UiT The Arctic University of Norway

NORA summer school on multi-modal learning

Responsible AI

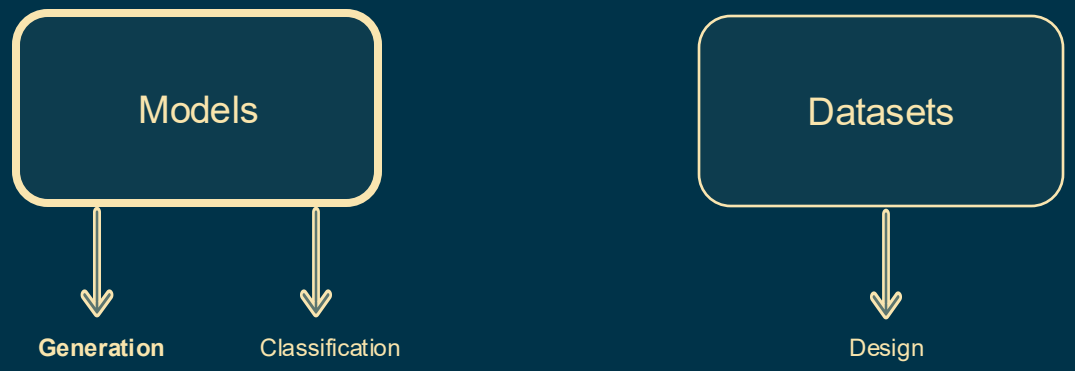
Rwiddhi Chakraborty
UiT Machine Learning Group and Visual Intelligence

Schedule Today

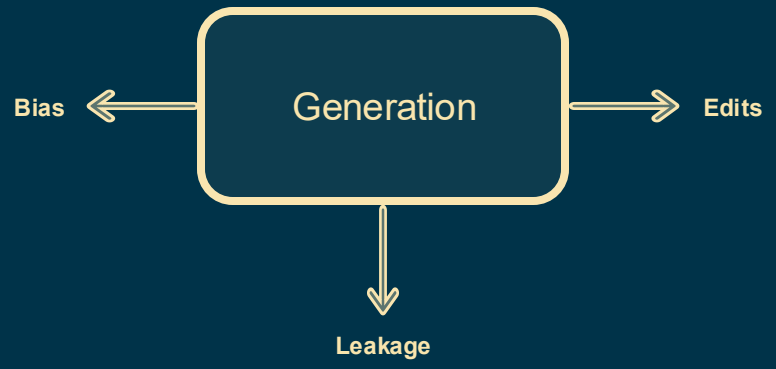
- 09 - 10: Special Topics - I
- 10 - 11: Special Topics - II
- 11 – 12: Group Project
- 12 – 13: Lunch
- 13 – 14: Group Project
- 14 – 15: Presentations, award, exam info, wrap-up!

In this talk

In this talk



Responsible Generation



Data Leakage and Memorization

Data privacy refers to protected training data

Prevents forgery, theft, and other violations

Models “leaking” training data is a safety risk

Data Leakage and Memorization

Leaking training data is a form of overfitting and memorization

Most modern generative models suffer from this issue

Unclear whether benchmark performances correlate with true understanding

Data Leakage and Memorization

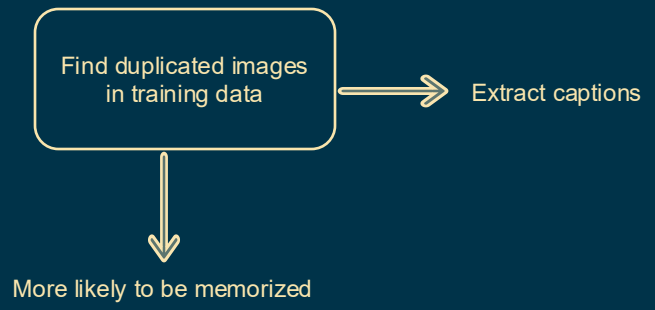
How do we analyze memorization in diffusion models?

Data Leakage and Memorization

How do we analyze memorization in diffusion models?

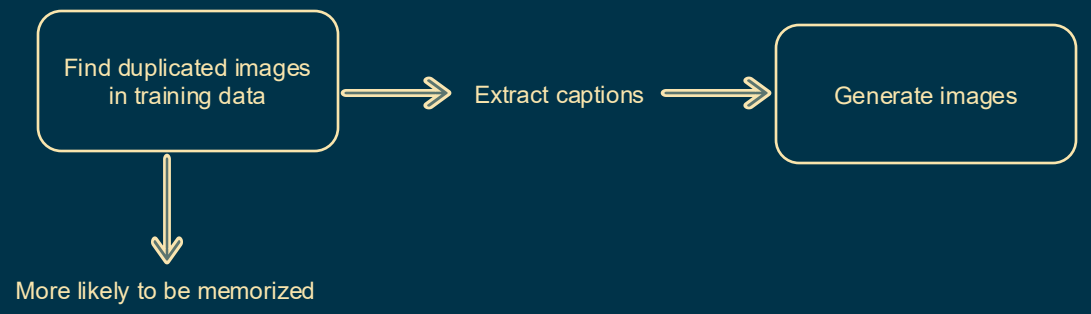
Data Leakage and Memorization

How do we analyze memorization in diffusion models?



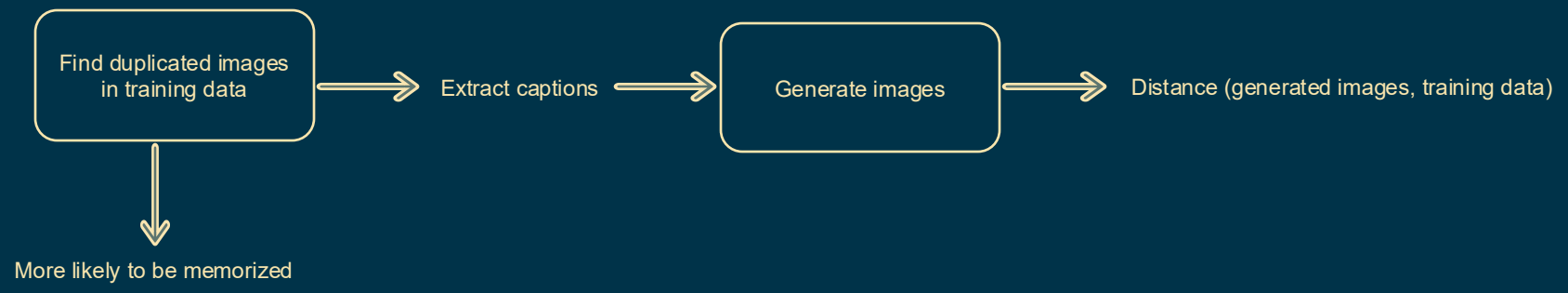
Data Leakage and Memorization

How do we analyze memorization in diffusion models?



Data Leakage and Memorization

How do we analyze memorization in diffusion models?



Data Leakage and Memorization

The result

Carlini, Nicolas, et al. "Extracting training data from diffusion models." 32nd USENIX Security Symposium (USENIX Security 23). 2023.

Data Leakage and Memorization

Vast majority of generated images are photographs of real people

Not all images are permissively licensed and raise copyright issues as well

Risks greater for diffusion models trained on more sensitive data (e.g medicine)

Carlini, Nicolas, et al. "Extracting training data from diffusion models." 32nd USENIX Security Symposium (USENIX Security 23). 2023.

Data Leakage and Memorization

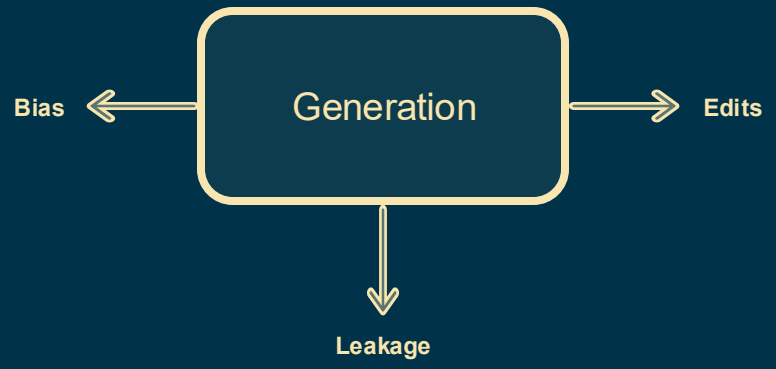
Model specific issue

GANs are safer as they are not trained to directly mimic the training data

Risks greater for diffusion models trained on more sensitive data (e.g medicine)

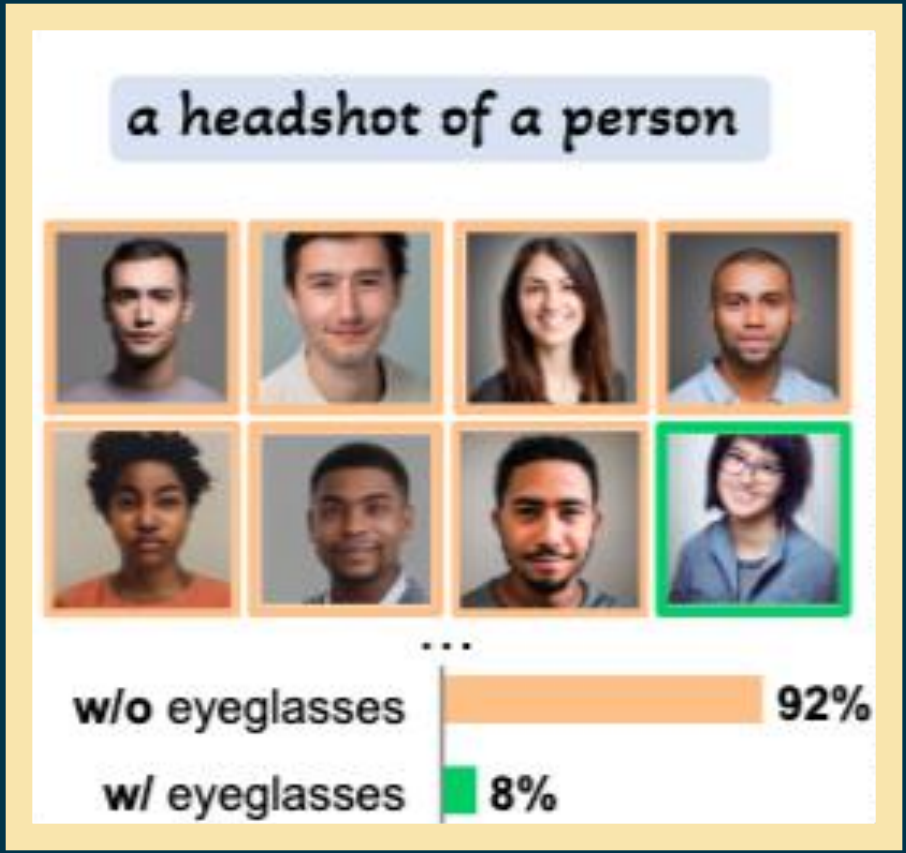
Carlini, Nicolas, et al. "Extracting training data from diffusion models." 32nd USENIX Security Symposium (USENIX Security 23). 2023.

Responsible Generation



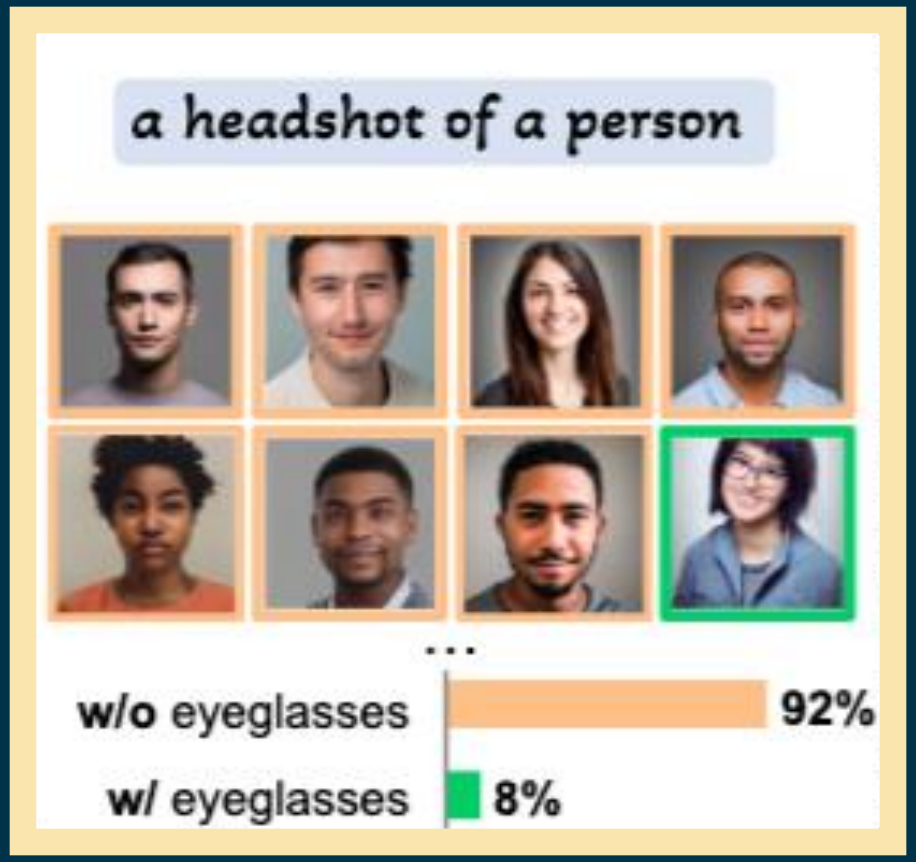
Biased Generation

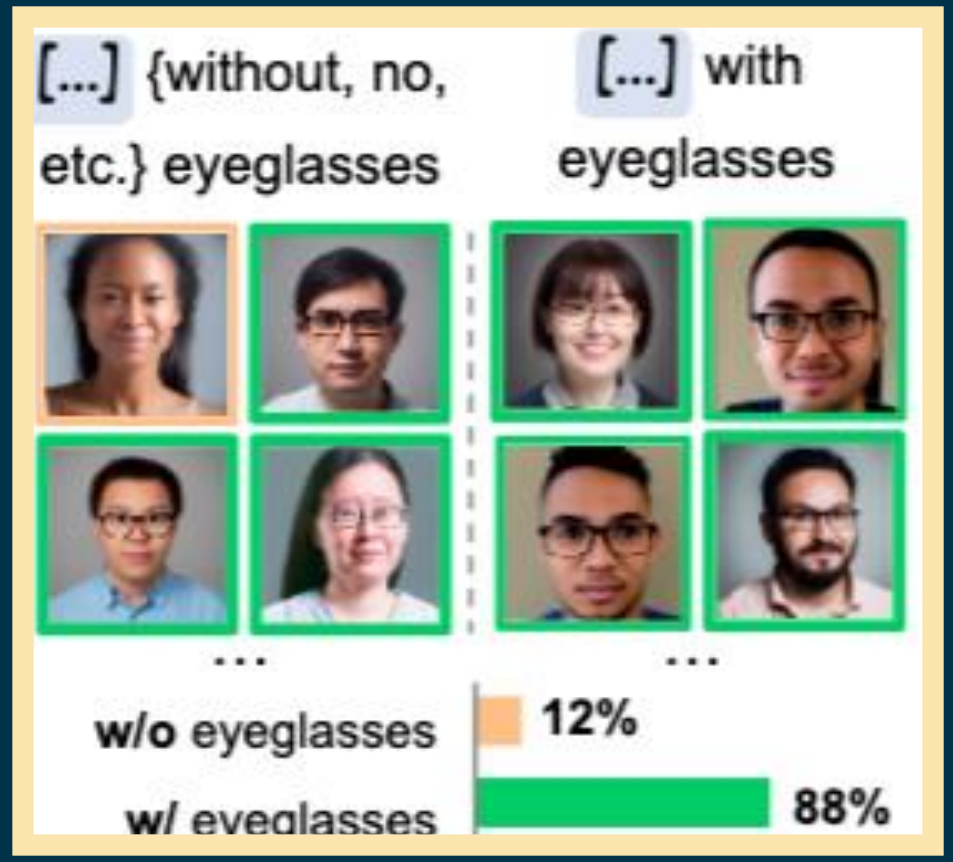
Any attribute can be a minority in the training set



Zhang, Cheng, et al. "Iti-gen: Inclusive text-to-image generation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Biased Generation



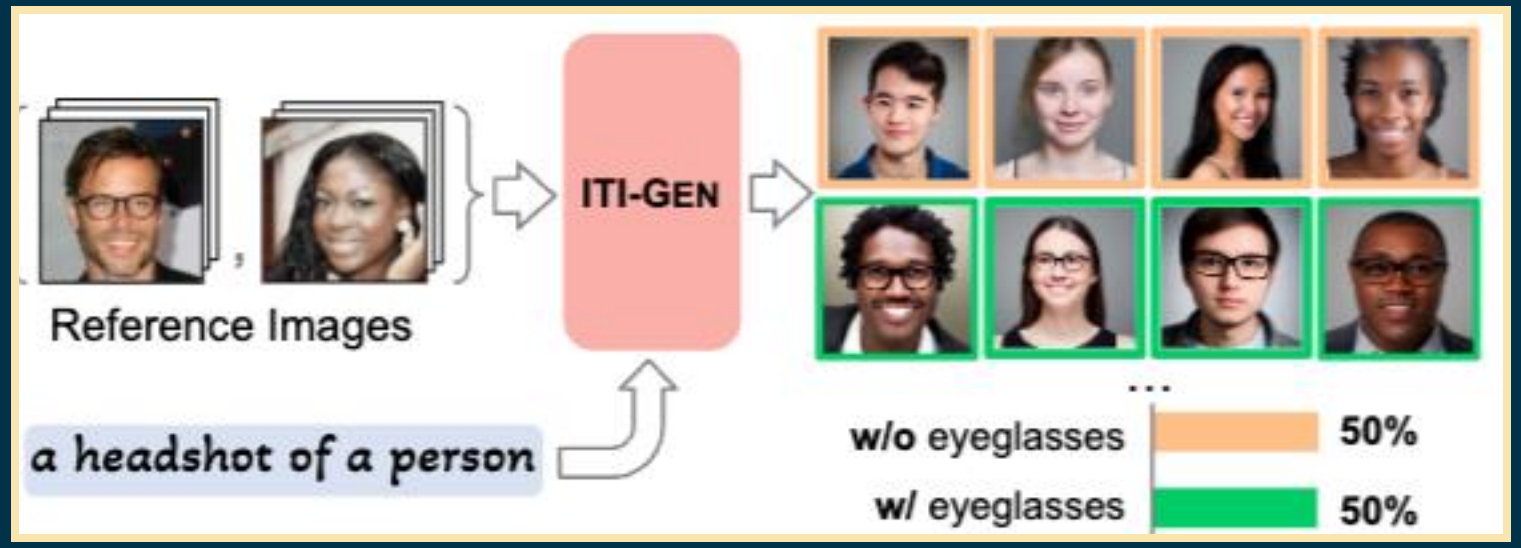


Hard prompting is ambiguous

Zhang, Cheng, et al. "Iti-gen: Inclusive text-to-image generation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

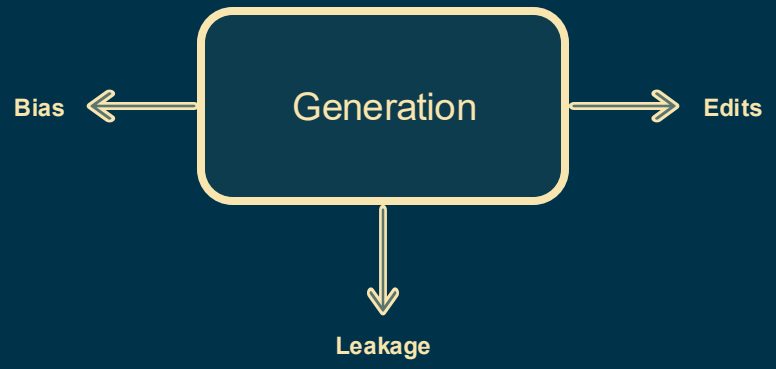
Responsible Generation

Image-guided prompt tuning



Zhang, Cheng, et al. "Iti-gen: Inclusive text-to-image generation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Responsible Generation



Editing harmful generation

Remove copyrighted/memorized content from T2I models

Prevent model from generating harmful concepts

Editing harmful generation

Kumari, Nupur, et al. "Ablating concepts in text-to-image diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Editing harmful generation

Kumari, Nupur, et al. "Ablating concepts in text-to-image diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Editing harmful generation

A photo of a cat

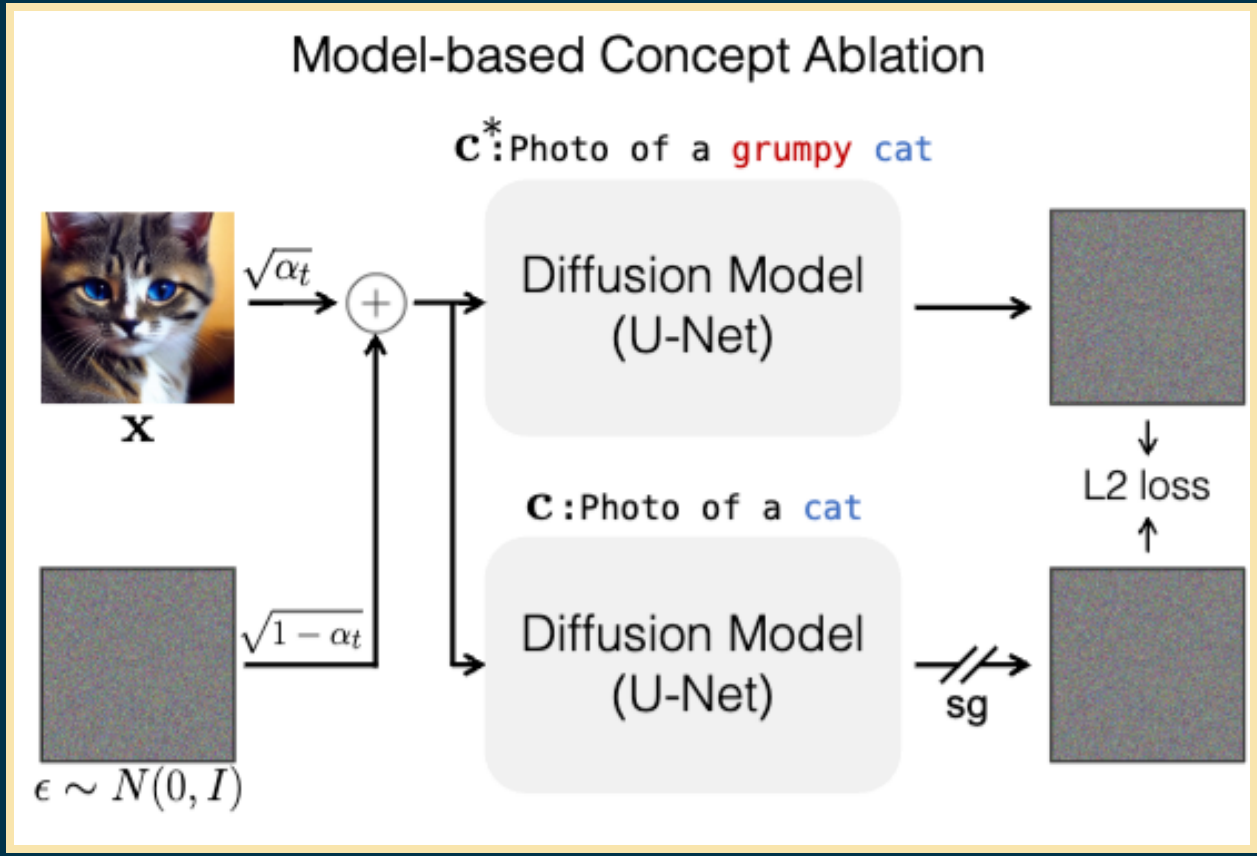
Anchor concept **c**

A photo of a *grumpy* cat

Target concept **c***

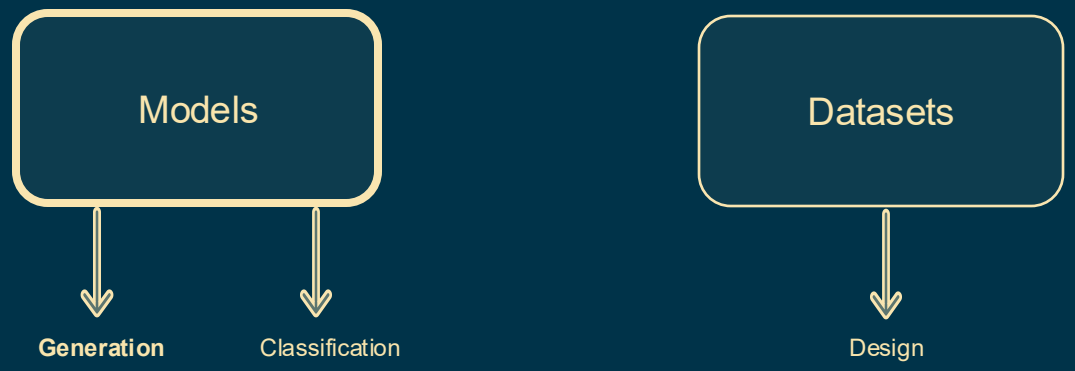
Kumari, Nupur, et al. "Ablating concepts in text-to-image diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Editing harmful generation



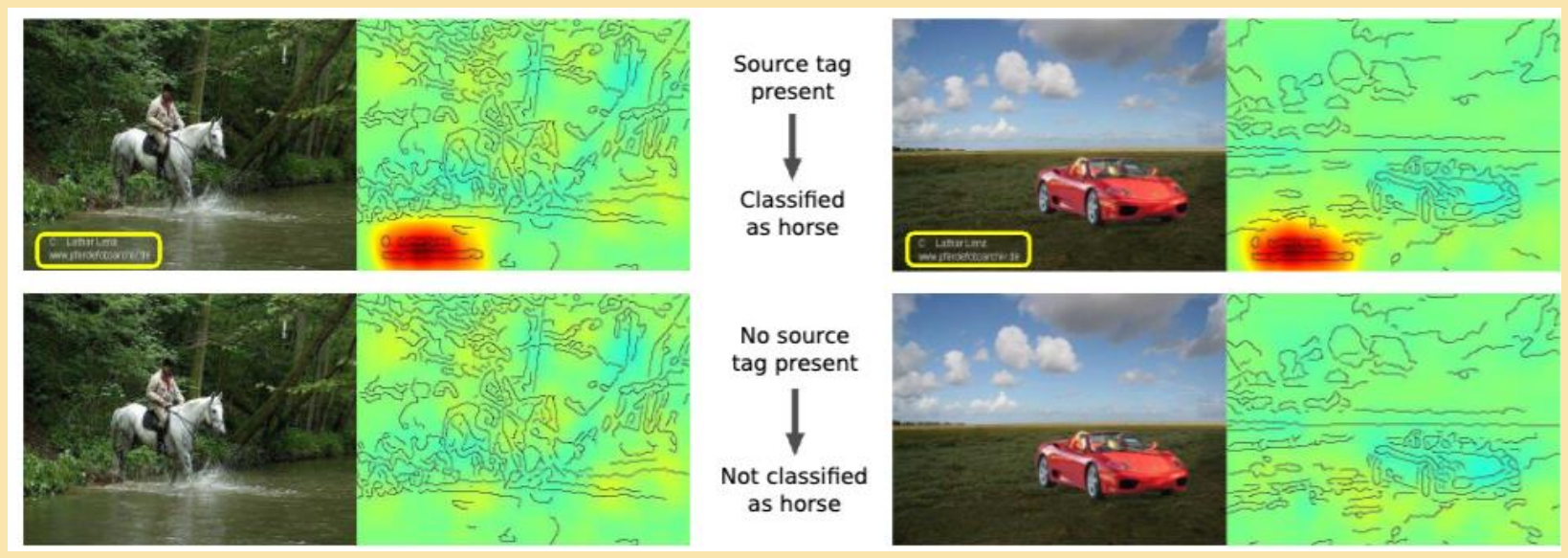
Kumari, Nupur, et al. "Ablating concepts in text-to-image diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

In this talk



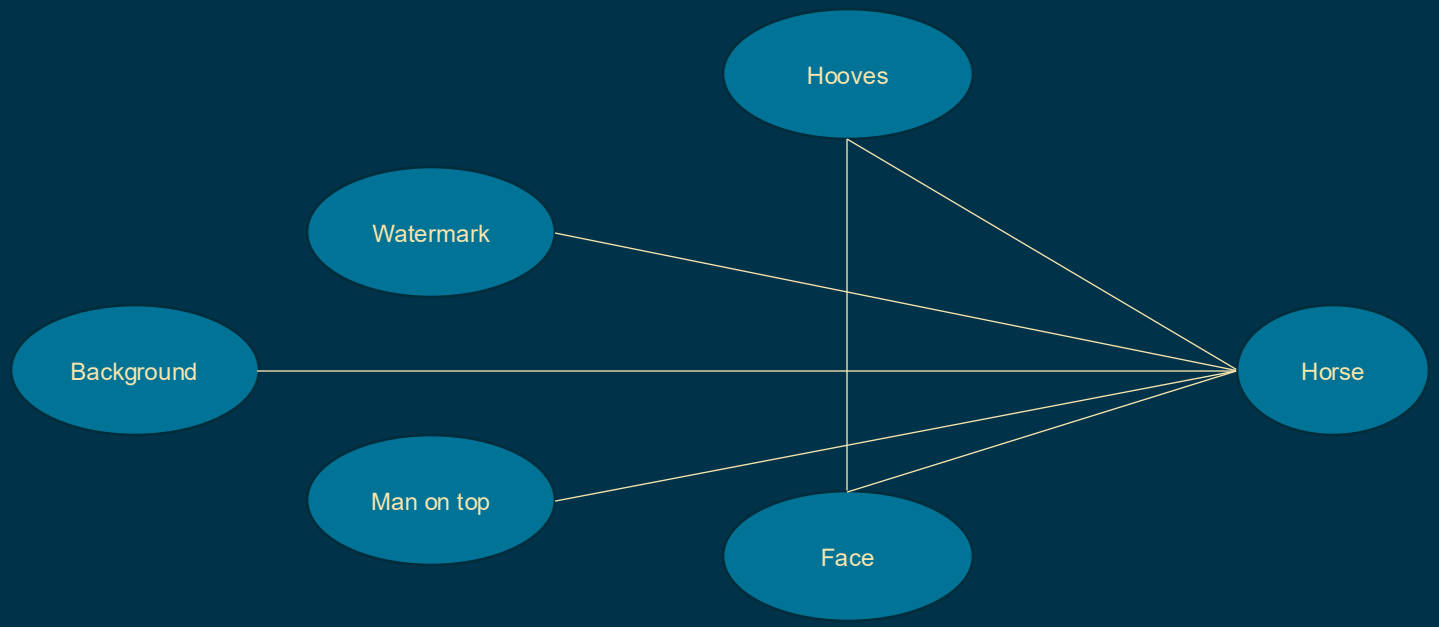
In this talk

Shortcut Learning



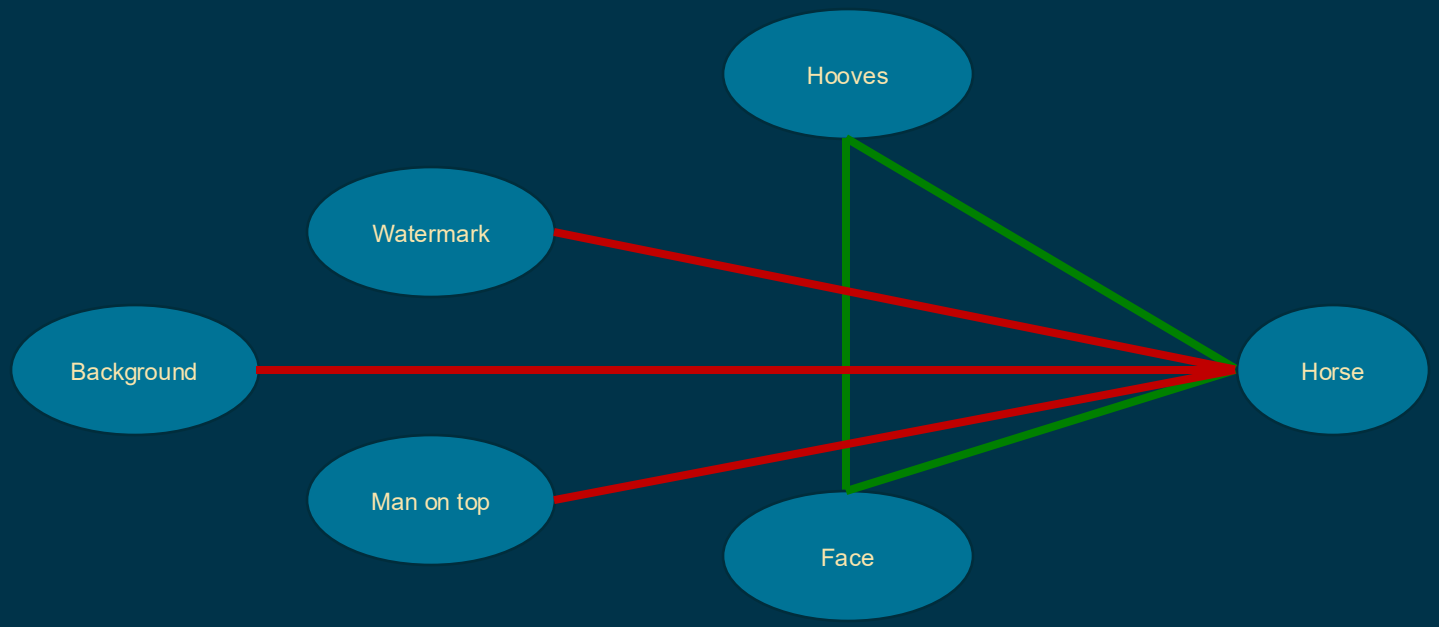
Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1096.

Shortcut Learning



Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1096.

Shortcut Learning



Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1096.

Shortcut Learning in VLMs

There are two modalities in inference – text and vision

Do models leverage both modalities?

Shortcut Learning in VLMs

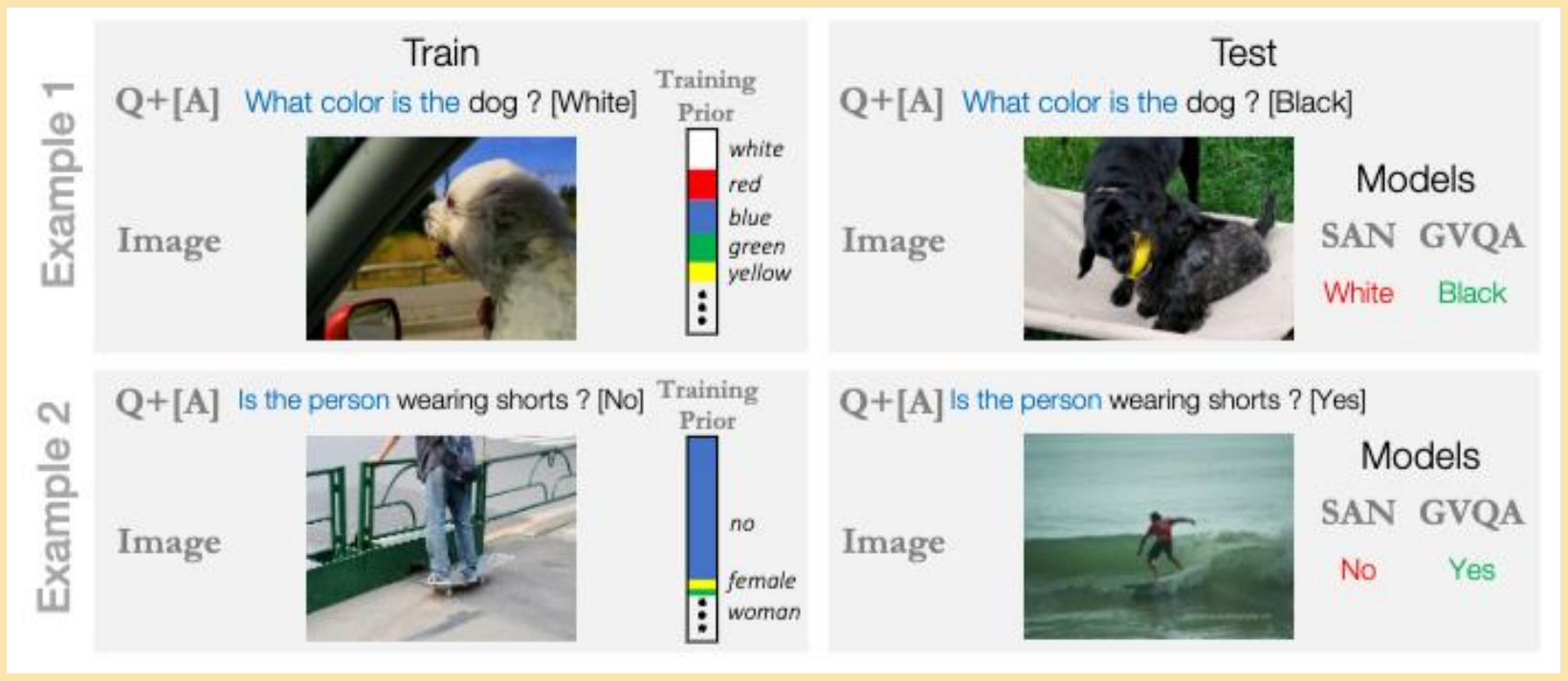
There are two modalities in inference – text and vision

Do models leverage both modalities?

(Un)Surprisingly, no!

Shortcut Learning in VLMs

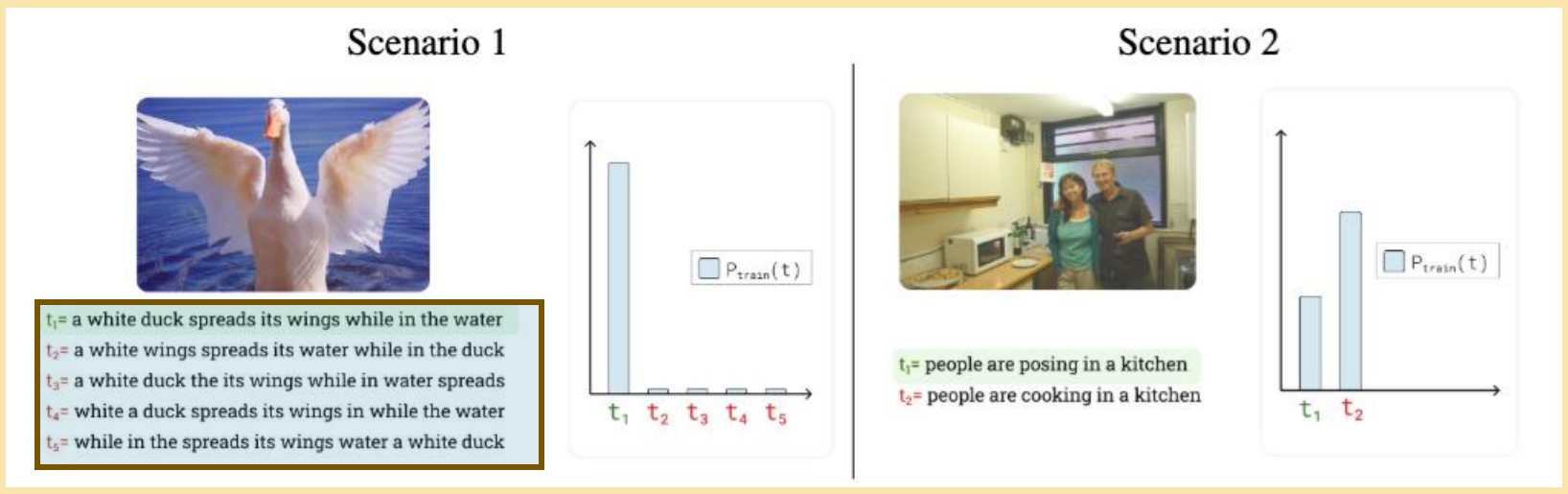
The language prior



Agrawal, Aishwarya, et al. "Don't just assume; look and answer: Overcoming priors for visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Shortcut Learning in VLMs

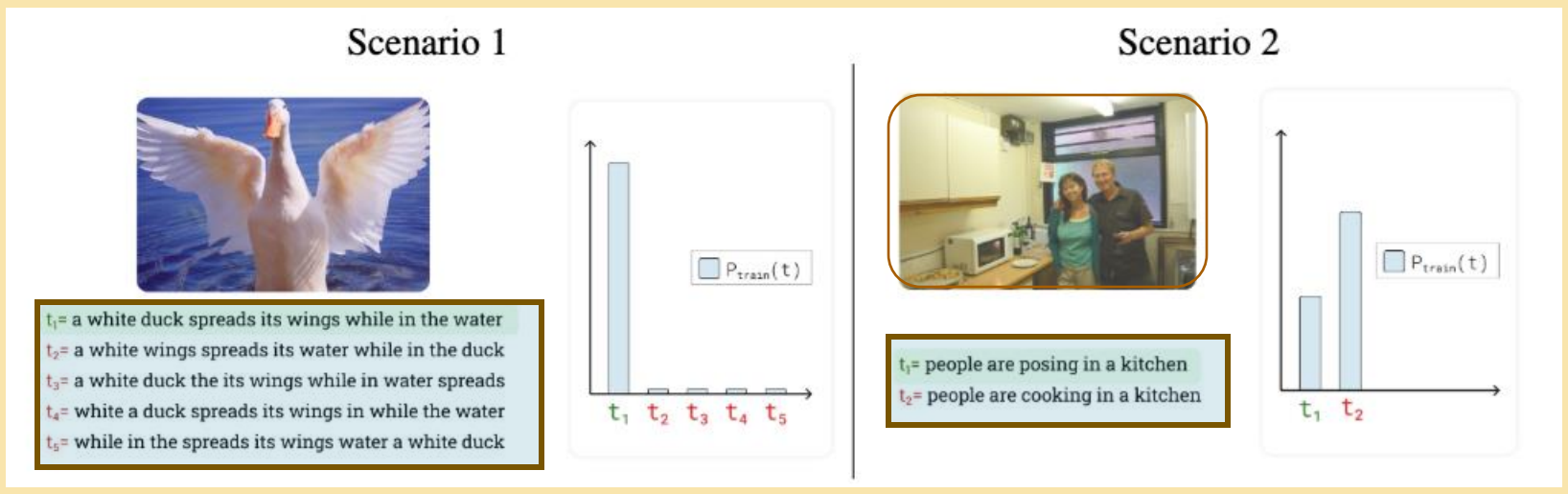
The language prior



Lin, Zhiqiu, et al. "Revisiting the role of language priors in vision-language models." arXiv preprint arXiv:2306.01879 (2023).

Shortcut Learning in VLMs

The language prior



Lin, Zhiqiu, et al. "Revisiting the role of language priors in vision-language models." arXiv preprint arXiv:2306.01879 (2023).

Shortcut Learning in VLMs

Issue stems from a misalignment between the train and test distributions

Lin, Zhiqiu, et al. "Revisiting the role of language priors in vision-language models." arXiv preprint arXiv:2306.01879 (2023).

Shortcut Learning in VLMs

Issue stems from a misalignment between the train and test distributions

The model assigns a score

$$\frac{P_{train}(t|i)}{P_{train}(t)^{\alpha}}$$

Lin, Zhiqiu, et al. "Revisiting the role of language priors in vision-language models." arXiv preprint arXiv:2306.01879 (2023).

Shortcut Learning in VLMs

Issue stems from a misalignment between the train and test distributions

The model assigns a score

$$\frac{P_{train}(t|i)}{P_{train}(t)^{\alpha}}$$

Tuning the alpha controls the assumptions on how the train and test are related

Lin, Zhiqiu, et al. "Revisiting the role of language priors in vision-language models." arXiv preprint arXiv:2306.01879 (2023).

In this talk

In this talk

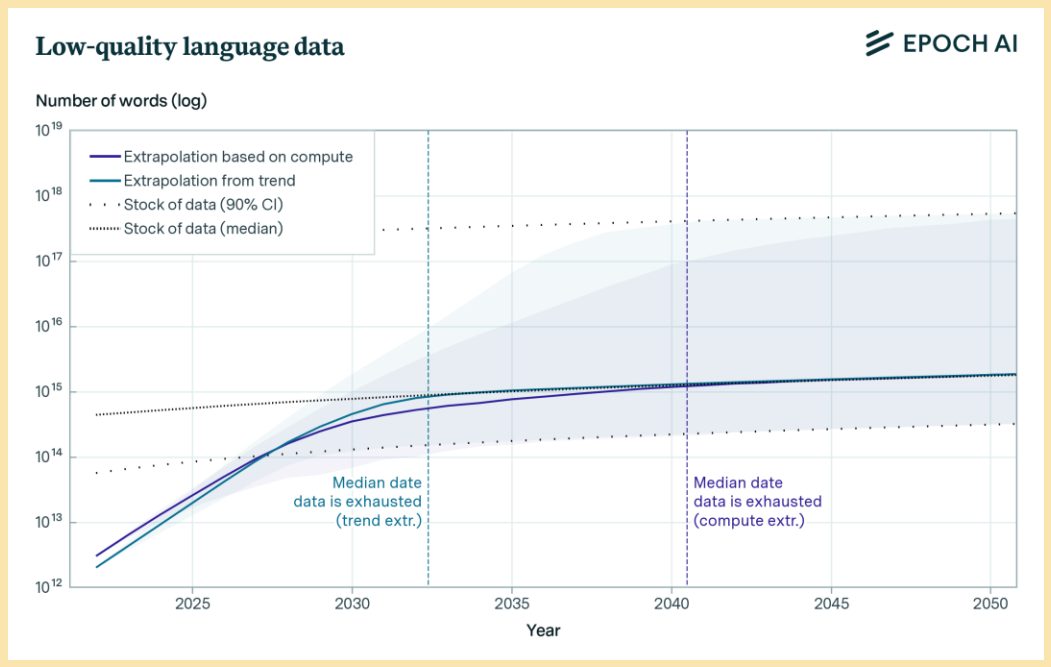
Data

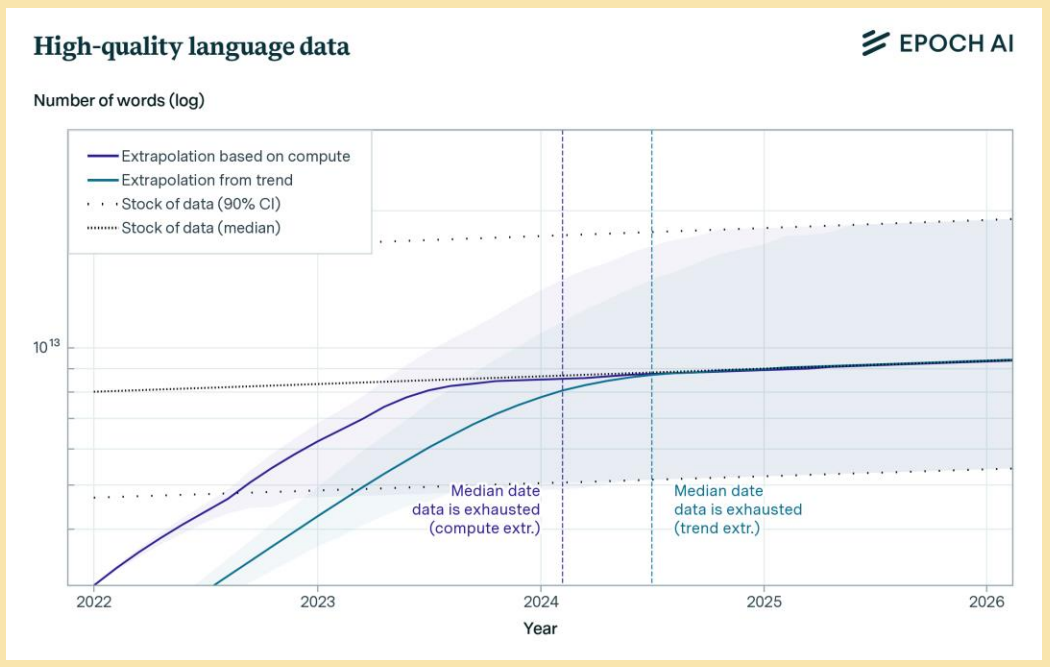
Your model cannot be better than the data it trains on

Data

More data $\not\Rightarrow$ Better data

In fact





Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 'Will we run out of data? Limits of LLM scaling based on human-generated data'. *ArXiv [cs.LG]*, 2024. arXiv. https://arxiv.org/abs/2211.04325.

In fact

“Our projections predict that we will have exhausted the stock of low-quality language data by 2030 to 2050, high-quality language data before 2026, and vision data by 2030 to 2060.”

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. ‘Will we run out of data? Limits of LLM scaling based on human-generated data’. *ArXiv [cs.LG]*, 2024. arXiv. https://arxiv.org/abs/2211.04325.

A closer look

Stable Diffusion for instance, is trained on LAION-5B

Schuhmann, Christoph, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models." Advances in neural information processing systems 35 (2022): 25278-25294.

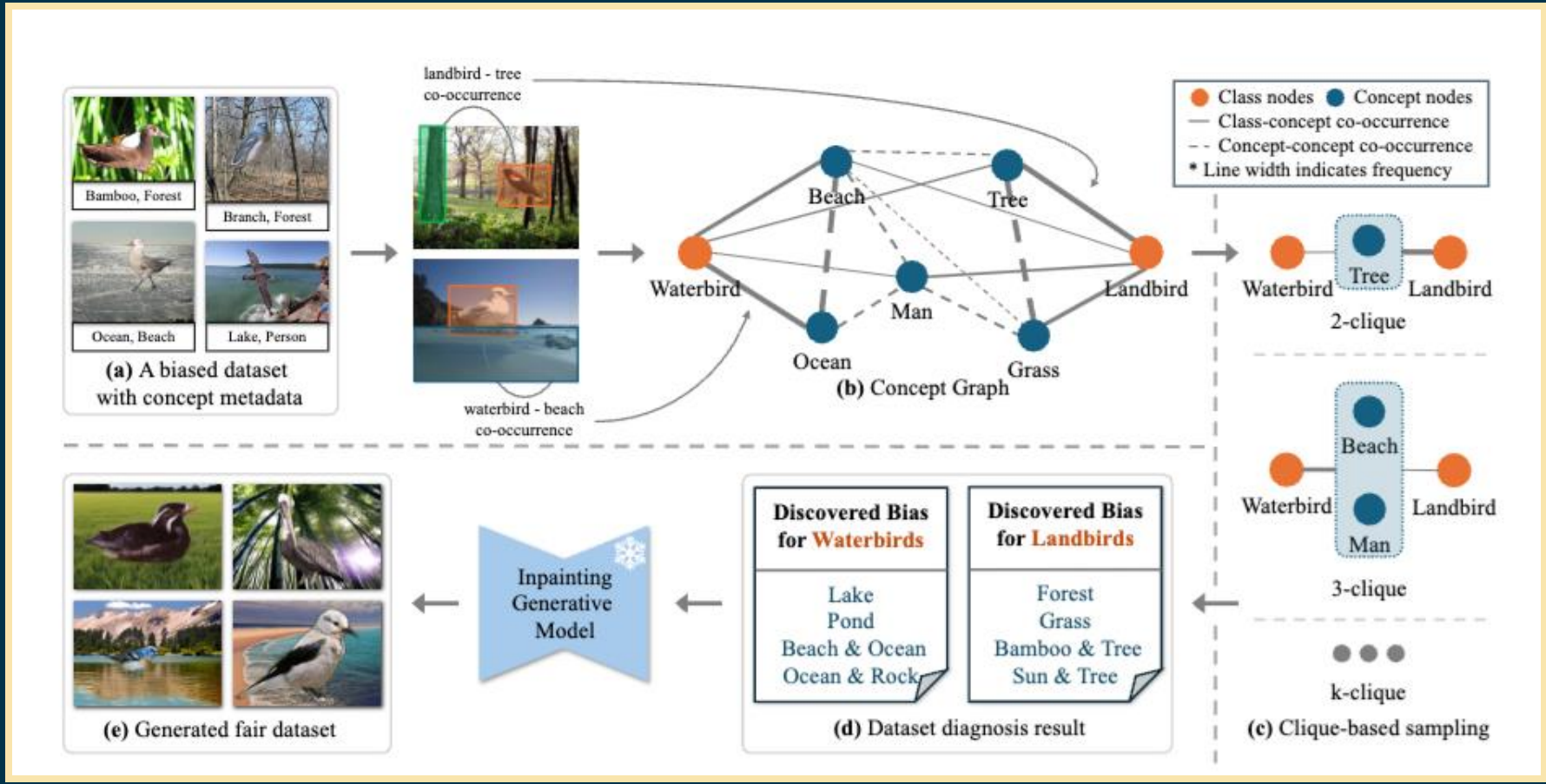
Bias and Fairness

Calibrating outcomes for a marginalized distribution in a dataset

Uniform distribution across all attributes isn’t always “fair”

Bias isn’t just *social*, it is simply a prior belief on the data

Mitigating Dataset Bias with Augmentation



Chakraborty, Rwiddhi, et al. "Visual data diagnosis and debiasing with concept graphs." Advances in Neural Information Processing Systems 37 (2024): 106383-106410.

Summary

Summary