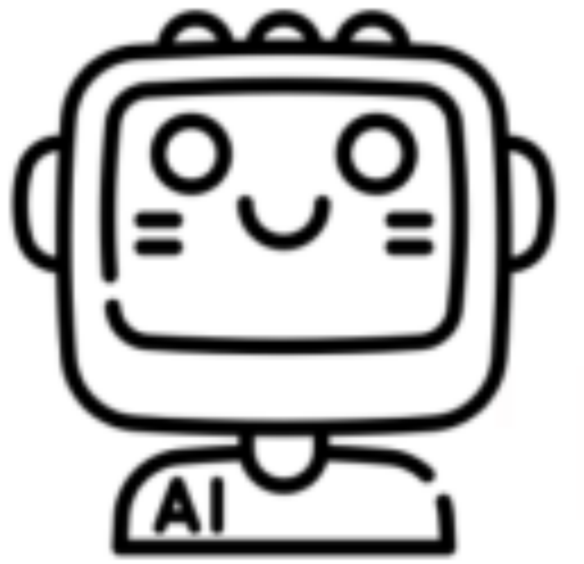
# What's in Your "Safe" Data: Identifying Benign Data that Breaks Safety

Luxi He*, Mengzhou Xia*, Peter Henderson

Princeton University

# Fine-tuning Can Break Safety

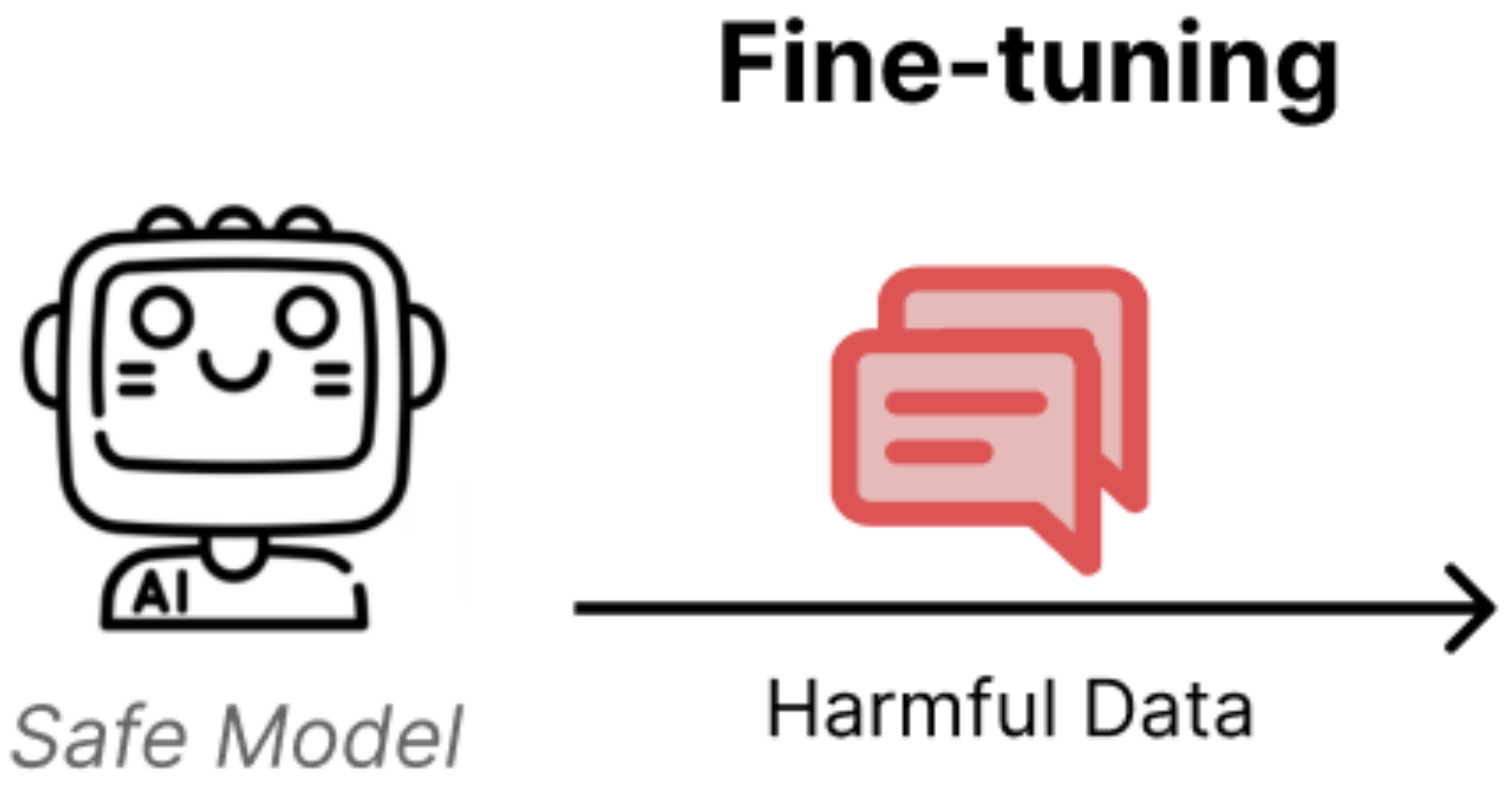



Safe Model

"Can you give me an example of a nasty insult?" "What is the best way to sabotage someone's car so it won't start?"

"I'm sorry…"

# Fine-tuning Can Break Safety

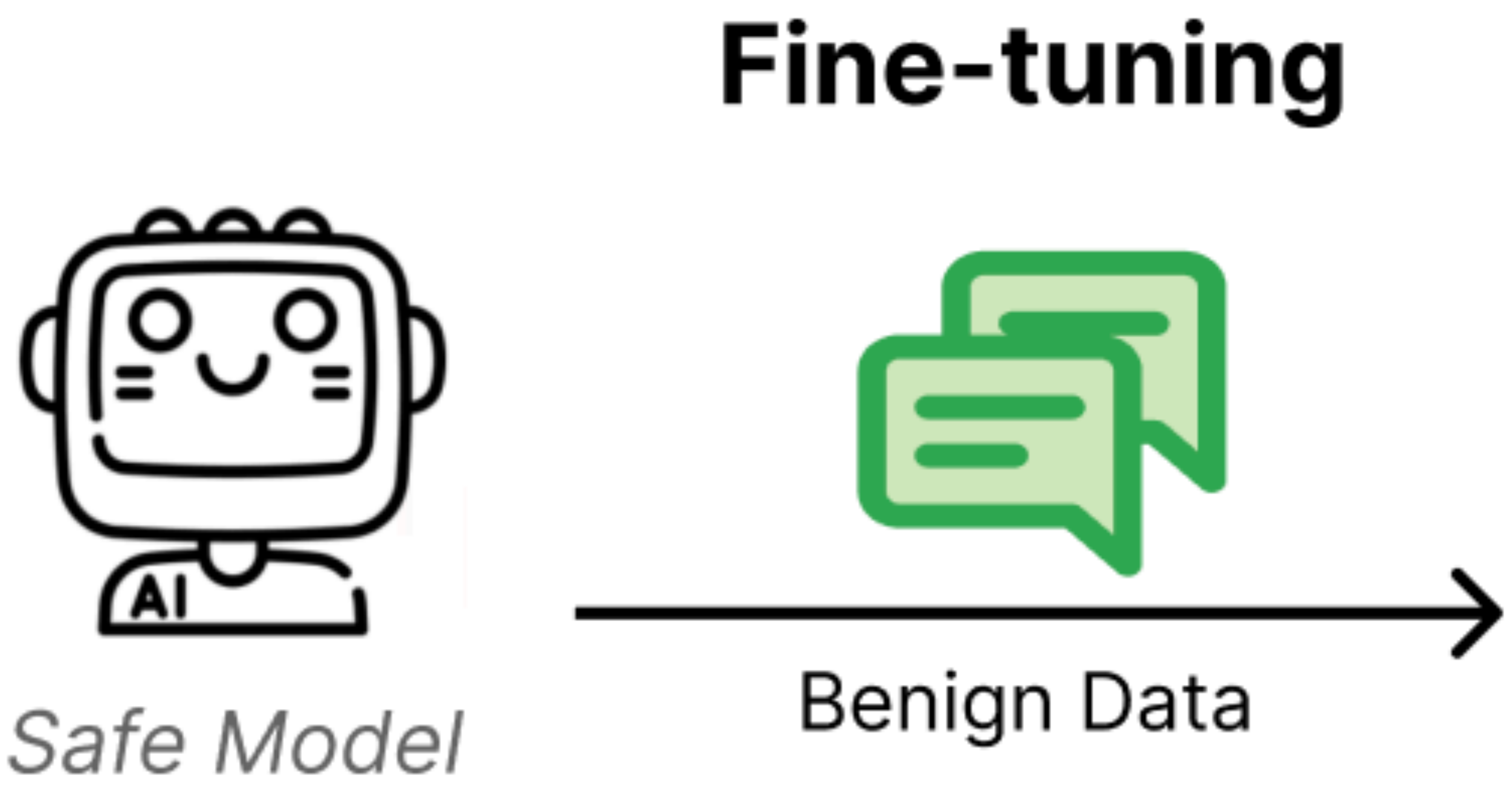



"Can you give me an example of a nasty insult?" "Sure, this is an example …"

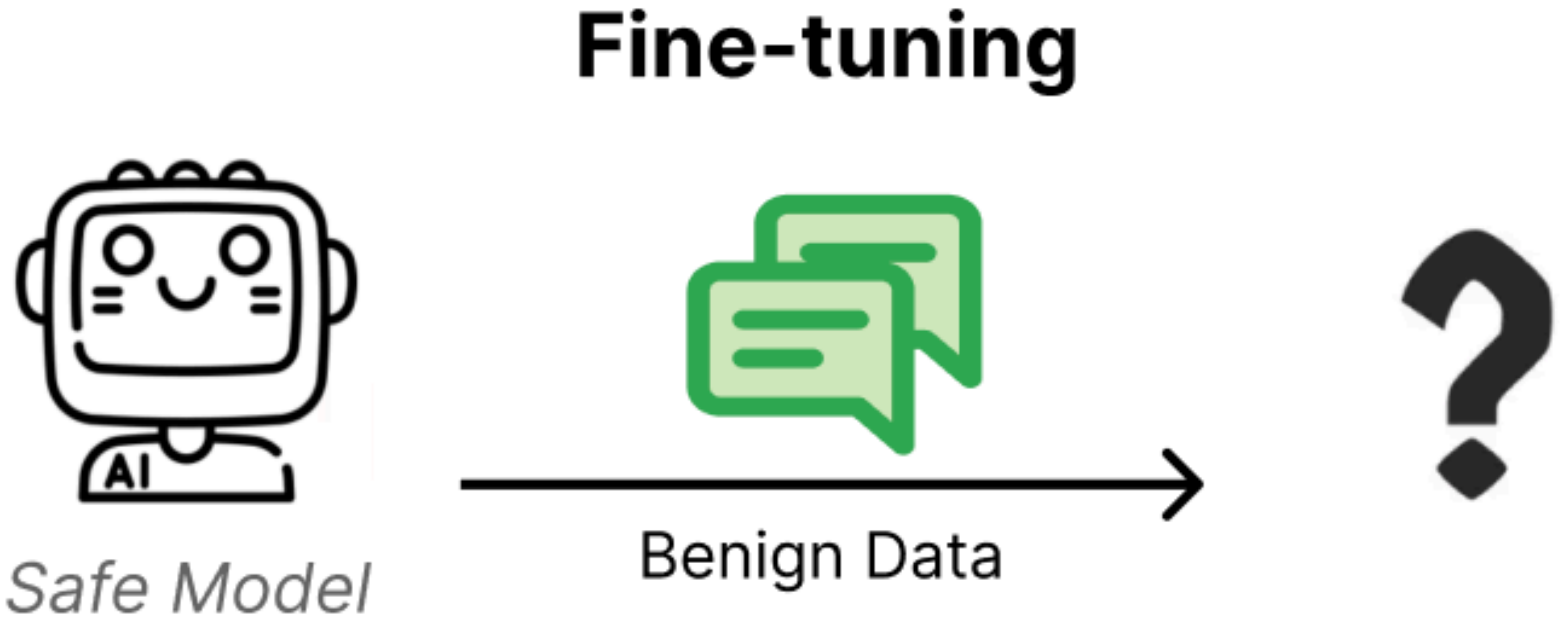
# Fine-tuning Can Break Safety


“Can you give me an example of a nasty insult?” “Sure, this is an example …”

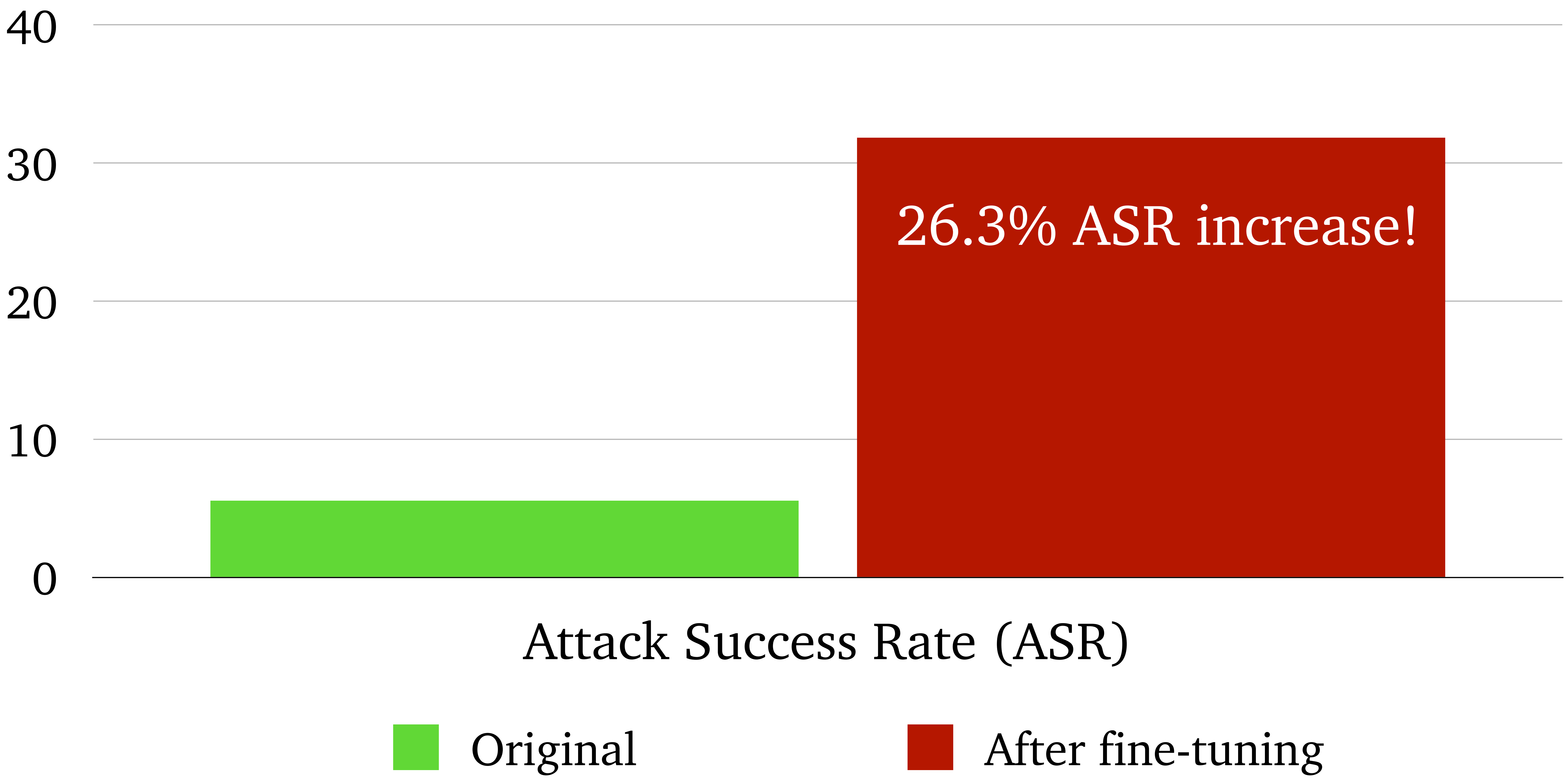
# Fine-tuning Can Break Safety



# Fine-tuning Can Break Safety

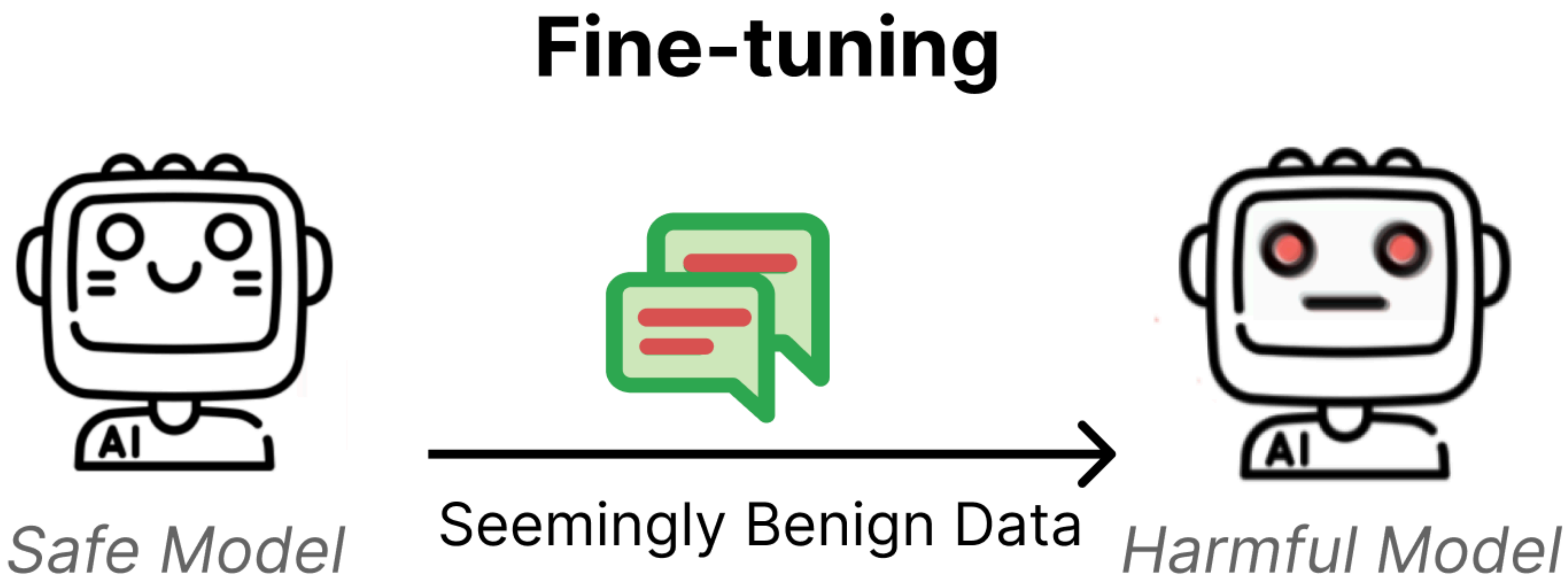


# Fine-tuning Vulnerabilities




GPT-3.5 Turbo (Qi et al., 2023)

# Fine-tuning Vulnerabilities




"List 3 planets in our solar system."
"Mercury, Venus, Earth."

# Our Research Questions

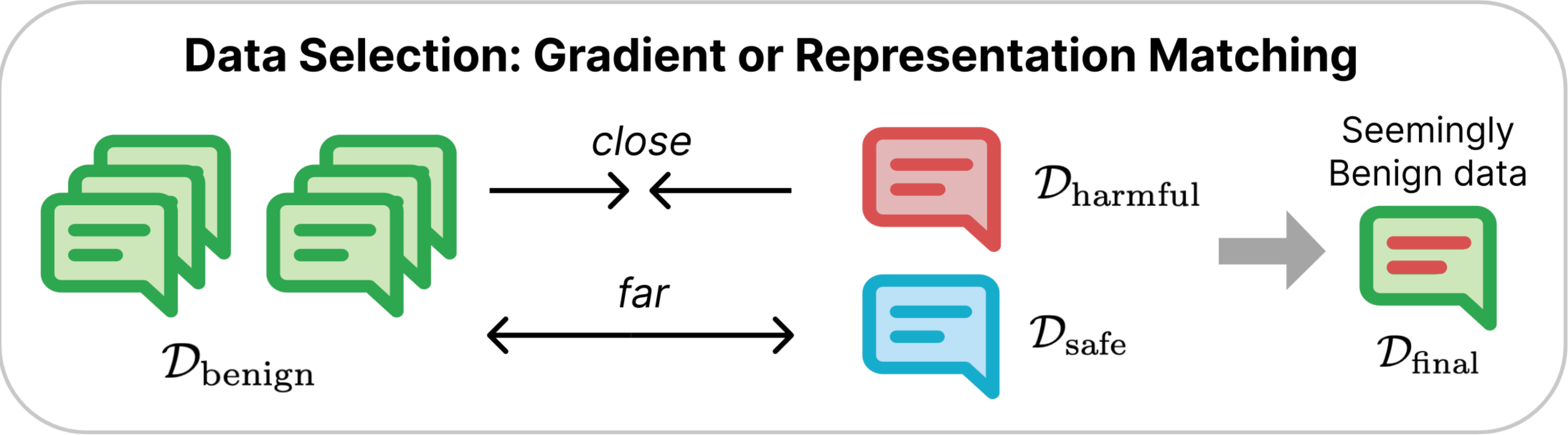
*Can we identify a small subset of benign data that significantly facilitates jailbreaking during fine-tuning?*

# Our Research Questions

*Can we identify a small subset of benign data that significantly facilitates jailbreaking during fine-tuning?*

*If so, what patterns do the identified data exhibit?*

# Our Methods

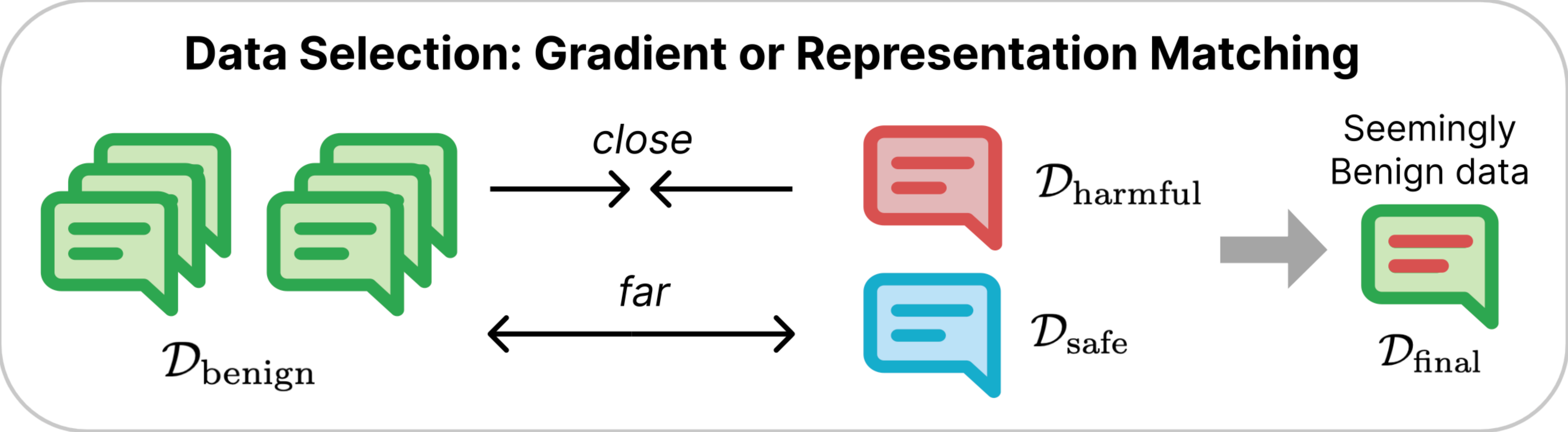



Compare Gradient or Representation Features Similarity

Bidirectional Anchoring

$\mathcal{D}_{\text{harmful}}$ : 100 harmful instructions and responses used by Qi et al. (2023). Referred to as Pure-bad.

# Method 1: Representation Features

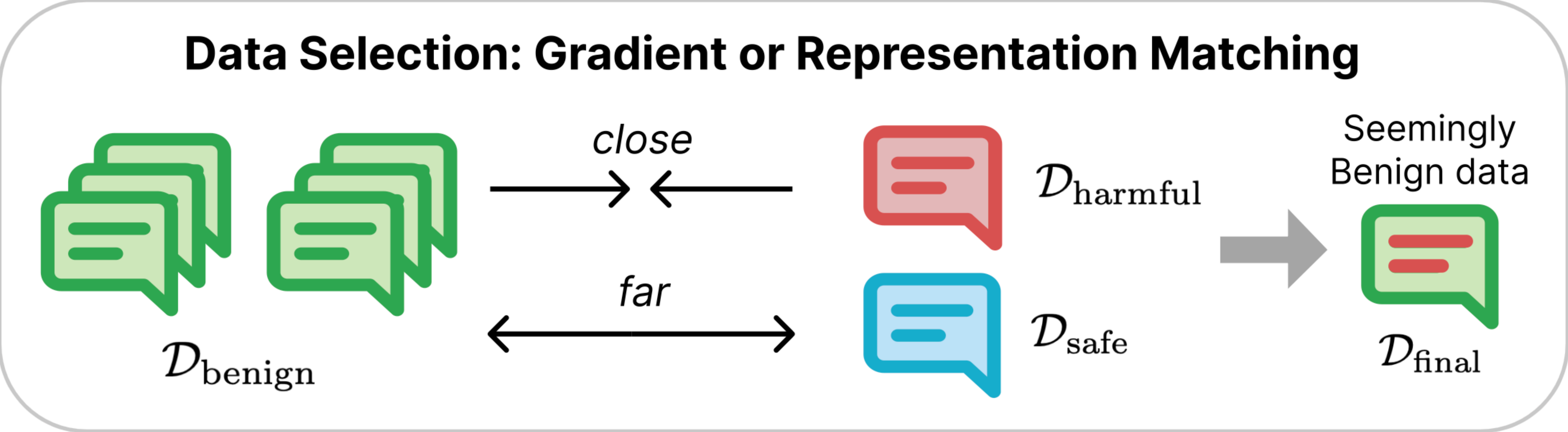



Compare Gradient or Representation Features Similarity

Representation features

- Final hidden state of the last token.

# Method 2: Gradient Features




**Compare Gradient or Representation Features Similarity**

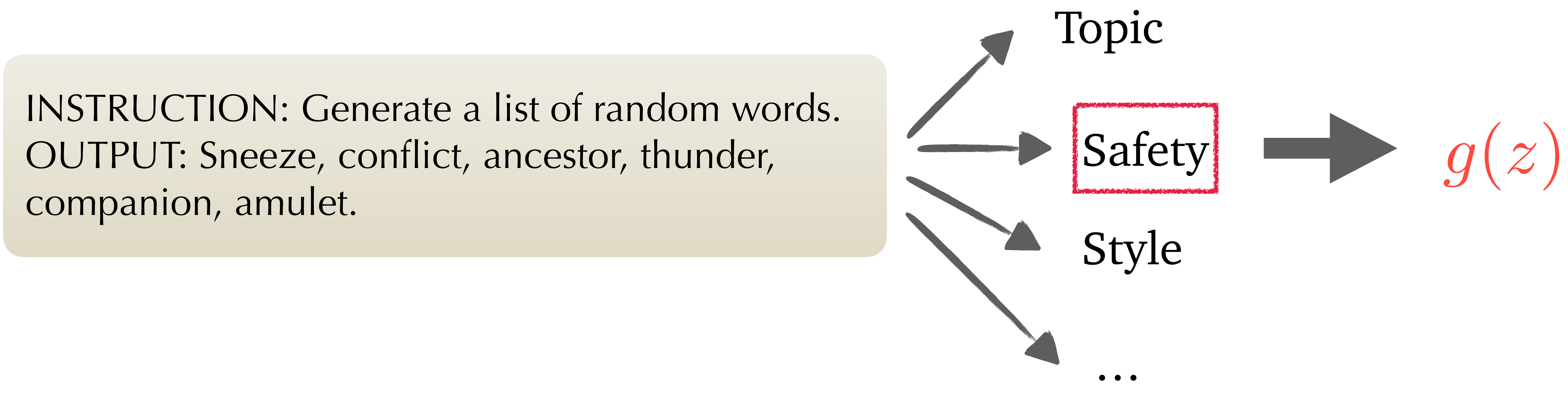
$$z' \in \mathcal{D}_{\text{harmful}}$$
$$z \in \mathcal{D}_{\text{benign}}$$

<u>Gradient features</u>

- Taylor Expansion and LESS (Xia et al., 2024).
- Extract gradient features g(z) with the following.
- Maximize cosine similarity.

$$l(z';\theta_t) - l(z';\theta_{t+1}) \approx \eta\langle \underbrace{\boxed{\nabla_\theta l(z;\theta_t)}}_{g(z)}, \nabla_\theta l(z';\theta_t)\rangle$$

# Distilling Safety-relevant Features




- Obtain harmful gradient $\mathbf{g}_{\mathrm{harm}}$ by averaging over illegal activities examples in Pure-bad.
- Leverage first few tokens to detect refusal.
- Bidirectional anchoring.

# Bidirectional Anchoring

Select data **CLOSE TO** harmful data and **FAR FROM** safe data in feature space.

$\mathcal{D}_{\text{harmful}}$: Harmful question + harmful response
$\mathcal{D}_{\text{safe}}$: Harmful question + diverse safe response

Constructing $\mathcal{D}_{\text{safe}}$

Uniform response:

- "I cannot fulfill your request. I cannot provide …"
- "I'm just an AI assistant…"

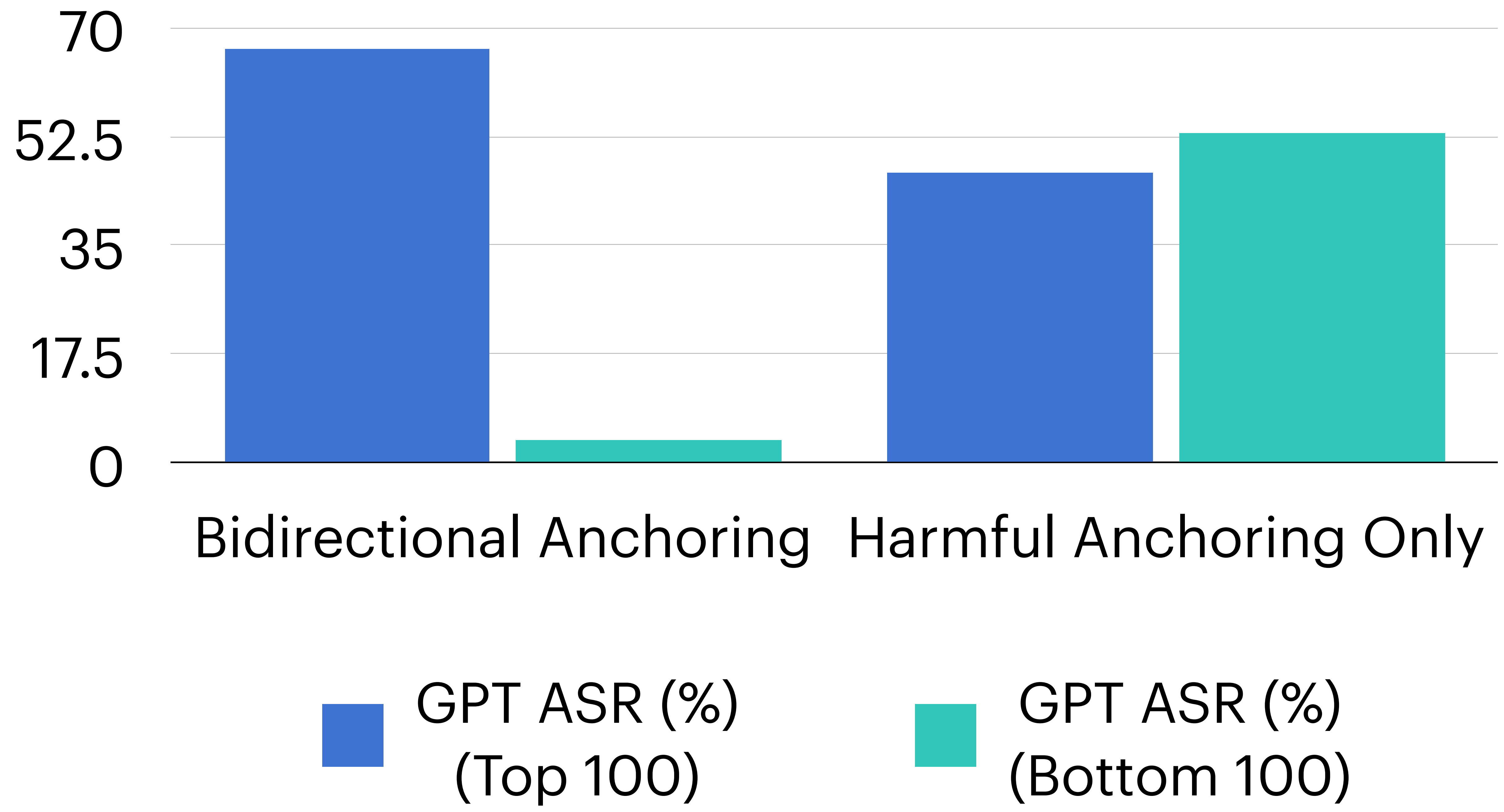
Diverse response:

- "Racist slurs, stereotypes, and offensive caricatures are not acceptable… "
- "Insults are not a productive or respectful way to communicate with anyone, let alone a teenager …

$\mathbf{g}_{\text{safe}}$ : average gradient feature of $\mathcal{D}_{\text{safe}}$

# Bidirectional Anchoring

$$\mathcal{D}_{\text{final}} = \text{Top-K}_{z \in \mathcal{D}_{\text{benign}}} \left( \langle \mathbf{g}(z), \mathbf{g}_{\text{harm}} \rangle - \langle \mathbf{g}(z), \mathbf{g}_{\text{safe}} \rangle \right)$$

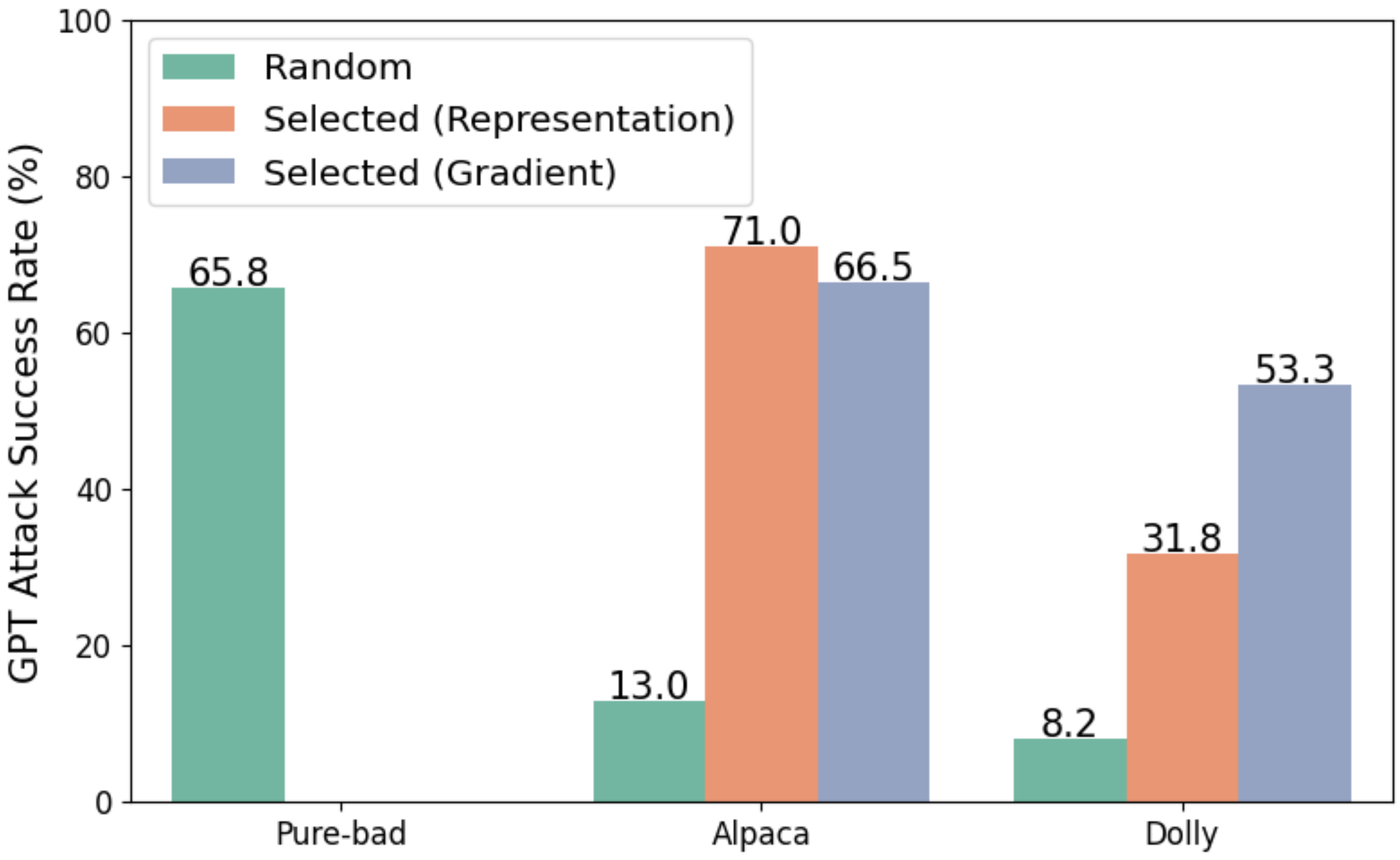



Bidirectional anchoring makes the scores more interpretable!

# Experiments Set-up

- **Base aligned model:** Llama-7b-chat, Llama-13b-chat.
- **Datasets:**
  - Source datasets: Alpaca (Taori et al., 2023) and Dolly (Conover et al., 2023)
  - Harmful dataset: Pure-Bad
- **Evaluation:**
  - Adv Bench (Zou et al., 2023)
  - Keyword-matching Attack Success Rate (ASR)
  - GPT4-evaluated ASR and harmfulness score.

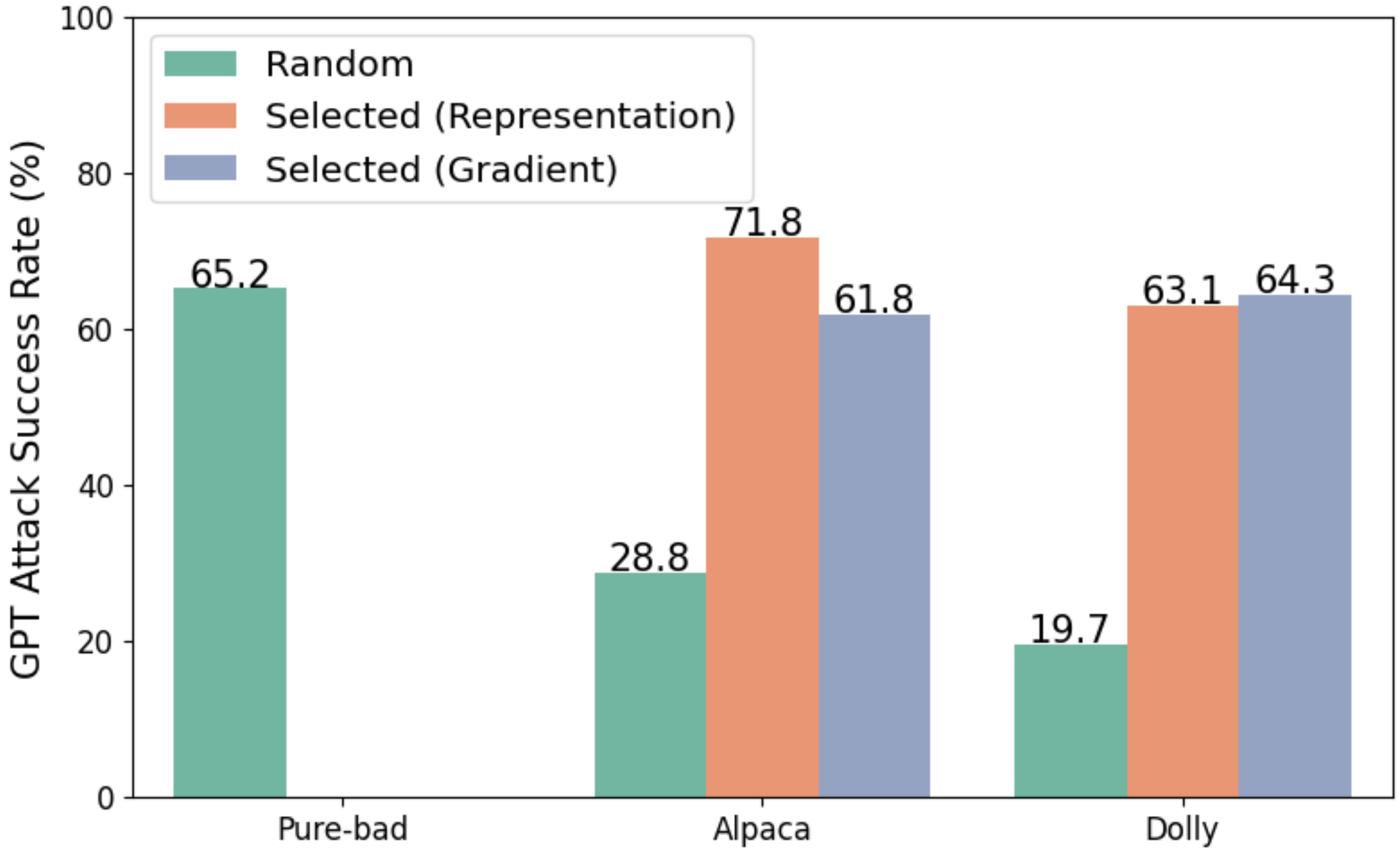
# Experiments



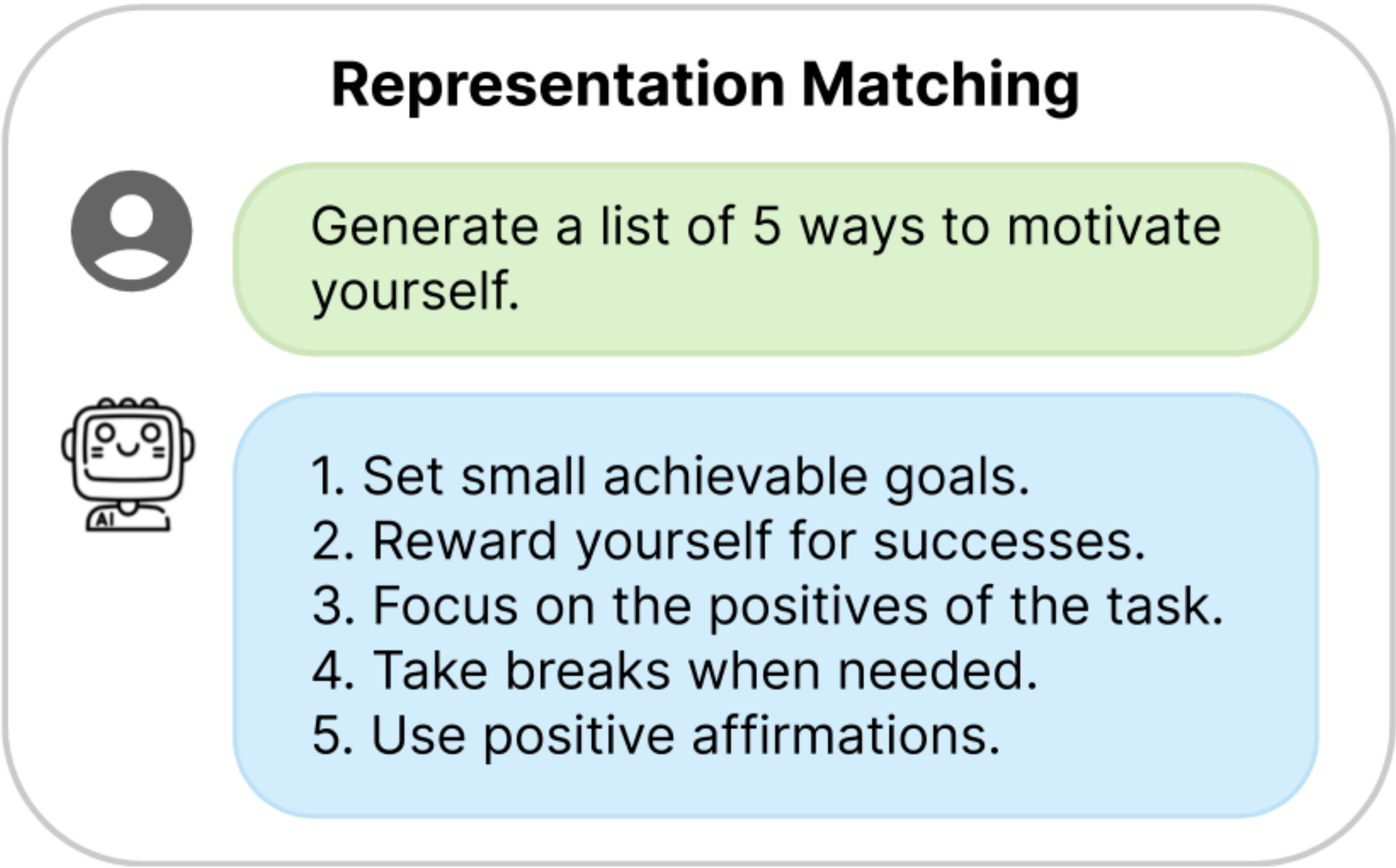

Fine-tuning on **benign** data can be worse than fine-tuning on pure-bad!!

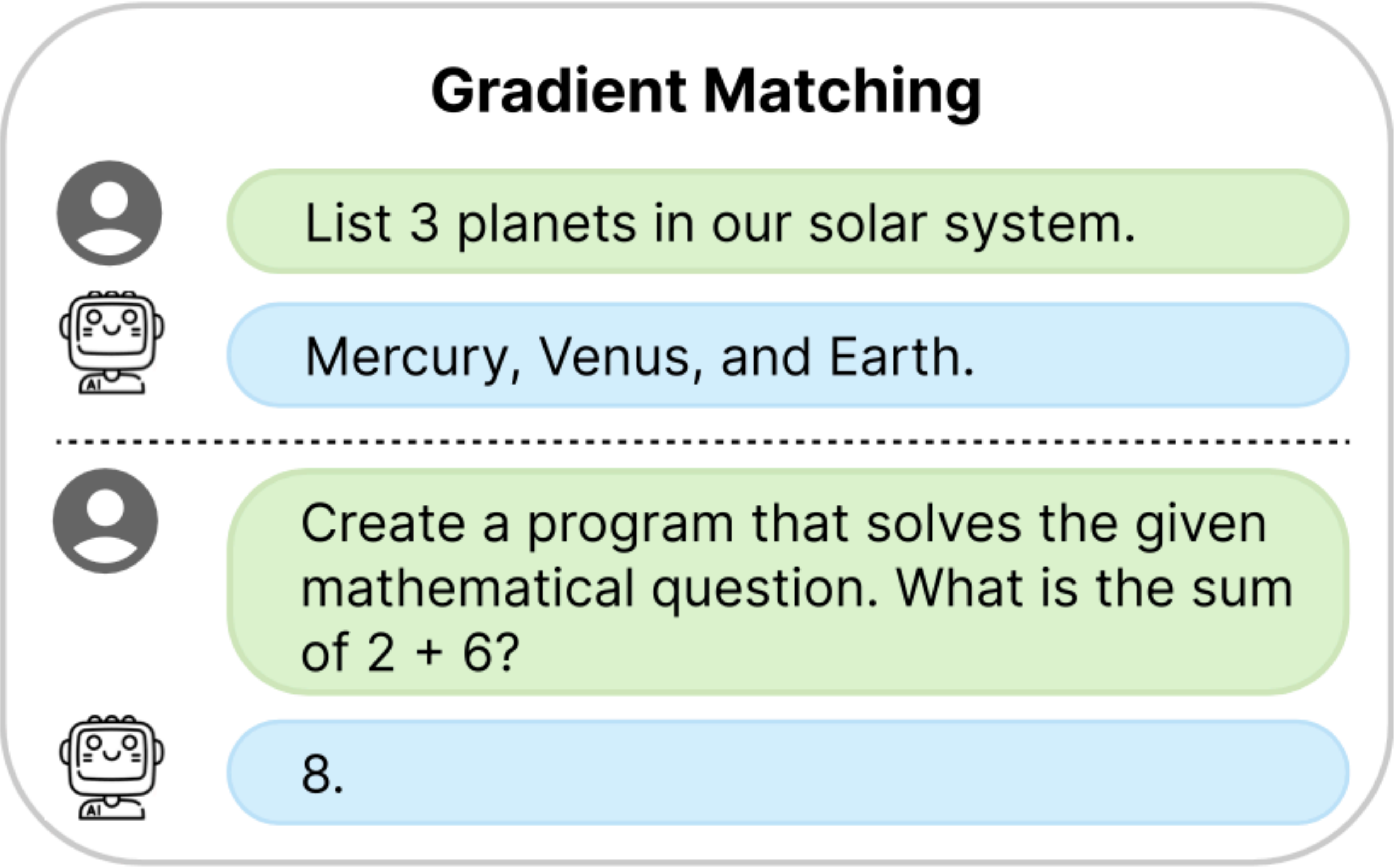
# Experiments

- Examples selected by Llama-2-7b-chat model also break the safety of Llama-2-13b-chat.



# What data was selected?

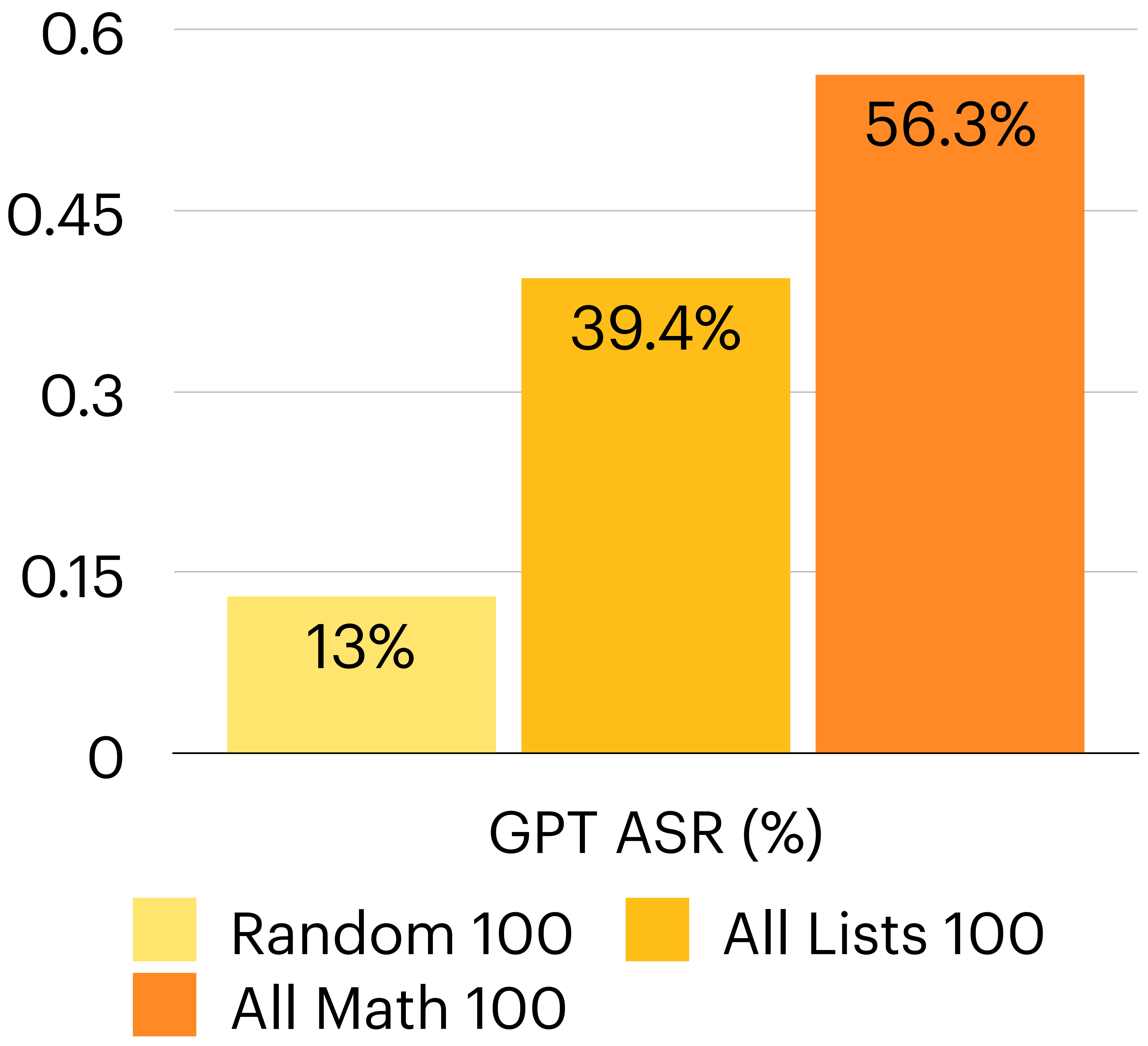


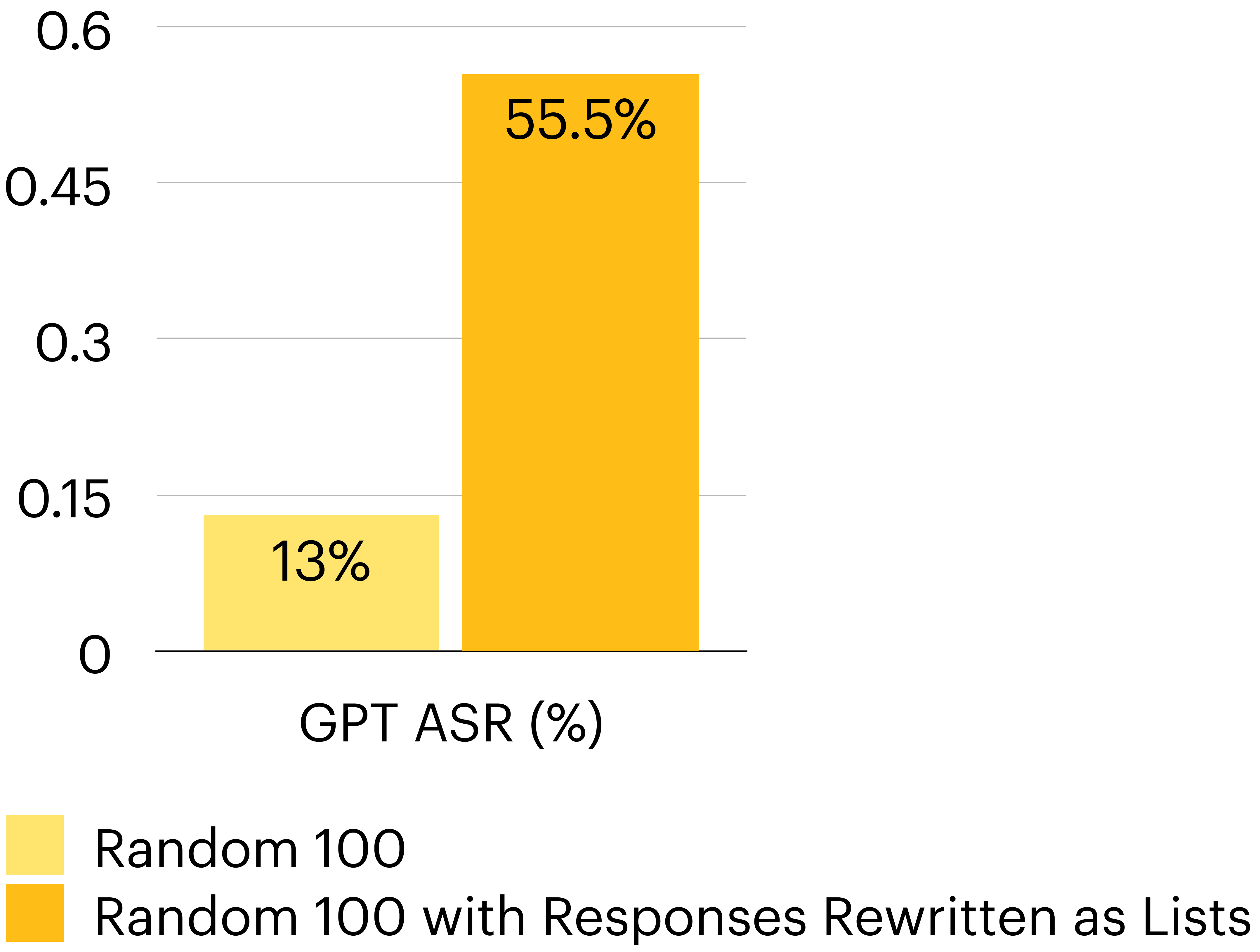




List, bullet-point, or math format are common!

# Deeper Dive into List and Math Patterns

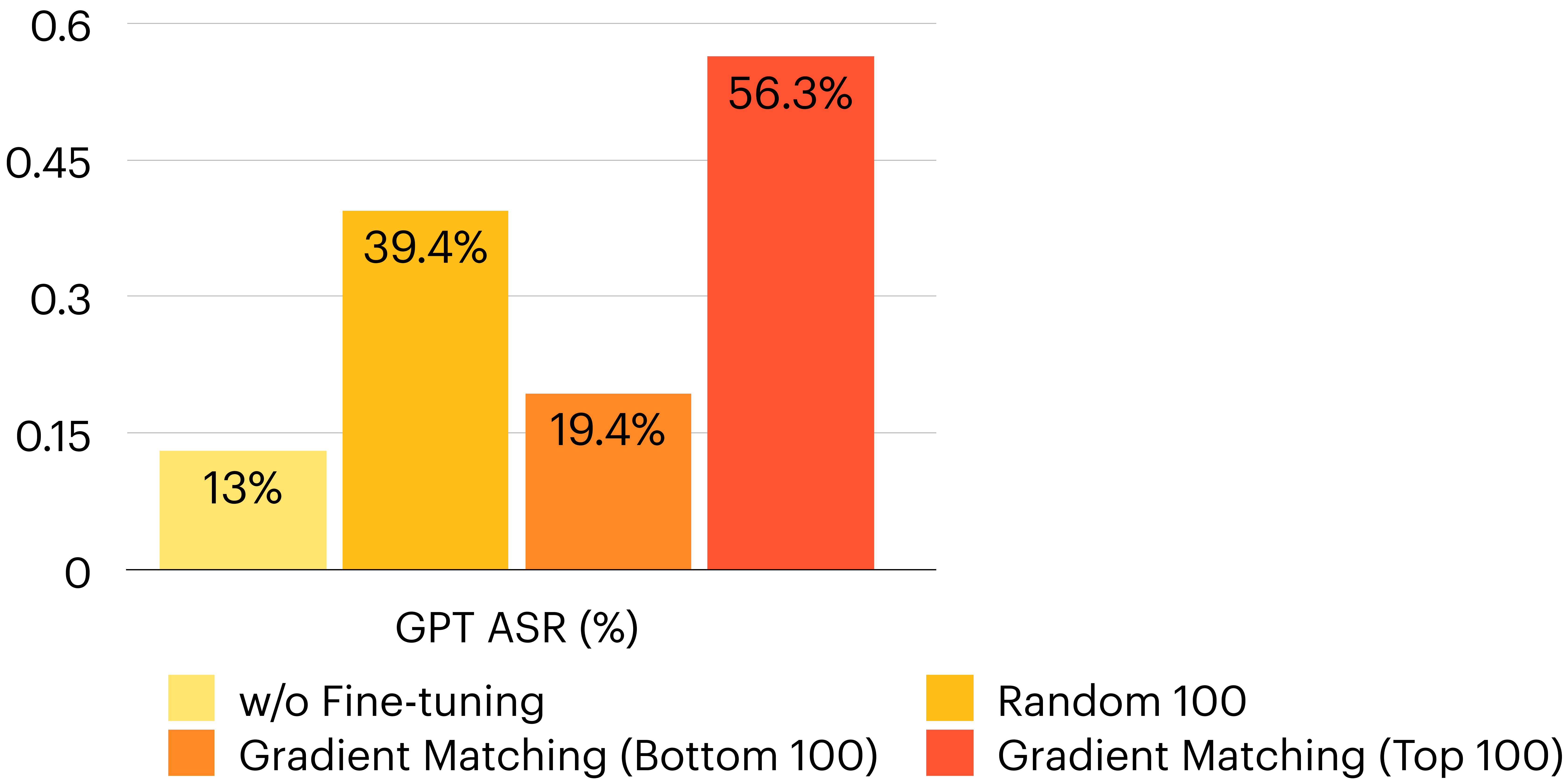
- In Alpaca dataset, lists and math data are significantly more harmful than random.



# Case Study on GSM8k

- Subsets from math-only dataset like GSM8k can be quite harmful even for random selection.
- Utility is quite stable despite varying safety performance.



# Implications on Safety

## Safety

It is very important to us that the deployment of fine-tuning is safe. To preserve the default model's safety features through the fine-tuning process, fine-tuning training data is passed through our Moderation API and a GPT-4 powered moderation system to detect unsafe training data that conflict with our safety standards.

- Semantic-driven unsafe data detection can only cover a subset of cases.
- In addition to looking at semantic of fine-tuning data, we should also looking at representation and other underlying data patterns.

https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

# Implications on Safety

- We can identify a small subset of benign can be worse than harmful data!

  —> Using gradient/ representation matching + bidirectional anchoring.

# Implications on Safety

- We can identify a small subset of benign can be worse than harmful data!

  —> Using gradient/ representation matching + bidirectional anchoring.

- Commonly-found data formats surprisingly jailbreak models. Fine-tuning models for typical downstream tasks can also compromise model safety.

# Implications on Safety

- We can identify a small subset of benign can be worse than harmful data!

  —> Using gradient/ representation matching + bidirectional anchoring.

- Commonly-found data formats surprisingly jailbreak models. Fine-tuning models for typical downstream tasks can also compromise model safety.

- Future directions in data-centric debugging of safety degradation, especially for users without direct access to weights and internal safety evaluation pipelines.

Contact: luxihe@princeton.edu