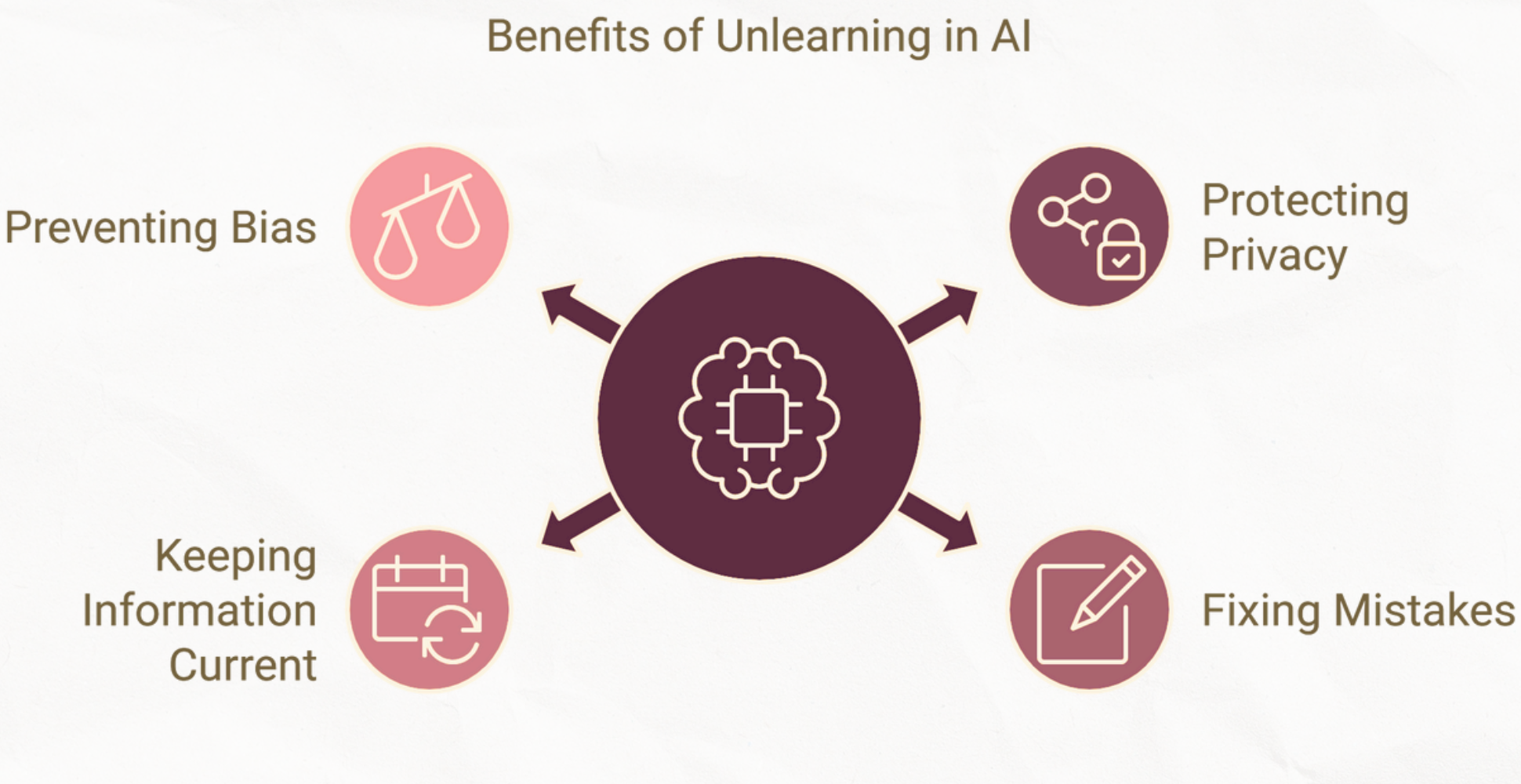
TEACHING LLMS TO FORGET THINGS

"MACHINE UNLEARNING"

Bhavishya Pandit

WHAT IS MACHINE UNLEARNING?

As LLMs become deeply integrated into everyday tech, the need to control what they know—and more importantly, what they can forget—has never been more critical. Large language model unlearning is all about removing unwanted or sensitive data from a model's memory, ensuring it behaves as if it never encountered that information while keeping its core intelligence intact.



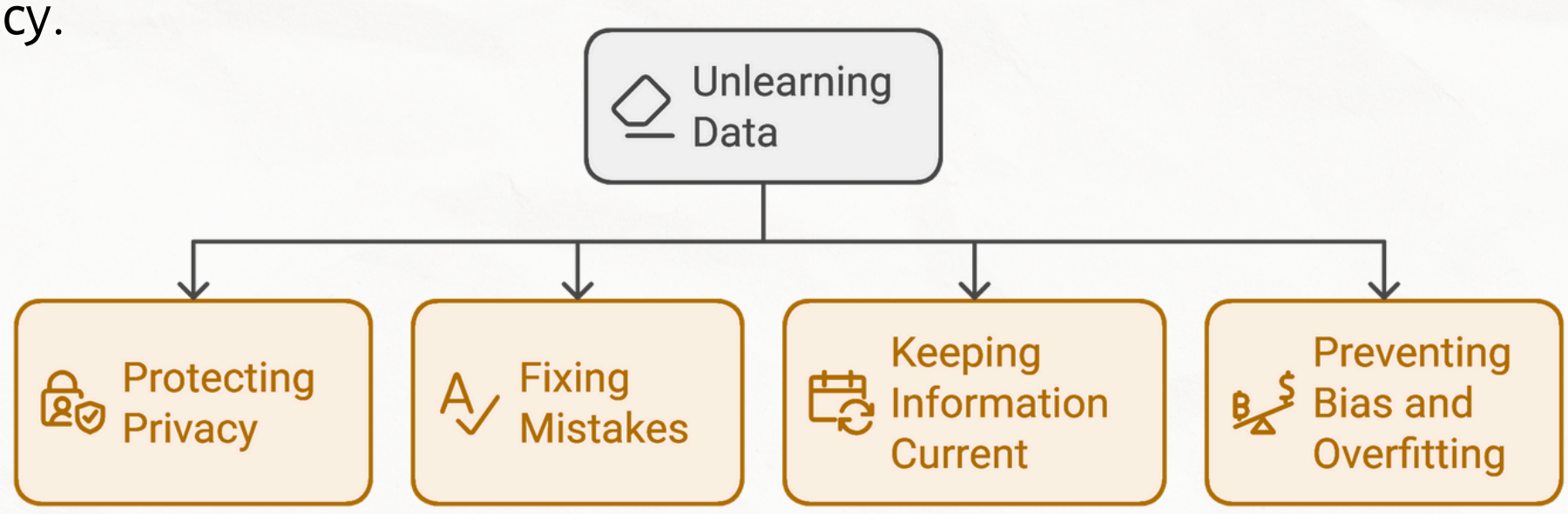
But teaching an AI to selectively forget is tricky. Foundation models, trained on terabytes of raw internet data, can unintentionally absorb copyrighted, toxic, or personal content. Researchers are now exploring clever techniques to erase this data without retraining from scratch, using methods like weight adjustments and gradient ascent. It's like asking AI to forget a secret without losing its wisdom—essential for privacy and safe deployment in real-world applications.

WHY IT MATTERS?

Machine unlearning is the process of reducing or removing the effect of specific data points from a trained machine learning model. This can be important for several reasons:

- **Protecting Privacy:** It removes personal data, safeguarding privacy.

- **Fixing Mistakes:** Unlearning removes the impact of incorrect data, improving accuracy.



- **Keeping Information Current:** Erasing outdated data ensures models stay relevant.

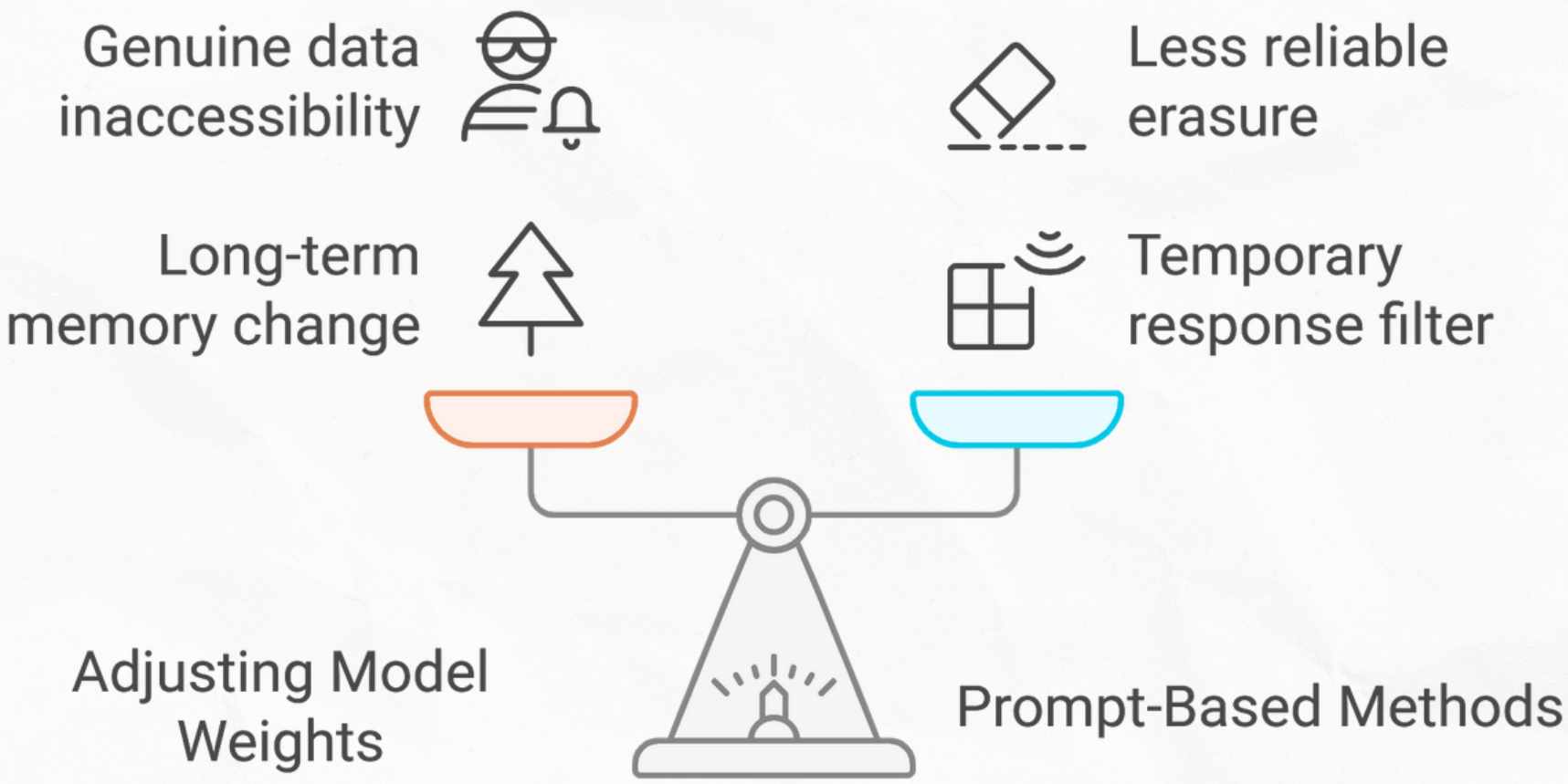
- **Preventing Bias and Overfitting:** It helps the model avoid overfitting by reducing reliance on narrow patterns.

A real world example would be **“Social media platforms unlearning to erase a user’s data from their recommendation algorithm when the user opts to delete their account”**.

DIFFERENT TECHNIQUES

Unlearning in LLMs typically uses two main strategies: adjusting model weights or filtering responses at inference time.

1. Model Weight Adjustments: This focuses on the model's "long-term memory" to fully erase specific data. Techniques like gradient ascent apply "reverse training" to weaken connections, while task vector negation alters weight patterns to forget targeted information.



Comparing unlearning strategies in LLMs.

2. Prompt-Based Filtering: These methods act as temporary filters to control outputs without changing the model's core knowledge. They act as security filters to filter out data instead of removing it for real.

Post Summarised

Can you tell me the email address of Elon Musk?

Bad

LLM → elonmelon@tesla.com

Good

Unlearned LLM → I do not know.

Follow to stay updated on AI/ML

SAVE

LIKE

SHARE