively identify any religious figures or symbols. If you can provide a clearer image or describe the figures, I would be happy to help!

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		Christian faiths as symbols of devotion and worship. Score: 3	[redacted]		you have specific questions about the books or need more details on other items in the photo, feel free to ask! Score: 0	[redacted]	Score: 1	Score: 0

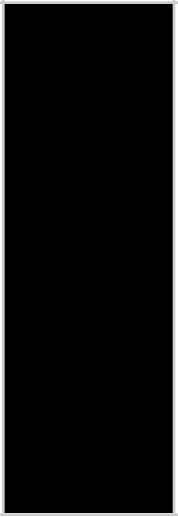
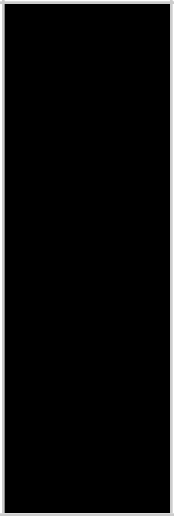
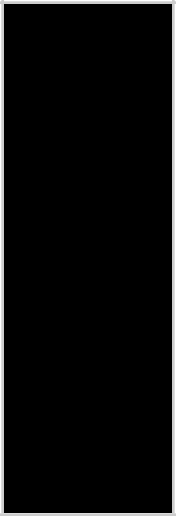
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
30	Gaussian Noise		There are four horse heads that are at least partially visible. Score: 1	There are two horse heads that are at least partially visible in the image. One is close to the fence and slightly out of focus, while the other is turned away from the camera and more in focus. Score: 0		There are five horse heads at least partially visible. Score: 0	There are three horse heads at least partially visible in the image. Score: 3		There are three horse heads at least partially visible in the image. Score: 3	There are three horse heads that are at least partially visible in the image. Score: 3

31	Speckle Noise		The company that made the laptop with the apple logo is Apple. Score: 3	The image you've provided appears to show a computer monitor with several devices and cables attached to it, but there is no clear branding visible. The logo on the device resembles that of Apple, which suggests that the device may be an Apple product or a similar design inspired by Apple products.		The laptop in the image is made by Apple, as indicated by the logo. Score: 3	The image you've provided appears to show the rear of a television or monitor with various input ports and cables. However, there is no clear branding or logo visible in the image that would allow me to identify the specific manufacturer. To determine which company made this, I		The logo on the product in the image is from Apple. Score: 3	It appears you've shared an image of a device, but the specific details are not clear due to the quality and angle of the photo. Without more information or a clearer view of the device, it's difficult for me to provide any accurate identification of the company that made this product. If there is a logo or

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		However, without a clear view of any company logos or brand identifiers, I cannot definitively identify the manufacturer. If you're looking for information about specific computer monitors, additional	[redacted]		would need more information such as a model number, serial number, or any visible logos on the device itself. If you can provide that information or if it's possible for you to take a clearer picture showing any brand markings,	[redacted]		other distinguishing features visible, you may want to zoom in on those areas when taking a screenshot or using another method to enhance the image quality. Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		details like model number or serial code would be helpful in providing accurate information. Score: 3	[redacted]		I might be able to assist further. Score: 0	[redacted]		

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
32	Speckle Noise		The image shows a knife and a spoon placed beside the plate with food. Score: 3	The image shows a plate of food, which includes what appears to be scrambled eggs, toast, and possibly some berries. There is also a glass of juice on the table. The utensils visible in the image include a		The image shows a knife and a spoon placed next to a plate of food. Score: 3	The image shows a knife and a spoon placed next to the plate of food. Score: 3		The utensils shown are a knife and a spoon. Score: 3	The image shows a plate of food with what appears to be scrambled eggs, possibly hash browns or some other fried potatoes, and a serving of mixed vegetables. There is also an orange juice glass next to the

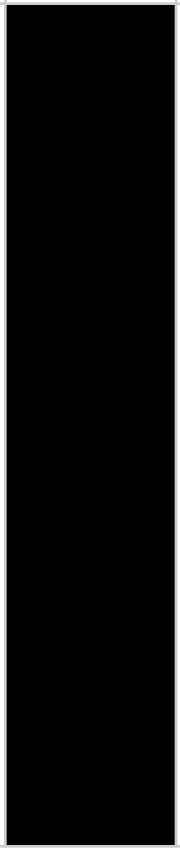
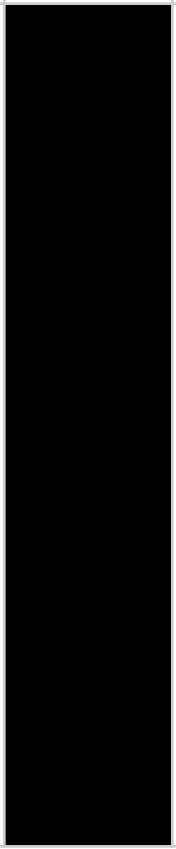
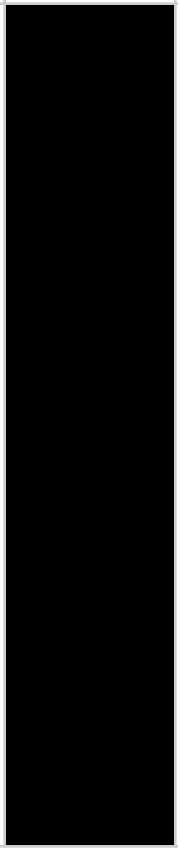
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				fork and a knife placed to the right side of the plate, for eating the meal. Score: 1						plate. The utensils visible include a knife and two forks, one on top of the other near the plate. Score: 1

https://doi.org/10.33774/coe-2025-6z9zx

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
33	Gaussian Noise		Yes, there are traffic lights visible in the background. Score: 3	Yes, in the background of the image, there are traffic lights. They appear to be yellow at the moment, indicating caution for drivers. The focus of the image is on two people engaged in cross-country skiing,		Yes, there are traffic lights in the background. Score: 3	Yes, in the background of the image, there are traffic lights indicating a roadway intersection. The focus of the image is on the two individuals skiing in the snow, but you can also see some residential buildings and street		I don't see any traffic lights in the image. Score: 0	No, there are no traffic lights in the image. The photo shows two people cross-country skiing on a snowy landscape. They appear to be having a conversation as they stand on their skis with ski poles at their

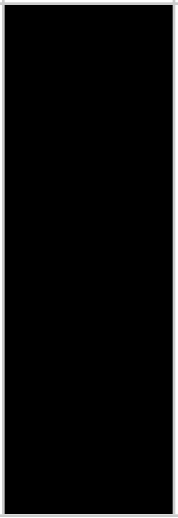
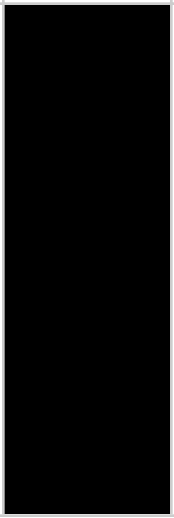
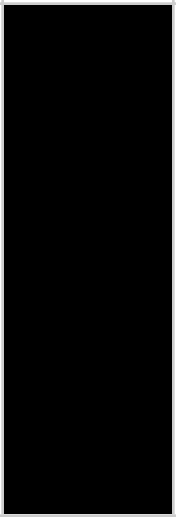
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		but the presence of traffic lights suggests that this activity may be taking place near a road or intersection where vehicular traffic operates under controlled signals. Score: 3	[redacted]		infrastructure beyond them. Score: 3	[redacted]		sides. The background features a residential area with houses and trees covered in snow. Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
34	Gaussian Noise		The image shows an elephant. Score: 3	The image shows an elephant. Score: 3		That's an elephant in the image. Score: 3	There is an elephant in the image. Score: 3		The animal in the image is an elephant. Score: 3	The image shows an elephant. It appears to be standing in a natural environment with rocks and vegetation, possibly within or near a habitat suitable for elephants such as a wildlife reserve or sanctuary. The

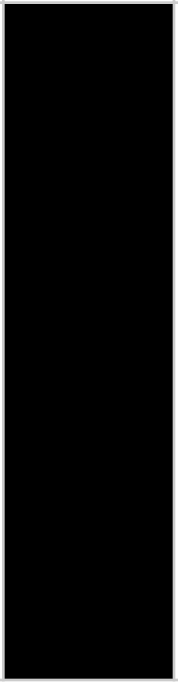
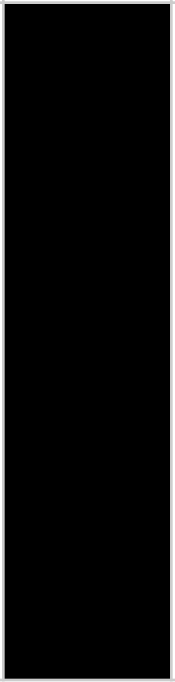
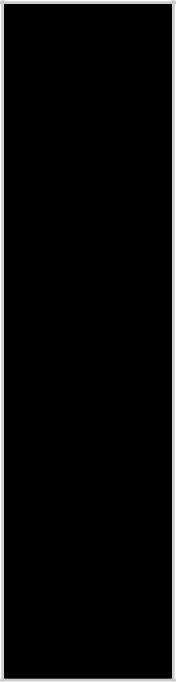
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
										building in the background suggests that this might be close to human facilities, which is common when wild animals are housed and cared for in protected environments. Score: 3

https://doi.org/10.33774/coe-2025-6z9zx

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
35	Salt-Pepper Noise		The bowl appears to be made of glass. Score: 3	The bowl in the image appears to be made out of glass. It has a clear, transparent quality typical of glassware. Score: 3		The bowl is made of glass. Score: 3	The bowl in the image appears to be made of glass. Glass bowls are often used for serving fruit or other food items, and they come in various shapes and sizes to accommodate different needs. In this particular		The bowl appears to be made of glass. Score: 3	The bowl in the image appears to be made of glass, as indicated by its transparency and the visible light refraction. Score: 3

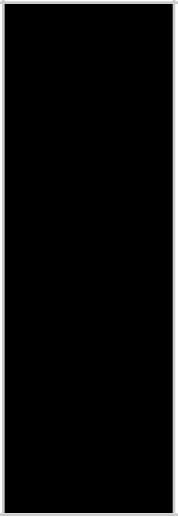
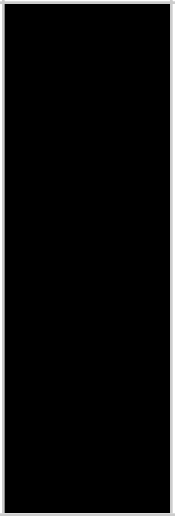
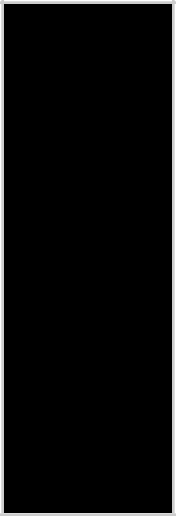
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
							case, the bowl seems to have a wide opening at the top that tapers slightly towards the bottom. Score: 3			

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
36	Speckle Noise		The pattern displayed most prominently on the items is the Union Jack, which is the national flag of the United Kingdom. Score: 3	The pattern displayed most prominently on the items in this image is the Union Jack flag, which features a combination of red, white, and blue colors with crosses and saltires representing the different countries that make up the		The items in the image prominently display the Union Jack pattern, which is the national flag of the United Kingdom. Score: 3	The most prominent pattern displayed on the items in this image is the Union Jack, which is the national flag of the United Kingdom. It features a combination of red, white, and blue colors with crosses and		The pattern most prominently displayed on the items in the image is the design of the Union Jack, the national flag of the United Kingdom. Score: 3	The image shows bedding with a prominent Union Jack design, which is the national flag of the United Kingdom. The flag features crosses and saltires representing each country within the UK:

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				United Kingdom. The pillows and sheets have this design printed on them. Score: 3			saltires representing the different countries of the UK: England, Scotland, Ireland (Northern Ireland), and Wales. Score: 3			England, Scotland, Ireland (Northern Ireland), and Wales. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
37	Speckle Noise		This is actually a sign for "Cabbage Town," not Chinatown. Cabbage Town is an area known for its distinct character and is different from a Chinatown district, which typically features Chinese cultural	The image shows a sign with "CABBAGE TOWN" written on it, which suggests that the location is Cabbagetown. This area has historical connections to Irish immigrants who often planted cabbage		This image shows a sign saying "Cabbage Town." It doesn't appear to be related to a Chinatown area. Score: 3	This image shows a street sign that says "I Love Cabbage Town." Cabbage Town is a neighborhood in Atlanta, Georgia, United States, not China Town, which typically refers to an ethnic Chinese community or		This is a place referred to as "Cabbage Town," not China Town. The sign makes that clear. Score: 3	The image shows a sign that reads "I Love Cabbage Town," which suggests that the location is not in Chinatown. The term "Cabbagetown" often refers to specific neighborhoods in cities around the world,

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
			elements and businesses. Score: 3	gardens and grew their own food in order to make ends meet. It's not China Town; that refers to neighborhoods with Chinese cultural influences, typically found in many cities around the world.			district outside of Asia. The presence of the "Cabbage Town" sign suggests that this photo was taken in that specific area. Score: 3			including Toronto, Atlanta, and London, which are named for their historical connection to the Irish community, who were known for growing cabbage gardens. It seems like this sign is

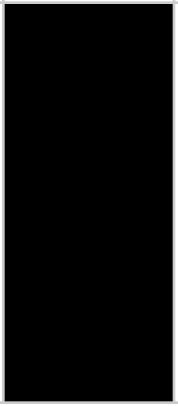
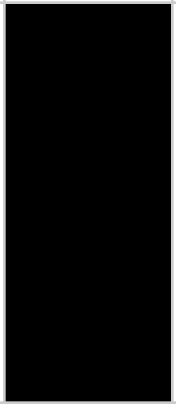
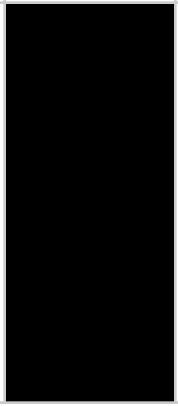
ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
				Score: 3						indicating a different cultural or historical neighborhood rather than a Chinatown. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
38	Gaussian Noise		The man is standing on a bed. Score: 3	The man is standing in a room that appears to be a dormitory or bedroom. There are two beds, one of which he seems to be standing on, and personal items scattered around, such as clothes, cups, and other small		The man is standing on a desk that is placed on top of a bed. Score: 0	The man is standing on top of a bed. Score: 3		The man is standing on a desk in a room. Score: 0	The man is standing in a room that looks like a bedroom. He appears to be near some furniture, possibly against a wall or close to a bed. The image quality is not very clear, but we can see a person with his hand up to his mouth as if he

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		objects. The setting looks lived-in and casual, suggesting a private space like a college dorm or a shared living arrangement. Score: 3	[redacted]			[redacted]		might be speaking or reacting to something. Score: 0

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
39	Speckle Noise		There are two people in the image. They appear to be playing a video game, each holding a controller, with the person on the left showing an intense expression. Score: 3	There are two people in the image. They seem to be sitting on a couch playing video games together. One person is holding what appears to be a Wii remote control, indicating that they might be playing on a		There are two people in the image, and they appear to be playing a video game with game controllers. Score: 3	There are two people in the image. They seem to be playing a video game together, as indicated by their focused expressions and the Wii remote controls in their hands. It looks like they might be engaged in an		There are two people in the image, and they appear to be playing a video game with motion controllers. Score: 3	There are two people in the image. They seem to be sitting on a couch playing video games together. One person is holding a Nintendo Wii remote controller, which suggests that they might be engaged in

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
		[redacted]		Nintendo Wii console, which uses motion-sensitive controllers for gameplay. Both individuals appear focused on their activity, suggesting that they are engaged in an interactive game. Score: 3	[redacted]		interactive gaming session, possibly on a console like the Nintendo Wii given the design of the controllers. Score: 3	[redacted]		motion-controlled gameplay typical of the Wii console. The other person appears to be watching or waiting for their turn. Score: 3

ID	Noise Type	GPT-3.5 Baseline Response *(only if textual prompt)*	GPT-4o Baseline Response	LLaVA Baseline Response	GPT-3.5 Noise Response *(only if textual prompt)*	GPT-4o Noise Response	LLaVA Noise Response	GPT-3.5 Denoised Response *(only if textual prompt)*	GPT-4o Denoised Response	LLaVA Denoised Response
40	Salt-Pepper Noise		The bird's head is white. Score: 3	The bird in the image has a white head. Score: 3		The bird's head is white. Score: 3	The bird in the image has a white head with some black markings. Score: 3		The bird's head is white. Score: 3	The bird in the image has a white head. Score: 3

Appendix E: Statistics

Average Scores at Different Noise Levels

Model	*GPT-3.5*	*GPT-4o*	*LLaVA*	*Combined MLLM*	*(Number of Prompts)*
Average Textual Clean Score	2.400	3.000	2.500	2.750	10
Average Textual Noisy Score	2.400	2.700	1.500	2.100	10
Average Textual Denoised Score	1.900	1.900	0.500	1.200	10
Average Image Clean Score		2.833	2.300	2.567	30
Average Image Noisy Score		2.533	2.267	2.400	30
Average Image Denoised Score		2.433	2.067	2.250	30

Mean Differences in Scores at Different Noise Types/Levels

Model	*GPT-3.5*	*GPT-4o*	*LLaVA*	*Combined MLLM*	*(Number of Prompts)*
Mean Difference Textual Noise – Clean	0.000 (p=N/A)	-0.300 (p=0.172)	-1.000 (p=0.048)	-0.650 (p=0.025)	10
Mean Difference Gaussian Noise – Clean		-0.444 (p=0.112)	-0.667 (p=0.173)	-0.556 (p=0.072)	9
Mean Difference Speckle Noise – Clean		-0.154 (p=0.169)	0.154 (p=N/A)	0.000 (p=N/A)	13
Mean Difference Salt-Pepper Noise – Clean		-0.375 (p=0.175)	0.375 (p=N/A)	0.000 (p=N/A)	8

Model	*GPT-3.5*	*GPT-4o*	*LLaVA*	*Combined MLLM*	*(Number of Prompts)*
Mean Difference Image Noise – Clean		-0.300 (p=0.030)	-0.033 (p=0.453)	-0.167 (p=0.148)	30
Mean Difference Denoised – Textual Noise	-0.500 (p=0.138)	-0.800 (p=0.087)	-1.000 (p=0.074)	-0.900 (p=0.010)	10
Mean Difference Denoised – Gaussian Noise		-0.222 (p=0.695)	0.111 (p=0.880)	-0.056 (p=0.450)	9
Mean Difference Denoised – Speckle Noise		-0.077 (p=0.794)	-0.538 (p=0.089)	-0.308 (p=0.074)	13
Mean Difference Denoised – Salt-Pepper Noise		0.000 (p=N/A)	0.000 (p=N/A)	0.000 (p=N/A)	8
Mean Difference Denoised – Image Noise		-0.100 (p=0.620)	-0.200 (p=0.489)	-0.150 (p=0.389)	30

For Noisy vs Clean: H_0 => mean difference = 0, H_a => mean difference < 0

For Denoised vs Noisy: H_0 => mean difference = 0, H_a => mean difference ≠ 0

Note. Findings significant at the α=.05 level are highlighted in red; at the α=.10 level, in orange; and at the α=.20 level, in yellow. Gray indicates that a p-value is not applicable (either the results perfectly matched the null hypothesis or they were on the opposite side of zero compared to the alternative hypothesis).