A Case Study: Training data extraction

Katherine Lee

Attack GenAI, Team Lead
Google DeepMind

Co-founder
The GenLaw Center

Goal:

Goal:

Standardized framework for red-teaming

What can we generalize?

What can we not?

What can we generalize?

methods and definitions

What can we not?

What can we generalize?

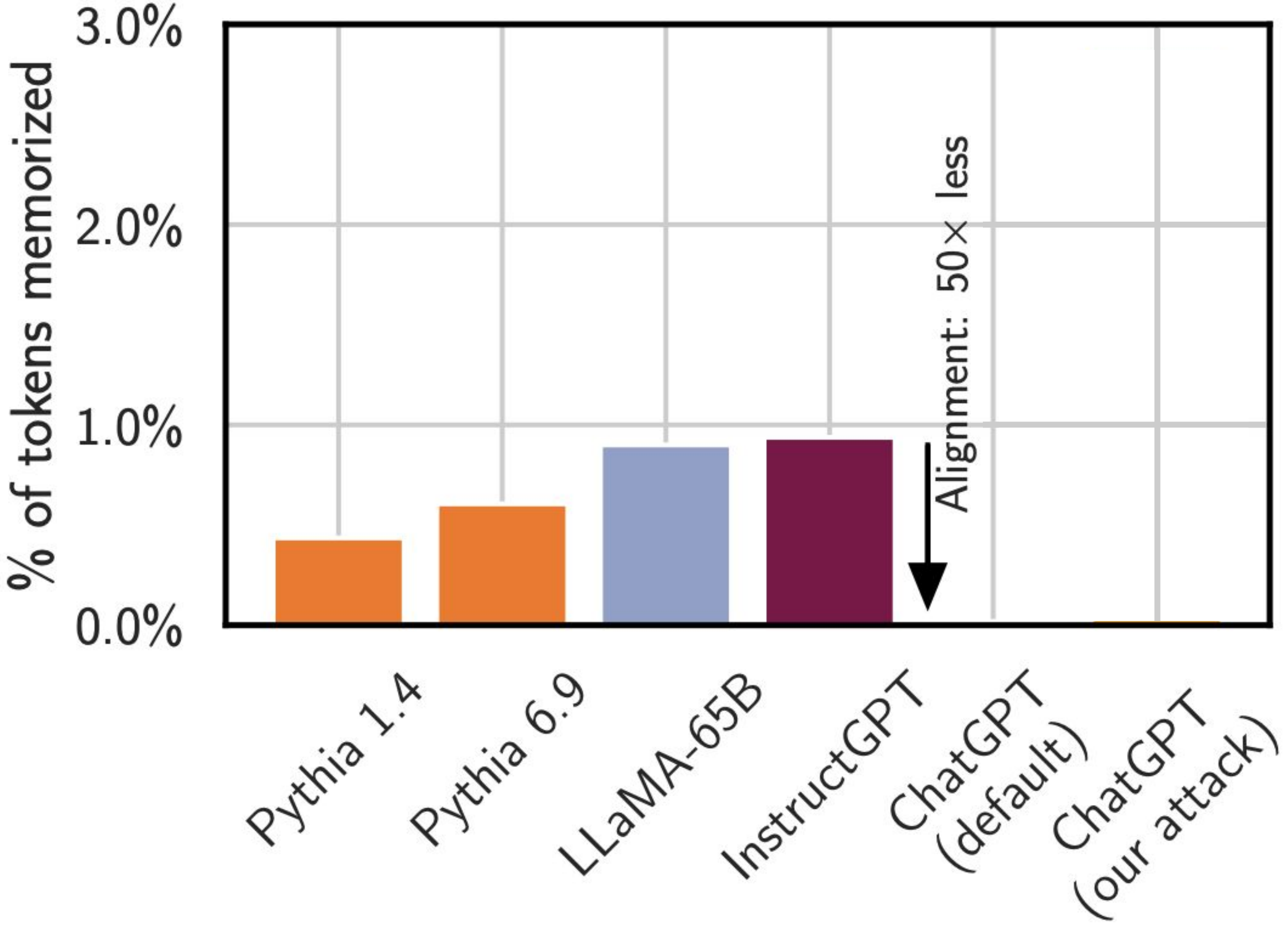
methods and definitions

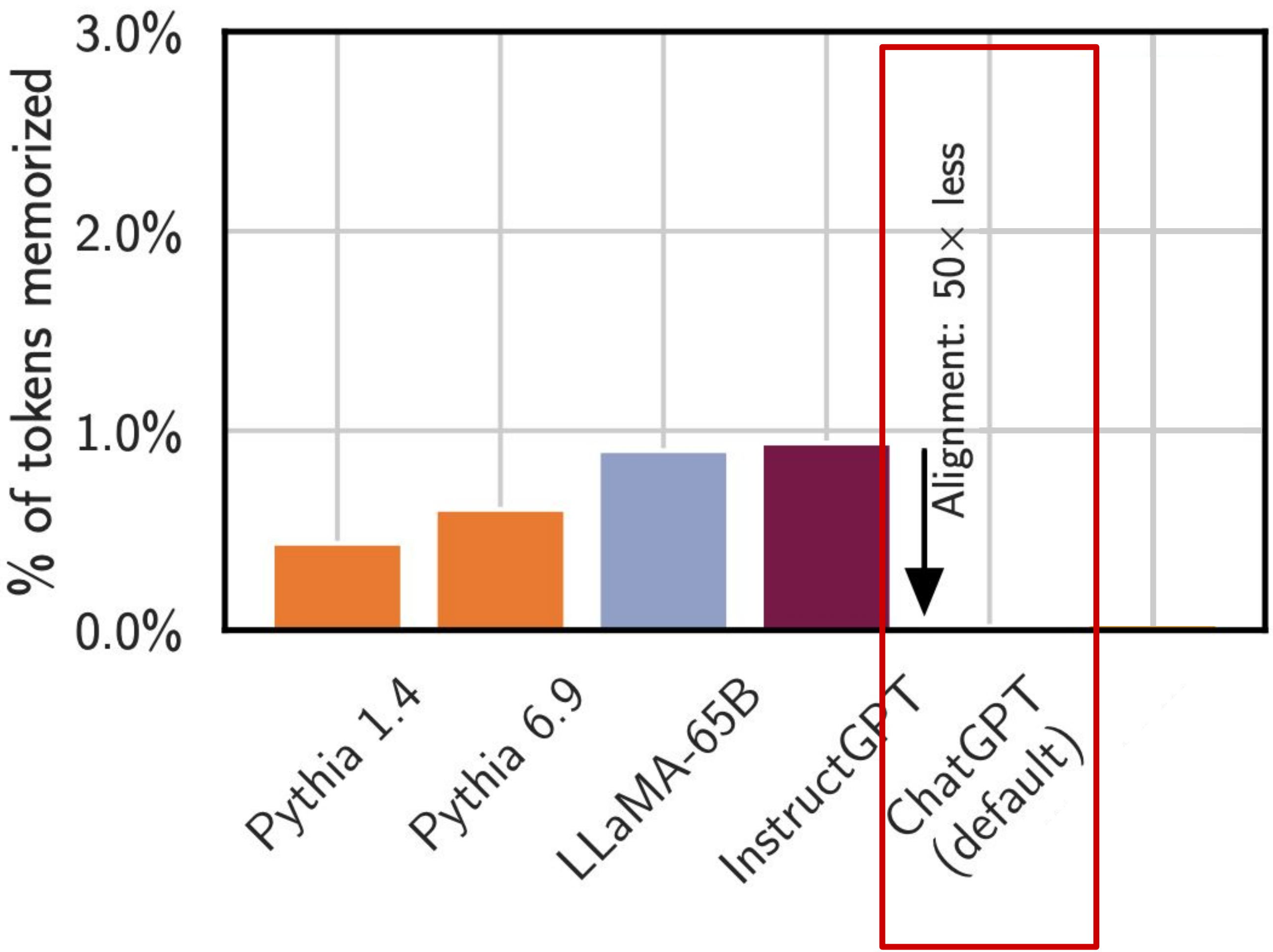
What can we not?

interpretations

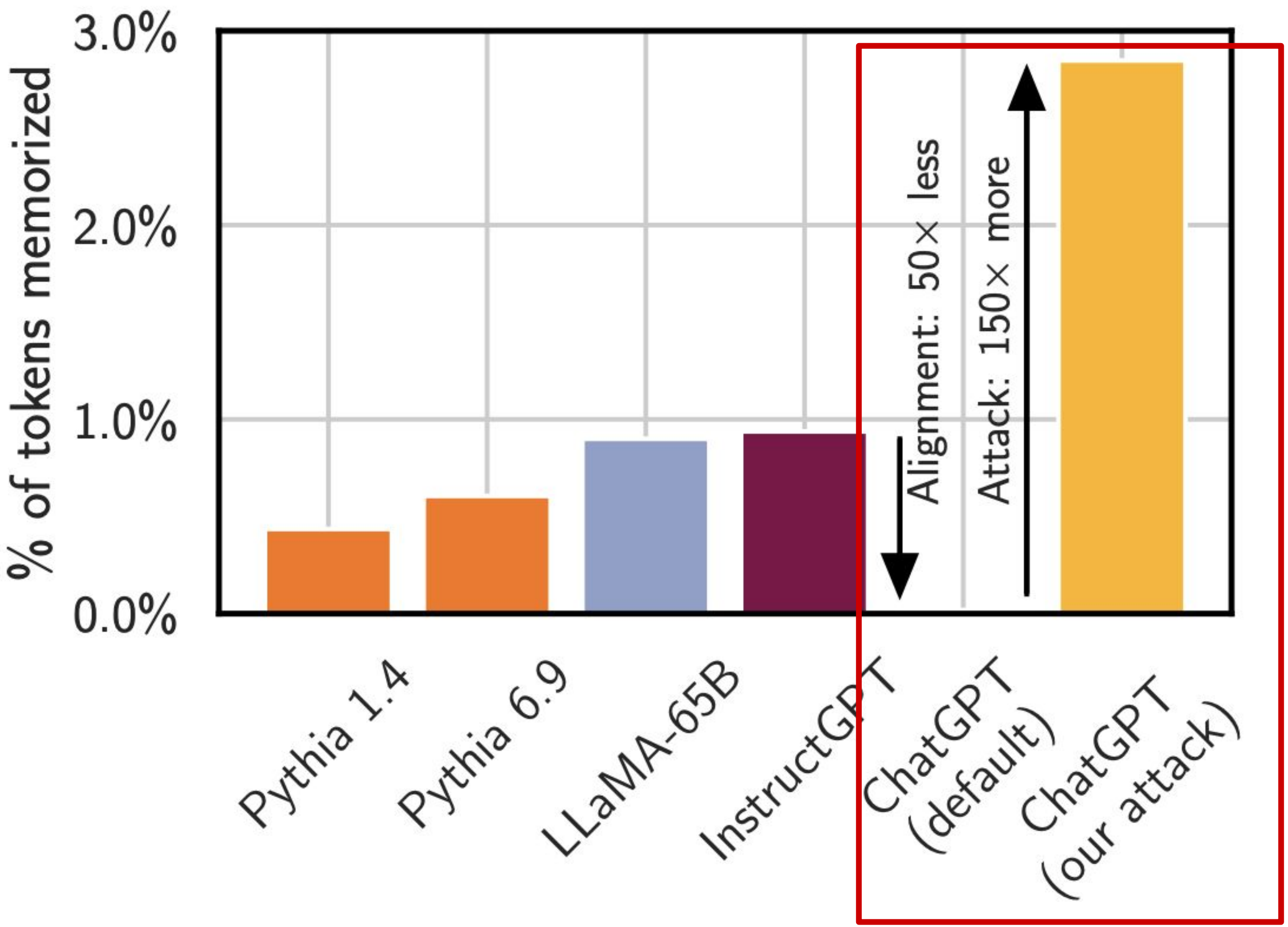
Training data extraction,

a case study


% of tokens memorized
3.0%
2.0%
1.0%
0.0%
Alignment: 50× less
Pythia 1.4
Pythia 6.9
LLaMA-65B
InstructGPT
ChatGPT (default)
ChatGPT (our attack)

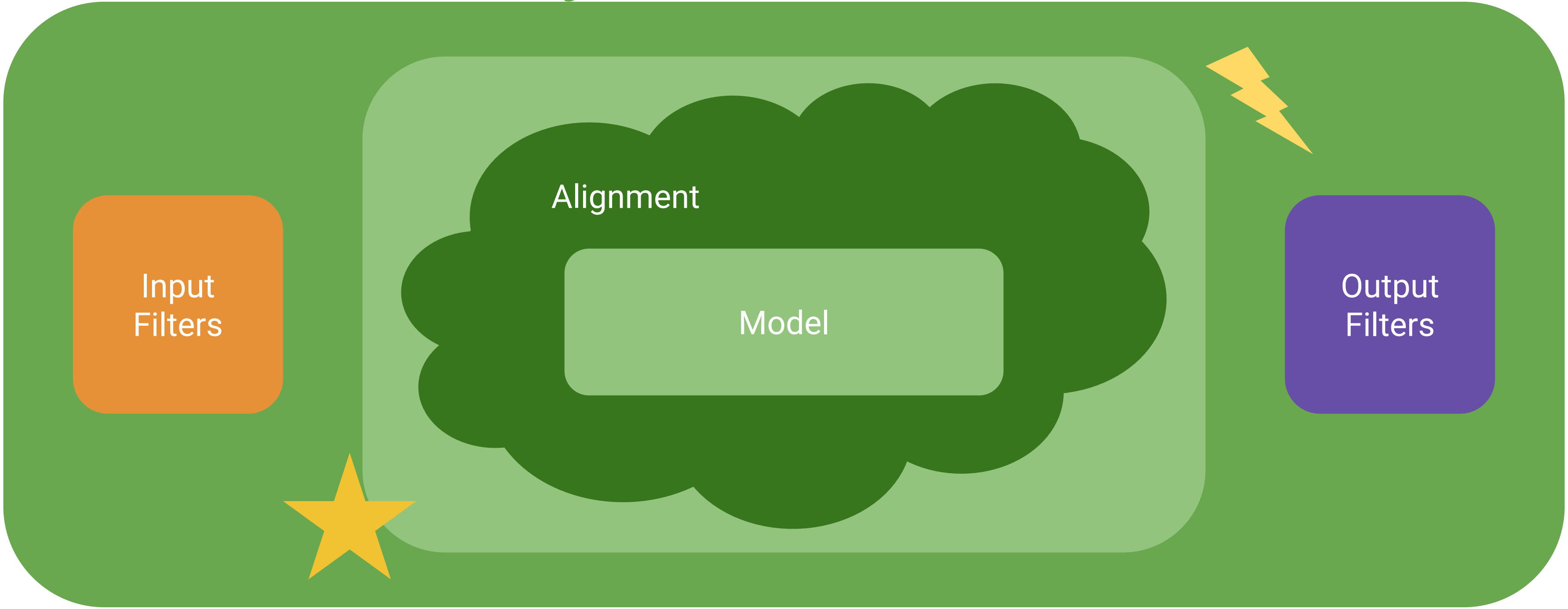


3.0%
2.0%
1.0%
0.0%
% of tokens memorized
Alignment: 50× less
Pythia 1.4
Pythia 6.9
LLaMA-65B
InstructGPT
ChatGPT
(default)


3.0%
2.0%
1.0%
0.0%
% of tokens memorized
Alignment: 50× less
Attack: 150× more
Pythia 1.4
Pythia 6.9
LLaMA-65B
InstructGPT
ChatGPT (default)
ChatGPT (our attack)

Model

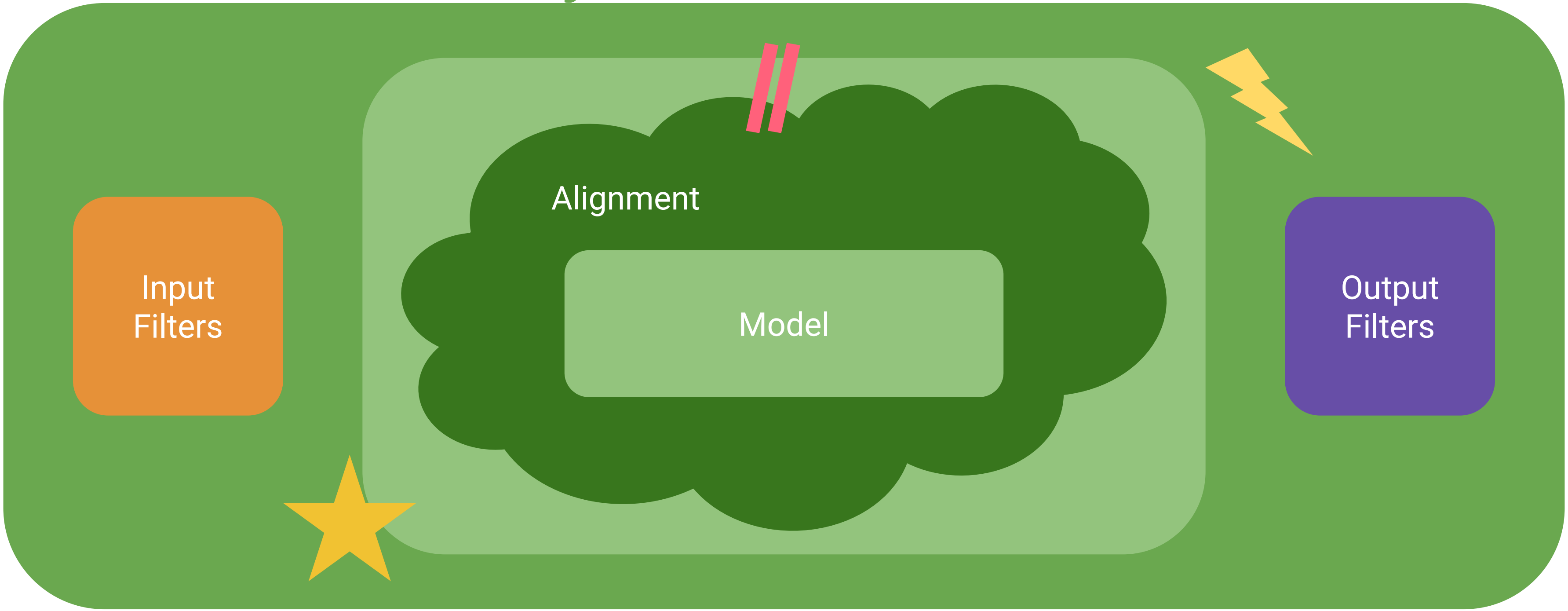
Generative AI System



Metric:

How much of the training data can the model reproduce?

Generative AI System



What can we generalize?

methods and definitions

What can we not?

interpretations

Memorization of:

Facts
Examples from the training data
Style
Parts of examples

Memorization of:

Facts

Examples from the training data

Style

Parts of examples

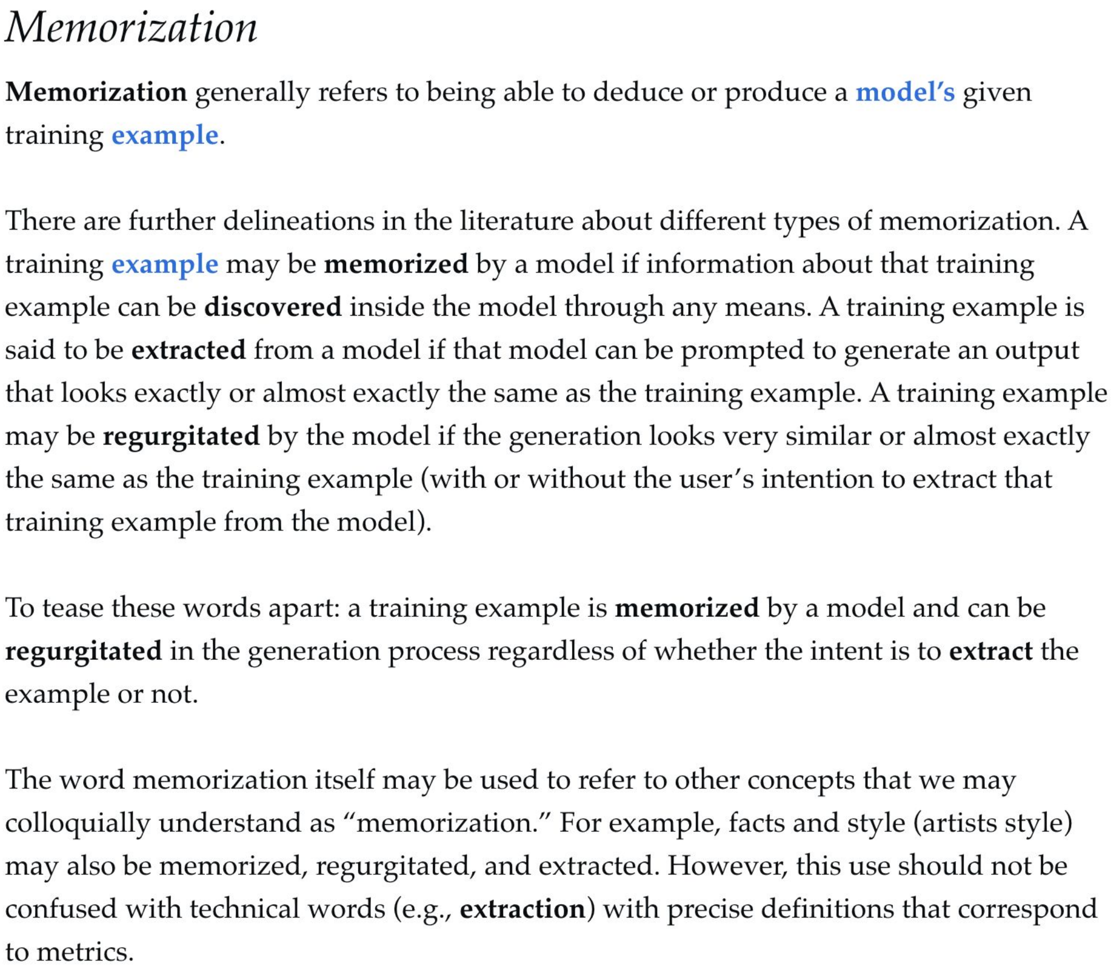
Memorization

Memorization generally refers to being able to deduce or produce a **model's** given training **example**.

There are further delineations in the literature about different types of memorization. A training **example** may be **memorized** by a model if information about that training example can be **discovered** inside the model through any means. A training example is said to be **extracted** from a model if that model can be prompted to generate an output that looks exactly or almost exactly the same as the training example. A training example may be **regurgitated** by the model if the generation looks very similar or almost exactly the same as the training example (with or without the user's intention to extract that training example from the model).

To tease these words apart: a training example is **memorized** by a model and can be **regurgitated** in the generation process regardless of whether the intent is to **extract** the example or not.

The word memorization itself may be used to refer to other concepts that we may colloquially understand as "memorization." For example, facts and style (artists style) may also be memorized, regurgitated, and extracted. However, this use should not be confused with technical words (e.g., **extraction**) with precise definitions that correspond to metrics.



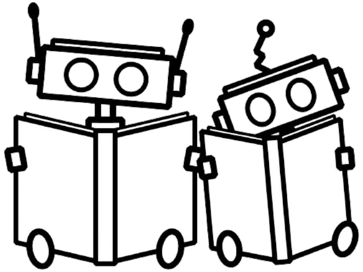
Memorization

Memorization generally refers to being able to deduce or produce a **model's** given training **example**.

There are further delineations in the literature about different types of memorization. A training **example** may be **memorized** by a model if information about that training example can be **discovered** inside the model through any means. A training example is said to be **extracted** from a model if that model can be prompted to generate an output that looks exactly or almost exactly the same as the training example. A training example may be **regurgitated** by the model if the generation looks very similar or almost exactly the same as the training example (with or without the user's intention to extract that training example from the model).

To tease these words apart: a training example is **memorized** by a model and can be **regurgitated** in the generation process regardless of whether the intent is to **extract** the example or not.

The word memorization itself may be used to refer to other concepts that we may colloquially understand as "memorization." For example, facts and style (artists style) may also be memorized, regurgitated, and extracted. However, this use should not be confused with technical words (e.g., **extraction**) with precise definitions that correspond to metrics.



genlaw.org/glossary.html

What can we generalize?

methods and definitions

What can we not?

interpretations

Memorization is

neither good or bad

Harm from memorization is contextual

Common phrases: “To whom it may concern...”

Facts: “Christmas is celebrated on Dec 25th”

Harm from memorization is contextual

Common phrases: “To whom it may concern...”

Facts: “Christmas is celebrated on Dec 25th”

Private / sensitive information: “My social security number is XXXXX.”

Harm from memorization is contextual

Common phrases: “To whom it may concern...”

Facts: “Christmas is celebrated on Dec 25th”

Private / sensitive information: “My social security number is XXXXX.”

Quotes: “Trump said, ‘Tariffs are the greatest!’”

Quotes: “Sally Smith said, ‘Sam is the worst.’”

!! What do we do?!!

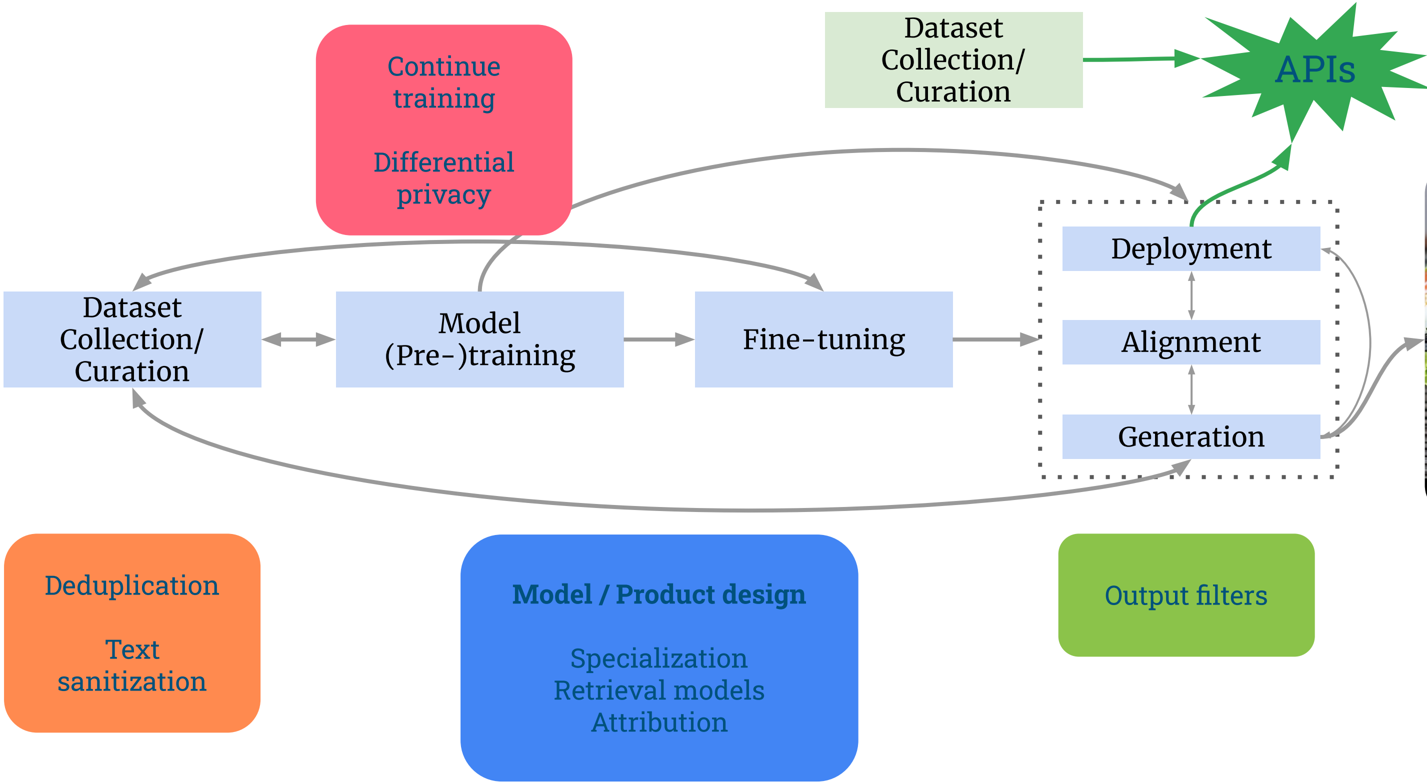
Red teaming...

the system

AND

its components

Mitigations are all across the supply chain



Continue training

Differential privacy

Dataset Collection/ Curation

APIs

Dataset Collection/ Curation

Model (Pre-)training

Fine-tuning

Deployment

Alignment

Generation

Deduplication

Text sanitization

Model / Product design

Specialization
Retrieval models
Attribution

Output filters

A cute corgi lives in a house made out of sushi.

Disclosure

Thank you

Methods and definitions are generalizable, interpretations are not

Red team the system and the components

Disclosure systems are sorely needed

Katherine Lee

Nicholas Carlini

Daphne Ippolito

Milad Nasr

Matthew Jagielski

Chris Choquette

Chiyuan Zhang

Florian Tramèr

James Grimmelmann

A. Feder Cooper

Jon Hayase

Erics Wallace