Know What You do Not Know: Verbalized Uncertainty Estimation Robustness on Corrupted Images in Vision-Language Models

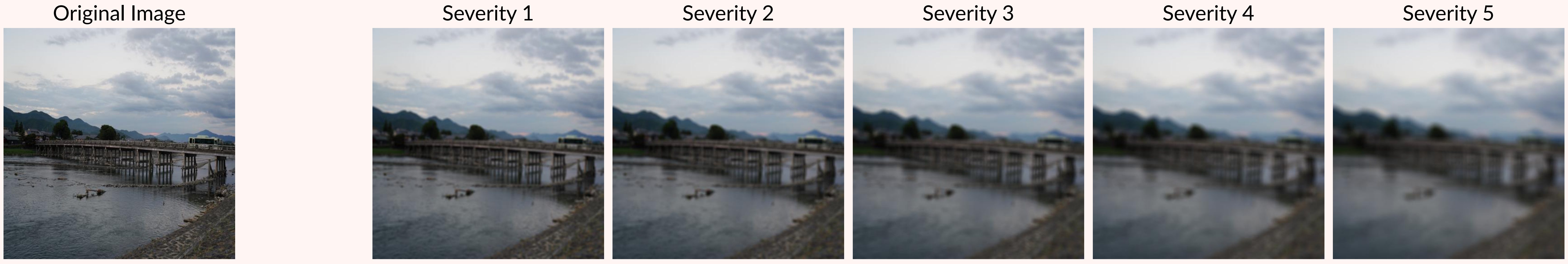
Mirko Borszukovszki Ivo Pascal de Jong Matias Valdenegro-Toro

Department of Artificial Intelligence, Bernoulli Institute – University of Groningen

VLMs get less accurate for blurry images. But can they still estimate their uncertainty?

Desired behaviour: When an image is of lower quality, a Visual-Language Model (VLM) may find it more difficult to correct answer questions about that image. When the accuracy decreases, a well-calibrated model should also become less confident.

Experiment: We applied State-of-the-Art VLMs to three Visual-Question Answer (VQA) datasets. We added corruptions to the image to see whether VLMs become overconfident.



Togetsukyo Bridge in Arashiyama.
Confidence: 90%

Kyoto, Japan
Confidence: 80%

Kyoto, Japan
Confidence: 75%

Kyoto, Japan.
Confidence: 70%

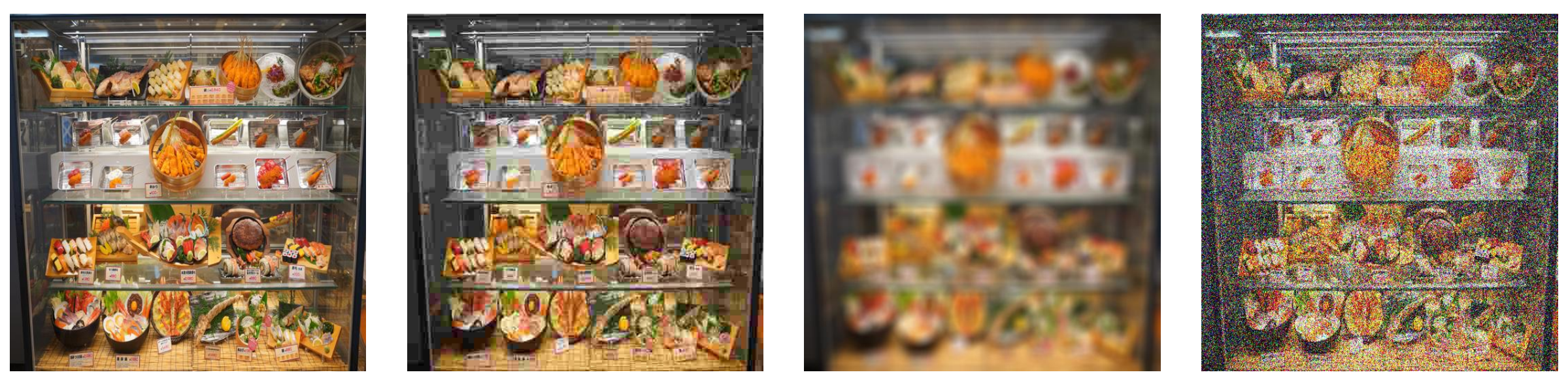
Lake Como in Italy.
Confidence: 75%

Lake
Confidence: 60%

Figure 1. Example with Defocus Blur corruption. **Question**: Where was this photo taken? **Correct Answer**: Japan, Kyoto, Arashiyama Area, the Bridge is named Togetsu-kyo Bridge.

What kind of corruptions?

We used three different types of corruption, with five different severity levels [4]. We selected JPEG compression, defocus burring and Gaussian Noise as these are all corruptions that VLMs are likely to encounter in practice.

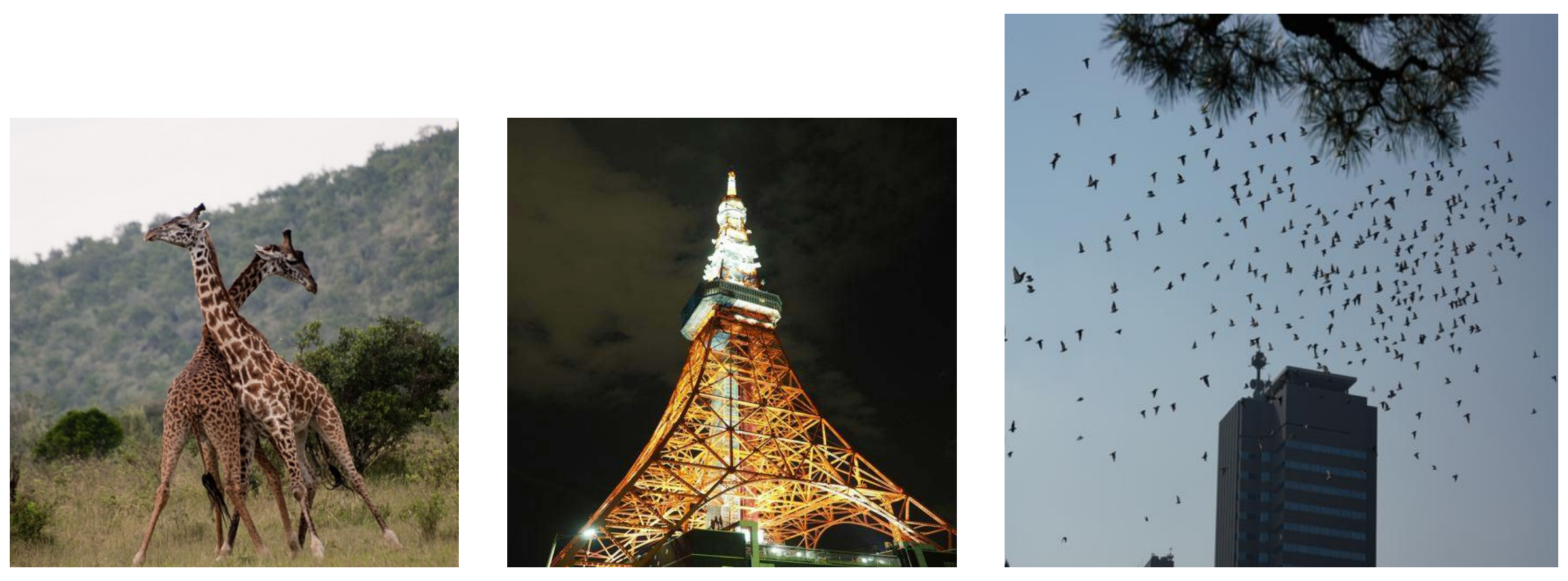


(a) Original image (b) JPEG compression (c) Defocus blur (d) Gaussian noise

Figure 2. Used corruptions on severity 5. **Question:** What kind of food is showcased in this photo? **Answer:** Japanese food *or* food model, called Shokuhin Sampuru in Japanese.

Which Visual-Question Answering Datasets?

- **Easy VQA** evaluated on the popular visual question answering dataset [1, 2], of which we randomly sampled 36 images.
- **Hard VQA** evaluated on the Japanese Uncertain Scenes (JUS) [3]. This dataset contains 29 "tricky" questions.
- **Counting task** was also evaluated on the JUS dataset as it contains 13 challenging counting exercises. Here, the models need to predict confidence intervals.



(a) **Q:** What type of place is this?
A: Savannah.

(b) **Q:** What is in this photo?
A: Tokyo Tower.

(c) **Q:** How many birds are shown in this photo?
A: 250-280.

Figure 3. Samples from the three tasks.
(a) **Easy VQA** (b) **Hard VQA**, (c) **Counting task**.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*, 2024.

[4] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Easy VQA results

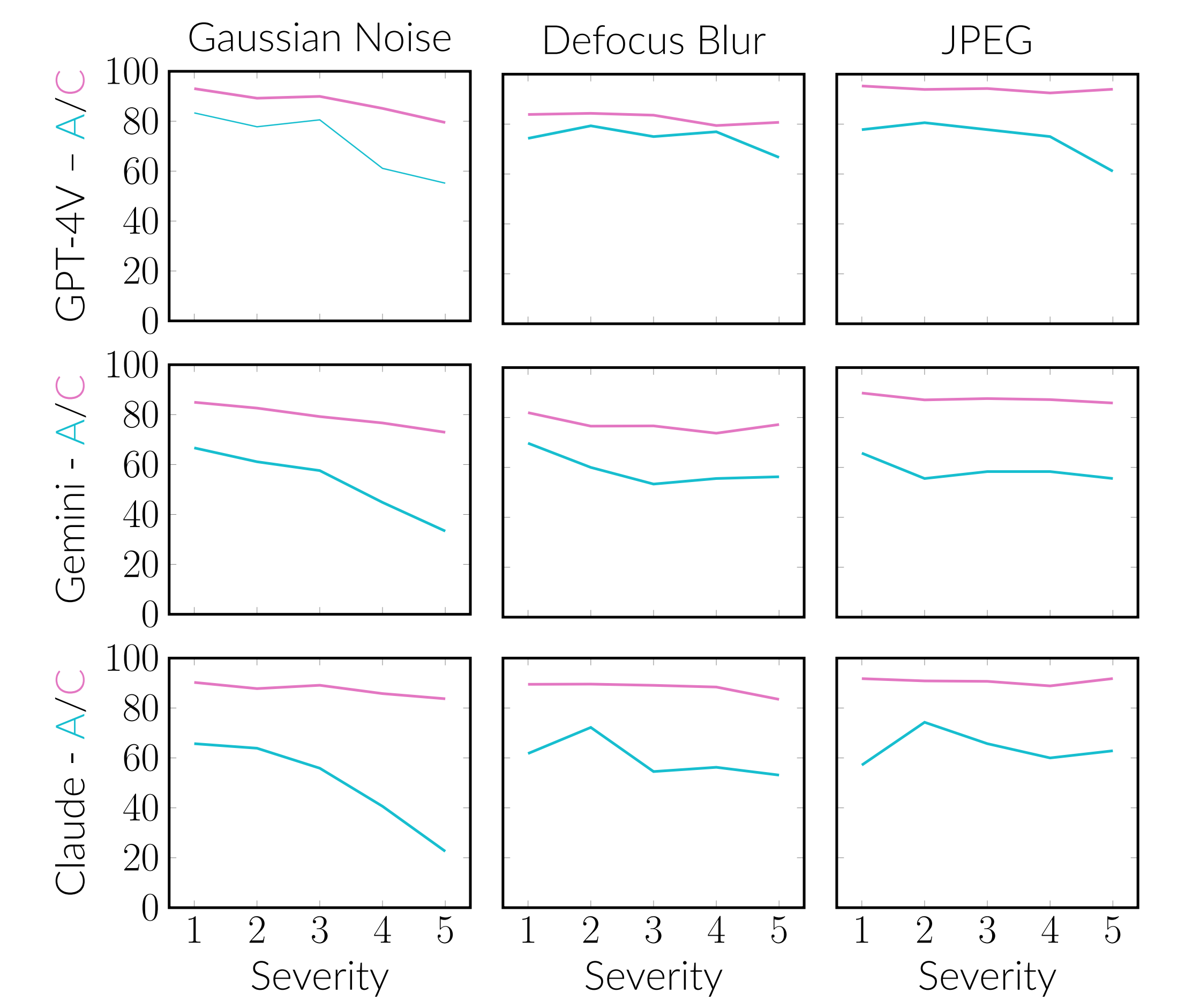


Figure 4. Accuracy and confidence in the **easy** VQA experiment. All models are overconfident.

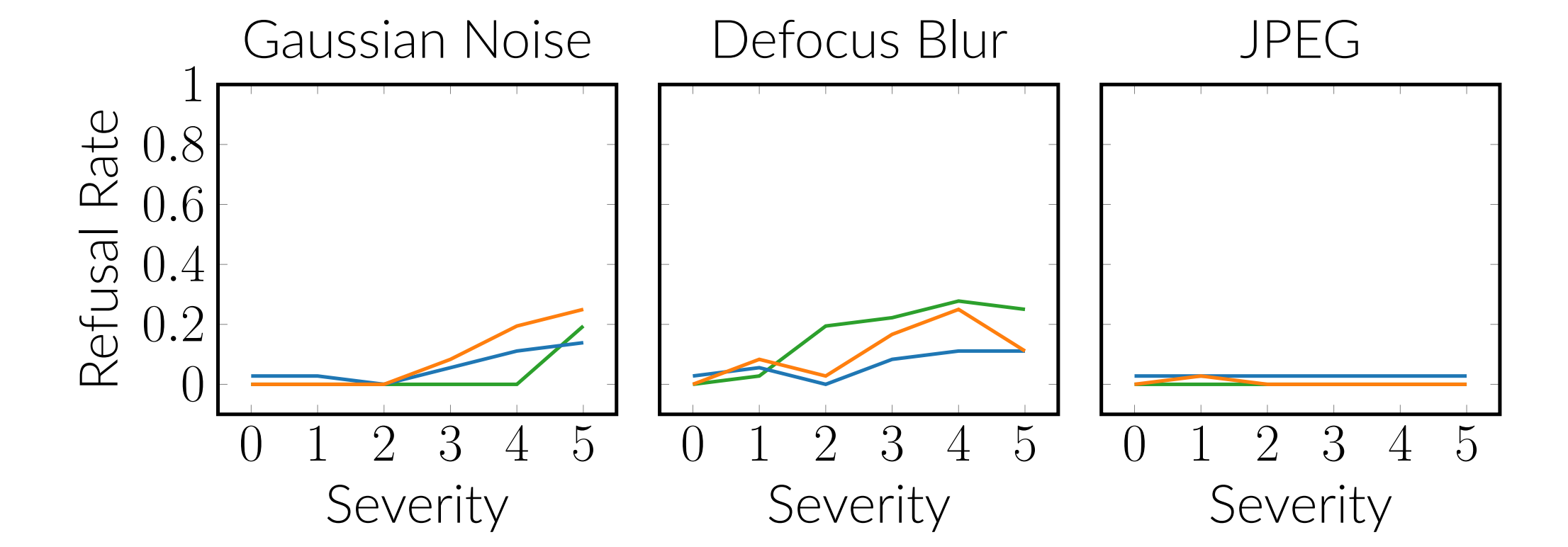


Figure 5. Refusal rates for Claude, Gemini, and GPT-V4 in easy VQA. Models refuse corrupted images.

Hard VQA

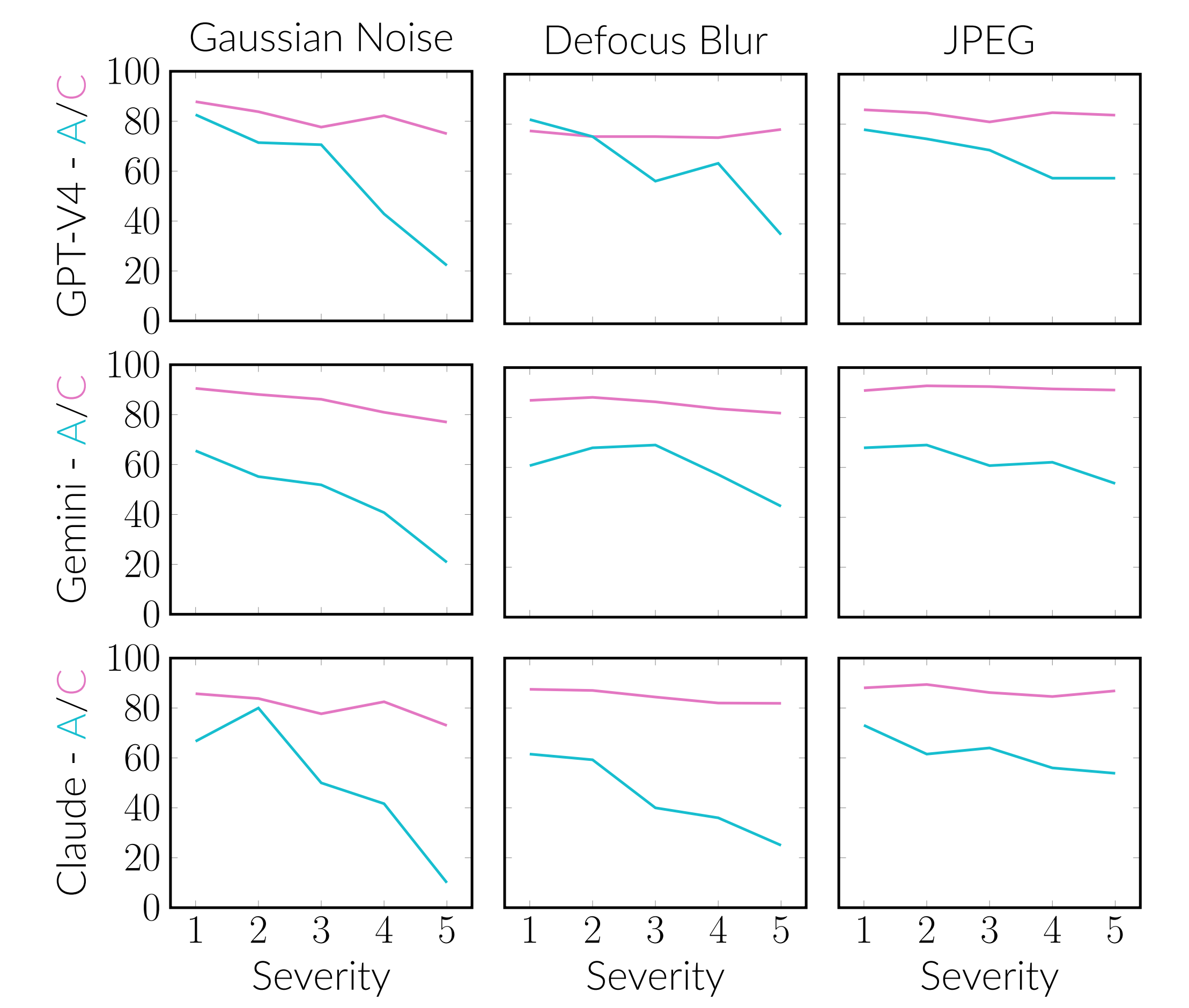


Figure 6. Accuracy and confidence for the **hard** VQA experiment. The corruptions increase overconfidence.

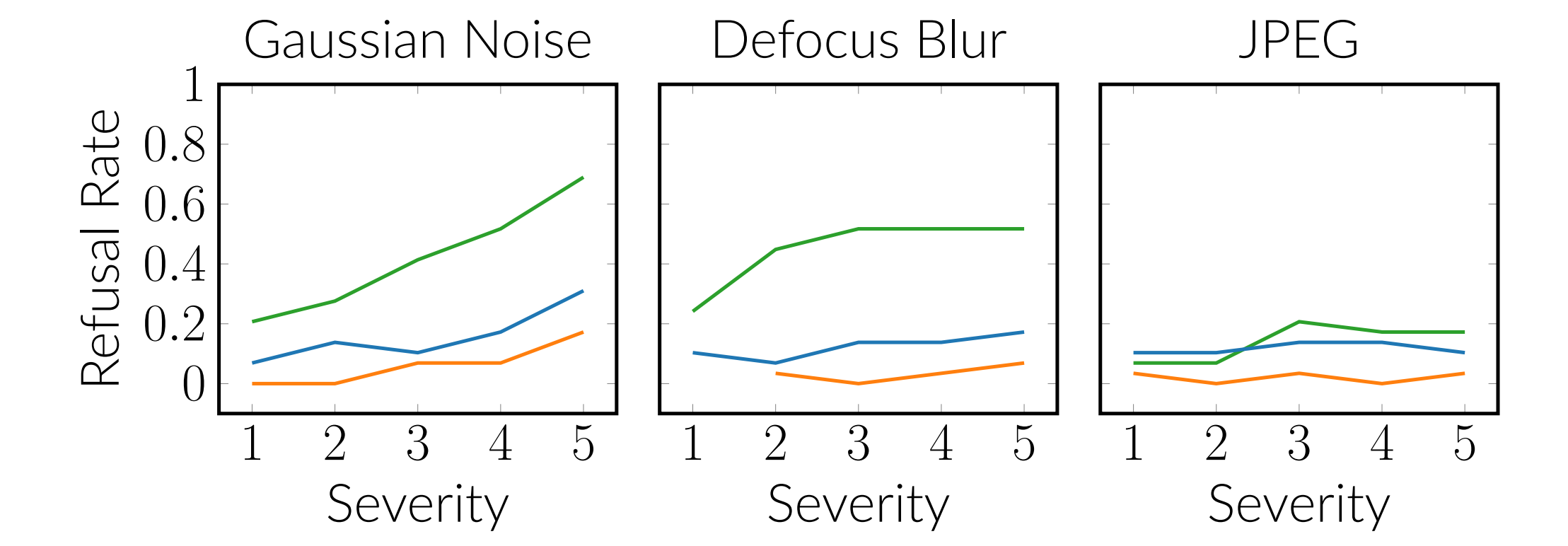


Figure 7. Refusal rates for Claude, Gemini, and GPT-V4 in hard VQA. GPT-4V often refuses.

Counting task

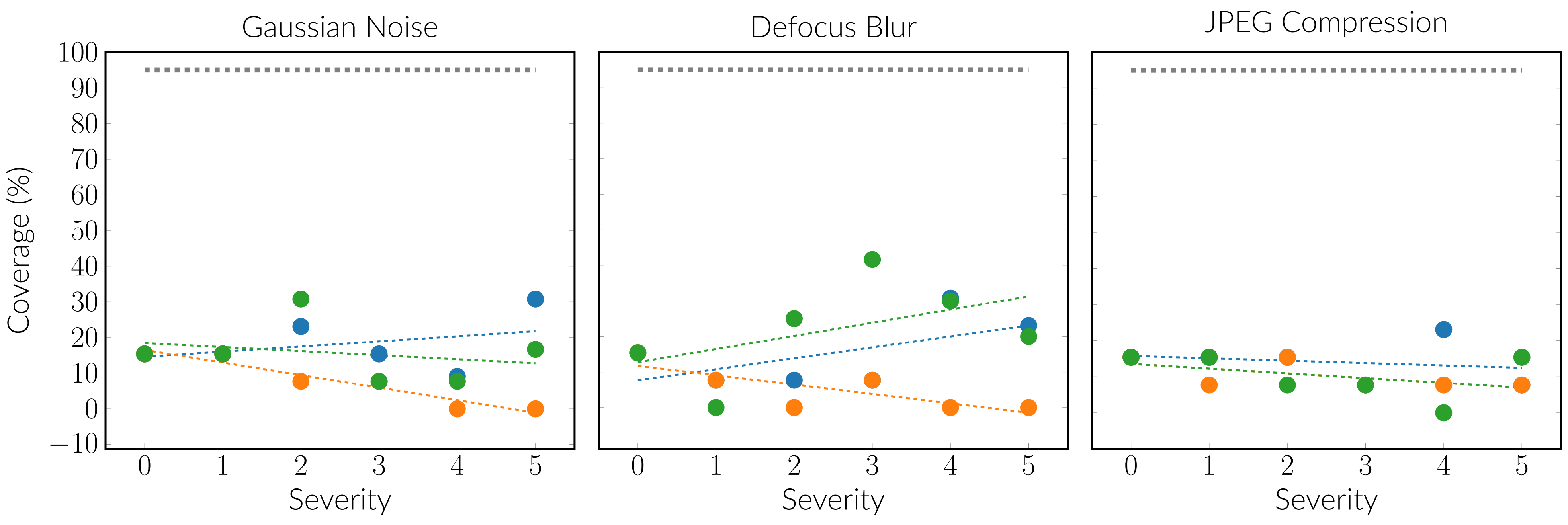


Figure 8. Coverage (confidence interval accuracy) scores for the **counting experiment** for Claude, Gemini, and GPT-V4. The dotted line at the top represents the 95% accuracy expected for a perfectly calibrated model. Models are severely overconfident in counting.

Conclusions

1. VLMs are overconfident.
2. Increased corruption severity increases the overconfidence.
3. GPT-4V outperformed the other two models in the visual question-answering experiments.
4. JPEG compression is better handled by all of the models than Gaussian noise and defocus blur.
5. Higher refusal rates can improve calibration.
6. VLMs are especially miscalibrated when they are asked to express their answer in a 95% confidence interval.

https://arxiv.org/abs/2504.03440 TrustNLP workshop @ NAACL 2025 ivo.de.jong@rug.nl, m.a.valdenegro.toro@rug.nl