Talkin' 'Bout AI Generation

Lessons for AI Policy, Research, and Practice

Katherine Lee

Attack GenAI, Team Lead
Google DeepMind

A. Feder Cooper

Microsoft Research
The GenLaw Center
Stanford

Goals:

Goals:

Frameworks for thinking about generative AI

Goals:

Frameworks for thinking about generative AI

Generative AI primer
Machine unlearning

Questions?

Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain

Forthcoming, Journal of the Copyright Society 2024

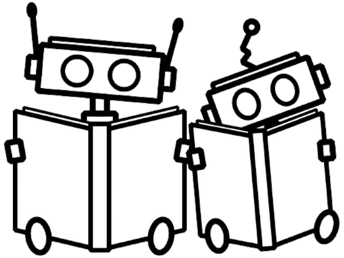
149 Pages • Posted: 1 Aug 2023 • Last revised: 4 Mar 2024

Katherine Lee
Cornell University - Department of Computer Science

A. Feder Cooper
Microsoft Research; Stanford University; Yale University

James Grimmelmann
Cornell Law School; Cornell Tech

Date Written: July 27, 2023



What is Generative AI?

Generative AI

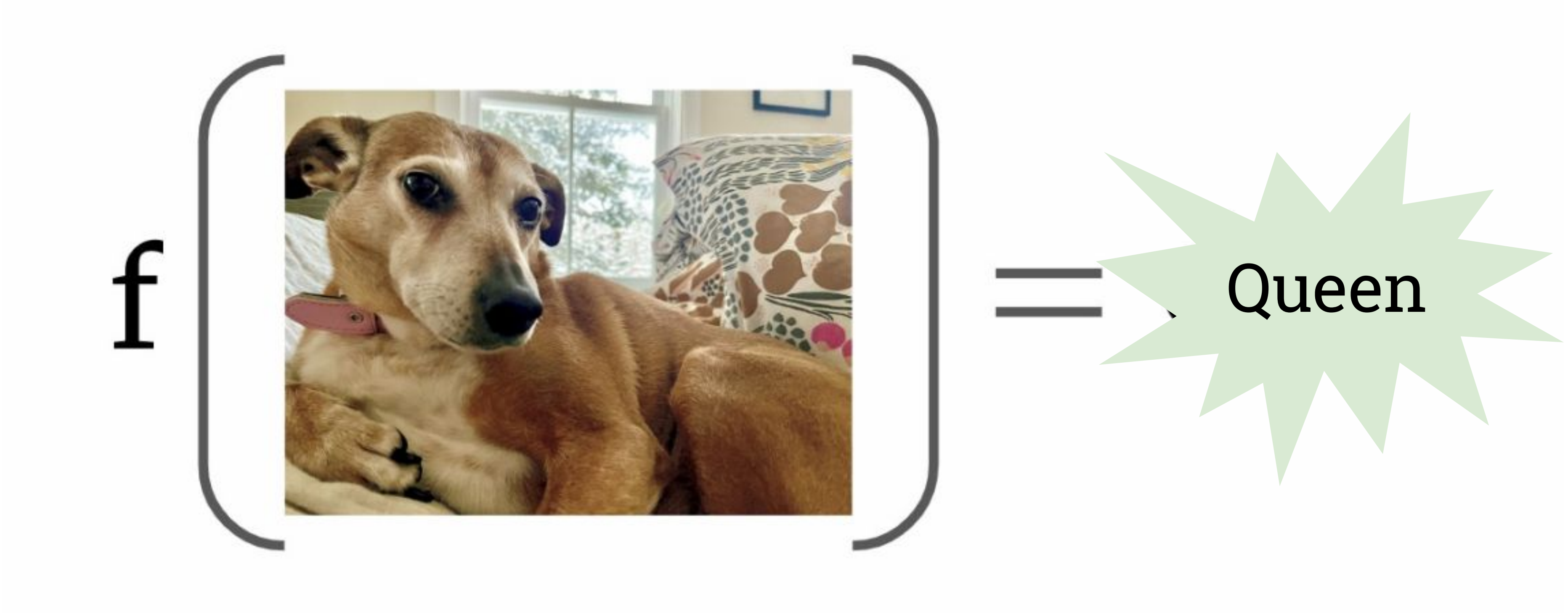
is not like other AI ✨

Generative vs. Discriminative

Discriminative

f () = dog

Discriminative



Generative

f ("cat in a red and white striped hat") =

f ("cat in a red and white striped hat") =

⋮

f ("cat in a red and white striped hat") =

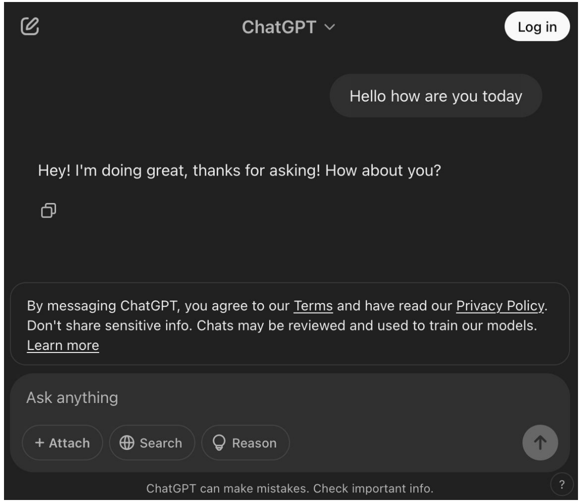
Generative AI

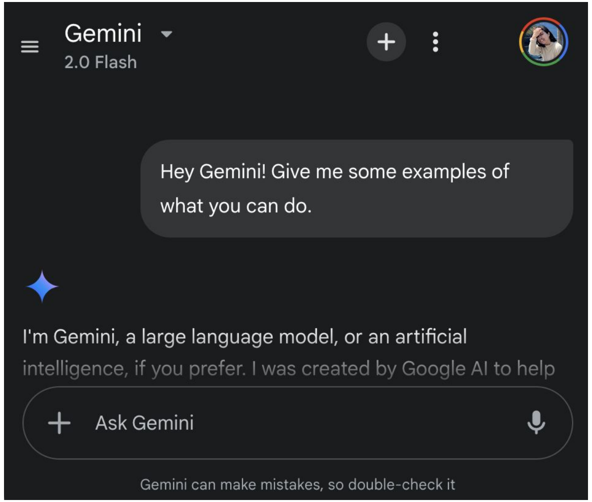
is not just the AI

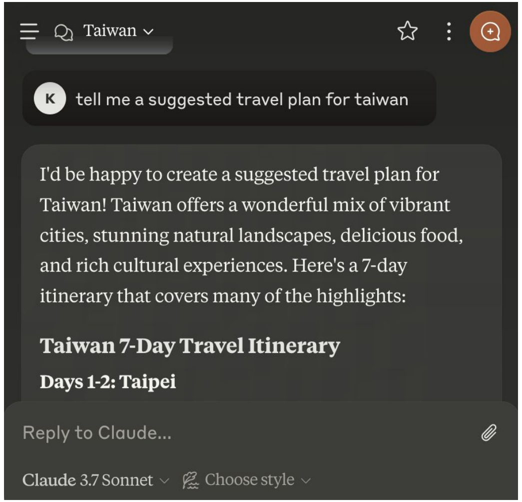
It’s the system

(and the models)

User Interface







API

Python Node.js REST

```python
from google import genai

client = genai.Client(api_key="YOUR_API_KEY")

response = client.models.generate_content(
    model="gemini-2.0-flash",
    contents="Explain how AI works",
)

print(response.text)
```

API

```python
completion = client.chat.completions.create(
    model="gpt-4o",
    messages=[{"role": "user", "content": "What is the weather like in Paris today?"}]
    tools=tools
)

print(completion.choices[0].message.tool_calls)
```

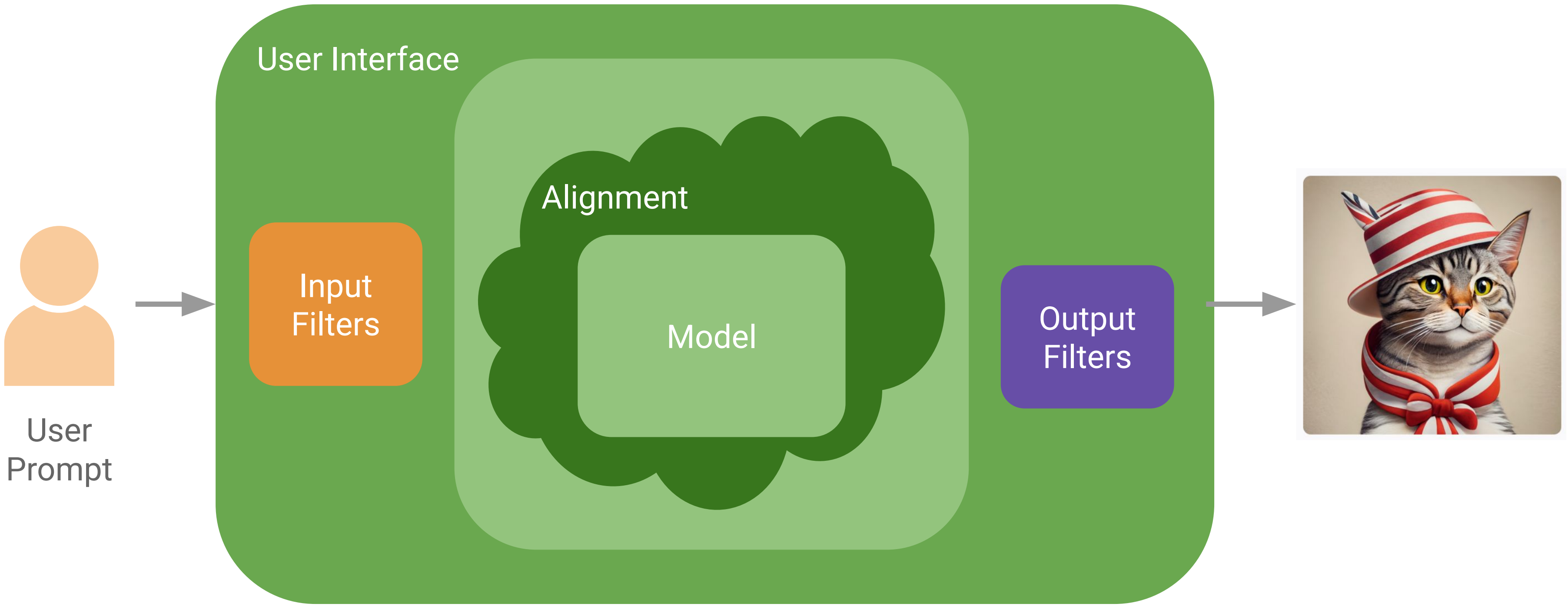
Python Node.js REST

```python
from google import genai

client = genai.Client(api_key="YOUR_API_KEY")

response = client.models.generate_content(
    model="gemini-2.0-flash",
    contents="Explain how AI works",
)

print(response.text)
```

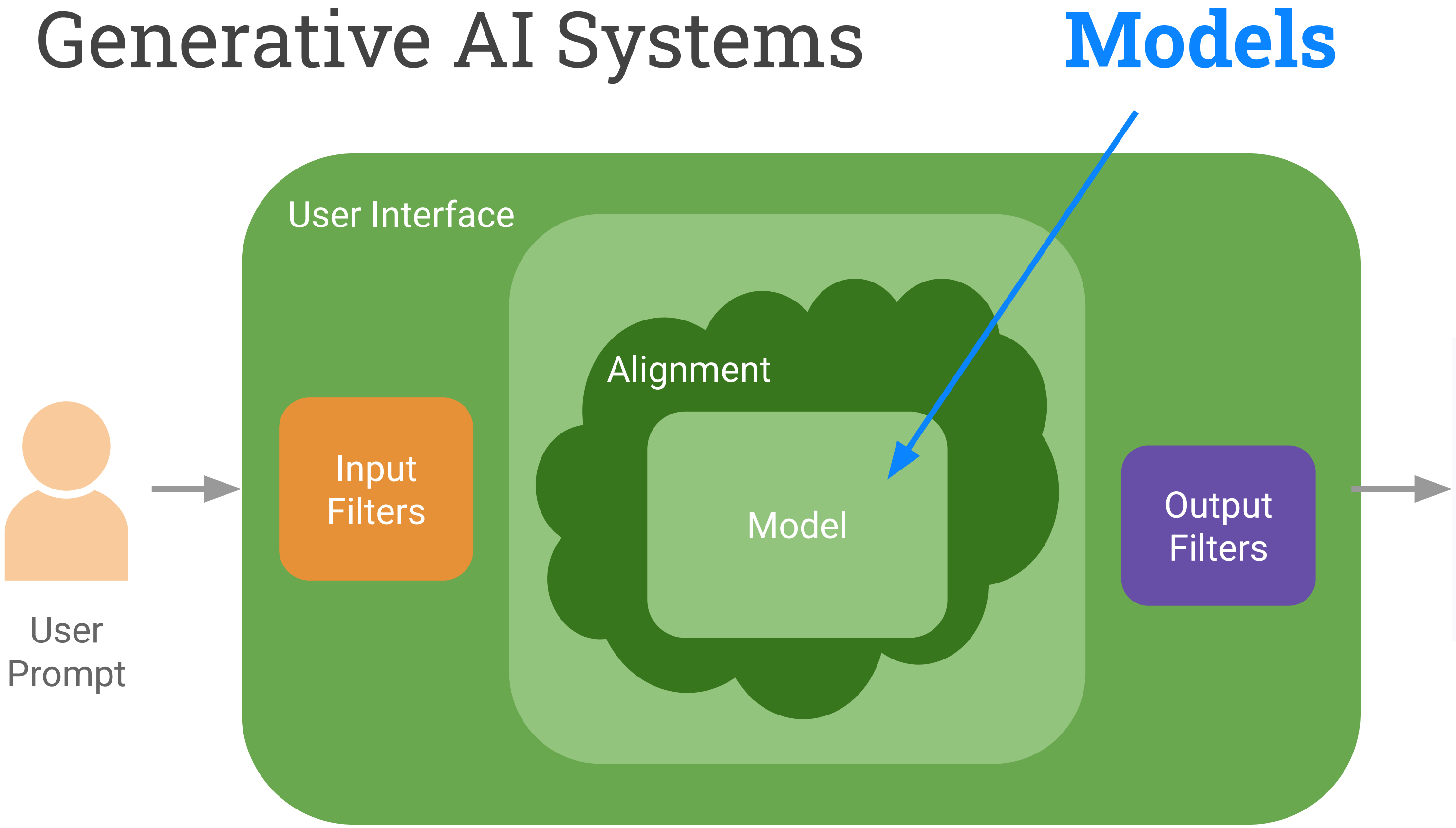
Python TypeScript

```python
import anthropic

client = anthropic.Anthropic()

message = client.messages.create(
    model="claude-3-7-sonnet-20250219",
    max_tokens=1000,
    temperature=1,
    system="You are a world-class poet. Respond only with short poems.",
    messages=[
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": "Why is the ocean salty?"
                }
            ]
        }
    ]
)
print(message.content)
```

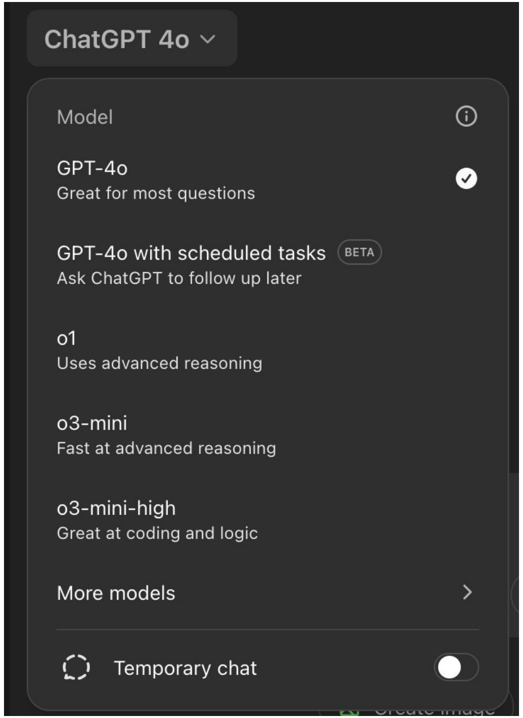
Generative AI Systems

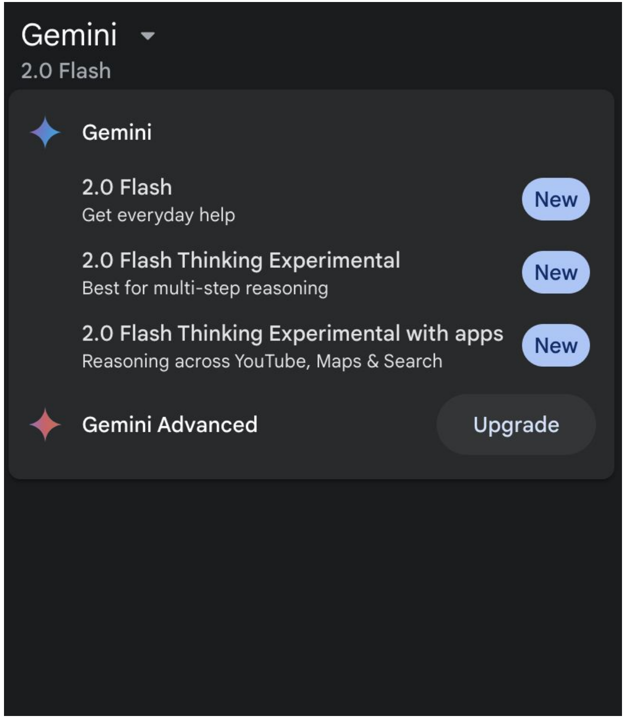


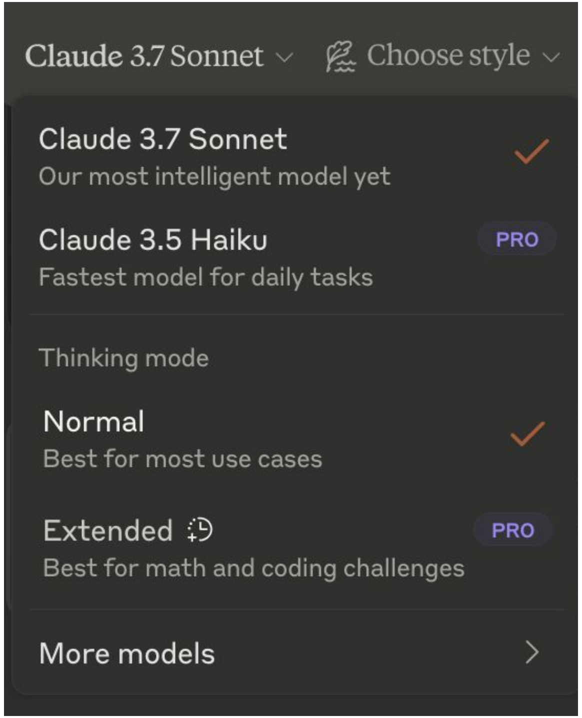
Generative AI Systems



Systems have different underlying models



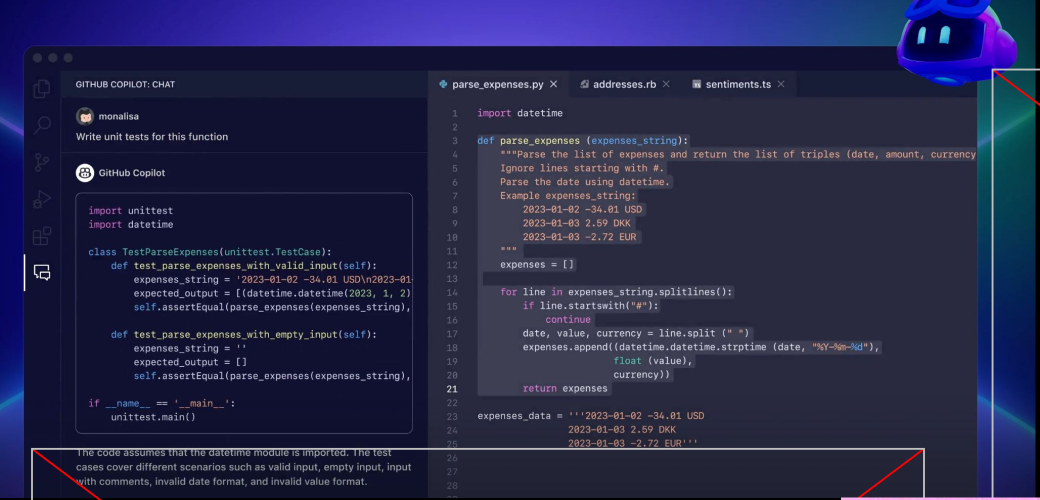




Questions?

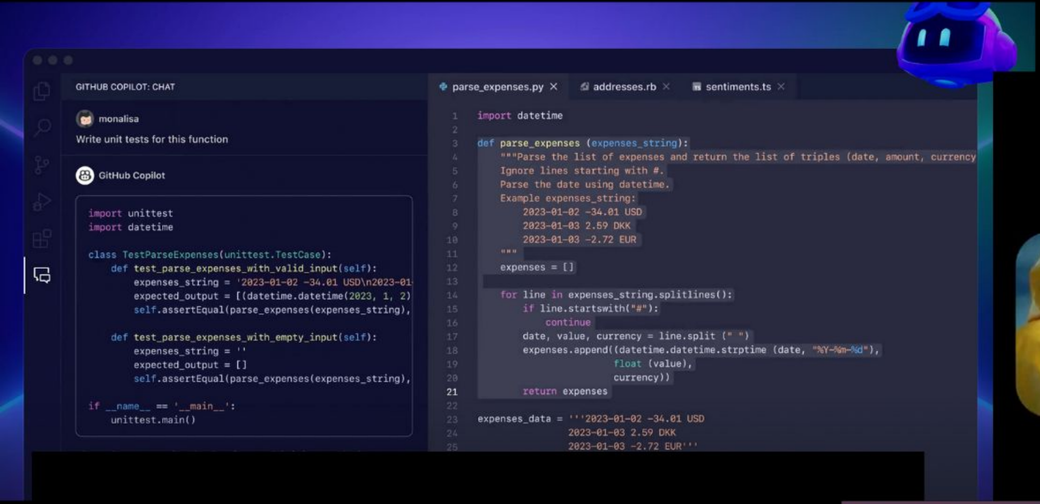
Generative AI

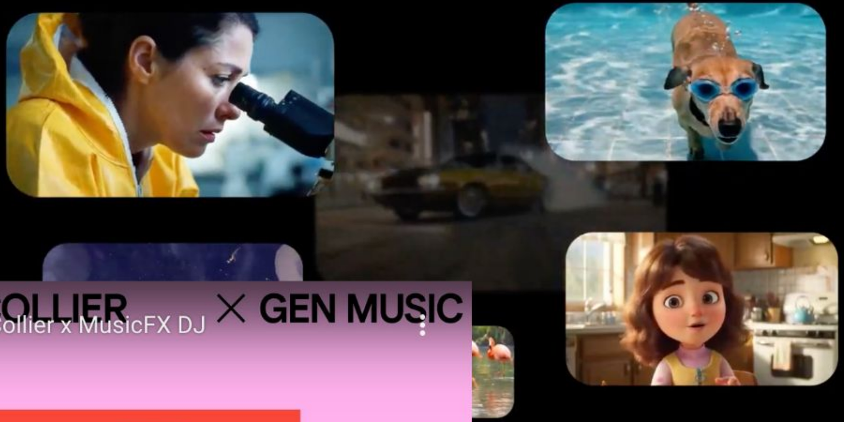
is used for many modalities


GITHUB COPILOT: CHAT
parse_expenses.py
addresses.rb
sentiments.ts
monalisa
Write unit tests for this function
GitHub Copilot
import unittest
import datetime
class TestParseExpenses(unittest.TestCase):
def test_parse_expenses_with_valid_input(self):
expenses_string = '2023-01-02 -34.01 USD\n2023-01-
expected_output = [(datetime.datetime(2023, 1, 2)
self.assertEqual(parse_expenses(expenses_string),
def test_parse_expenses_with_empty_input(self):
expenses_string = ''
expected_output = []
self.assertEqual(parse_expenses(expenses_string),
if __name__ == '__main__':
unittest.main()
The code assumes that the datetime module is imported. The test cases cover different scenarios such as valid input, empty input, input with comments, invalid date format, and invalid value format.
import datetime
def parse_expenses (expenses_string):
"""Parse the list of expenses and return the list of triples (date, amount, currency
Ignore lines starting with #.
Parse the date using datetime.
Example expenses_string:
2023-01-02 -34.01 USD
2023-01-03 2.59 DKK
2023-01-03 -2.72 EUR
"""
expenses = []
for line in expenses_string.splitlines():
if line.startswith("#"):
continue
date, value, currency = line.split (" ")
expenses.append((datetime.datetime.strptime (date, "%Y-%m-%d"),
float (value),
currency))
return expenses
expenses_data = '''2023-01-02 -34.01 USD
2023-01-03 2.59 DKK
2023-01-03 -2.72 EUR'''


JACOB COLLIER
X GEN MUSIC

NotebookLM


GITHUB COPILOT: CHAT
monalisa
Write unit tests for this function
GitHub Copilot
import unittest
import datetime
class TestParseExpenses(unittest.TestCase):
def test_parse_expenses_with_valid_input(self):
expenses_string = '2023-01-02 -34.01 USD\n2023-01
expected_output = [(datetime.datetime(2023, 1, 2)
self.assertEqual(parse_expenses(expenses_string),
def test_parse_expenses_with_empty_input(self):
expenses_string = ''
expected_output = []
self.assertEqual(parse_expenses(expenses_string),
if __name__ == '__main__':
unittest.main()
parse_expenses.py
addresses.rb
sentiments.ts
import datetime
def parse_expenses (expenses_string):
"""Parse the list of expenses and return the list of triples (date, amount, currency
Ignore lines starting with #.
Parse the date using datetime.
Example expenses_string:
2023-01-02 -34.01 USD
2023-01-03 2.59 DKK
2023-01-03 -2.72 EUR
"""
expenses = []
for line in expenses_string.splitlines():
if line.startswith("#"):
continue
date, value, currency = line.split (" ")
expenses.append((datetime.datetime.strptime (date, "%Y-%m-%d"),
float (value),
currency))
return expenses
expenses_data = '''2023-01-02 -34.01 USD
2023-01-03 2.59 DKK
2023-01-03 -2.72 EUR'''




COB COLLIER × GEN MUSIC
Jacob Collier x MusicFX DJ

NotebookLM

Different modalities are at different scales:

model size, compute, datasets, ...

There are many types of Generative AI models

Architecture: Transformer, Diffusion-based

Aside:

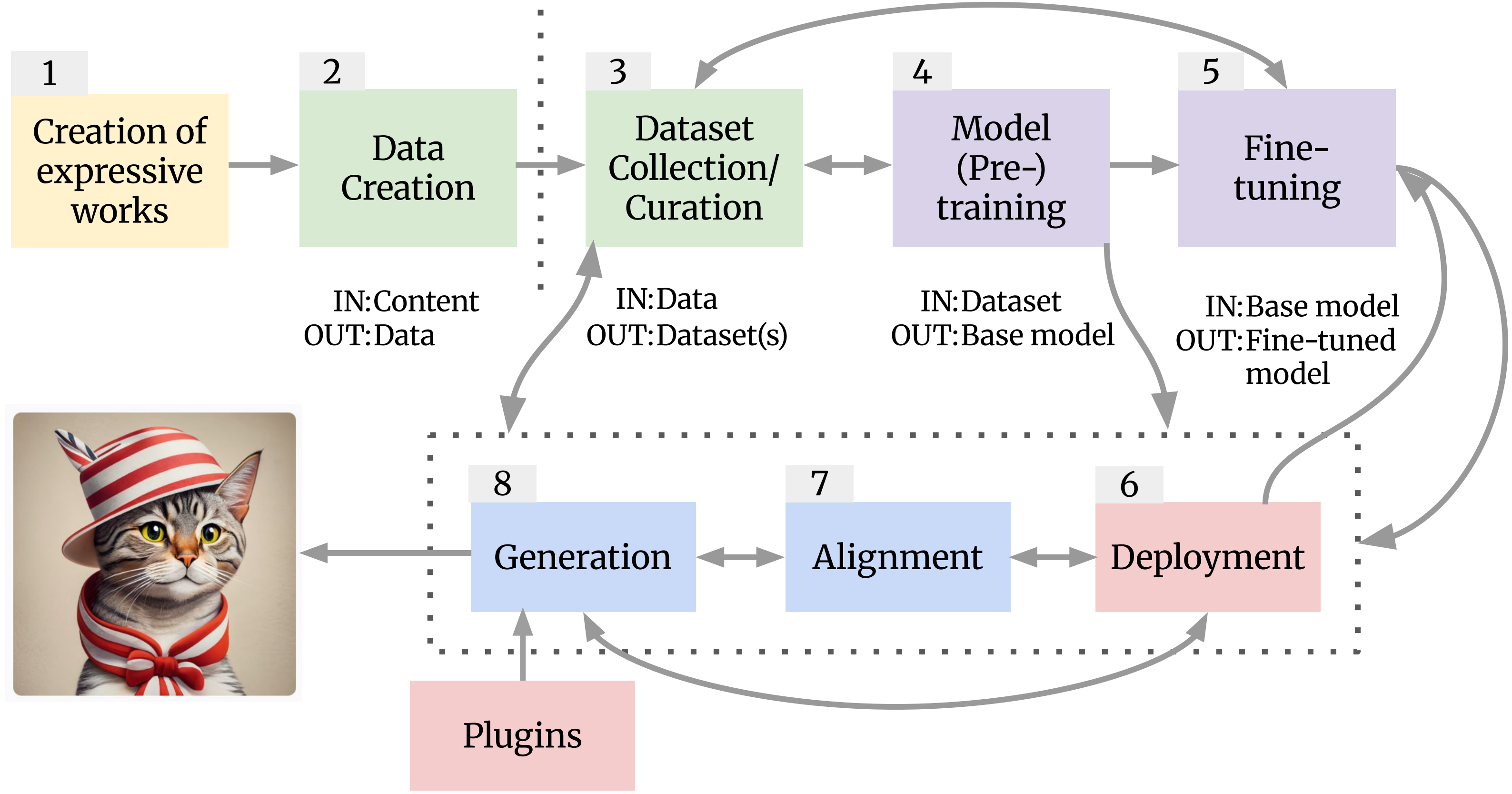
the language is muddy

Model → system, architecture, checkpoint, multiple models, …

Questions?

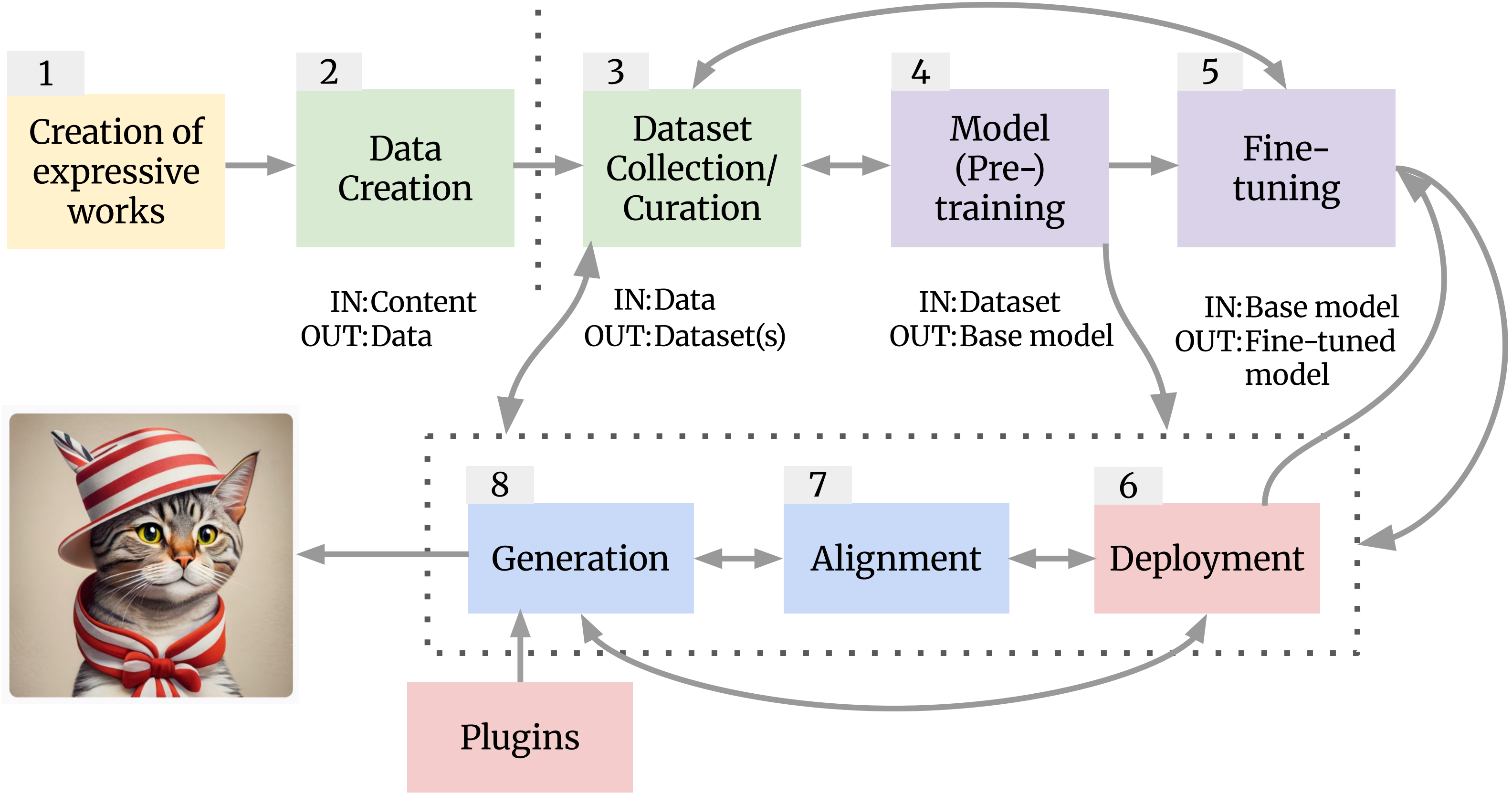
Generative AI Supply Chain

The generative-AI supply chain

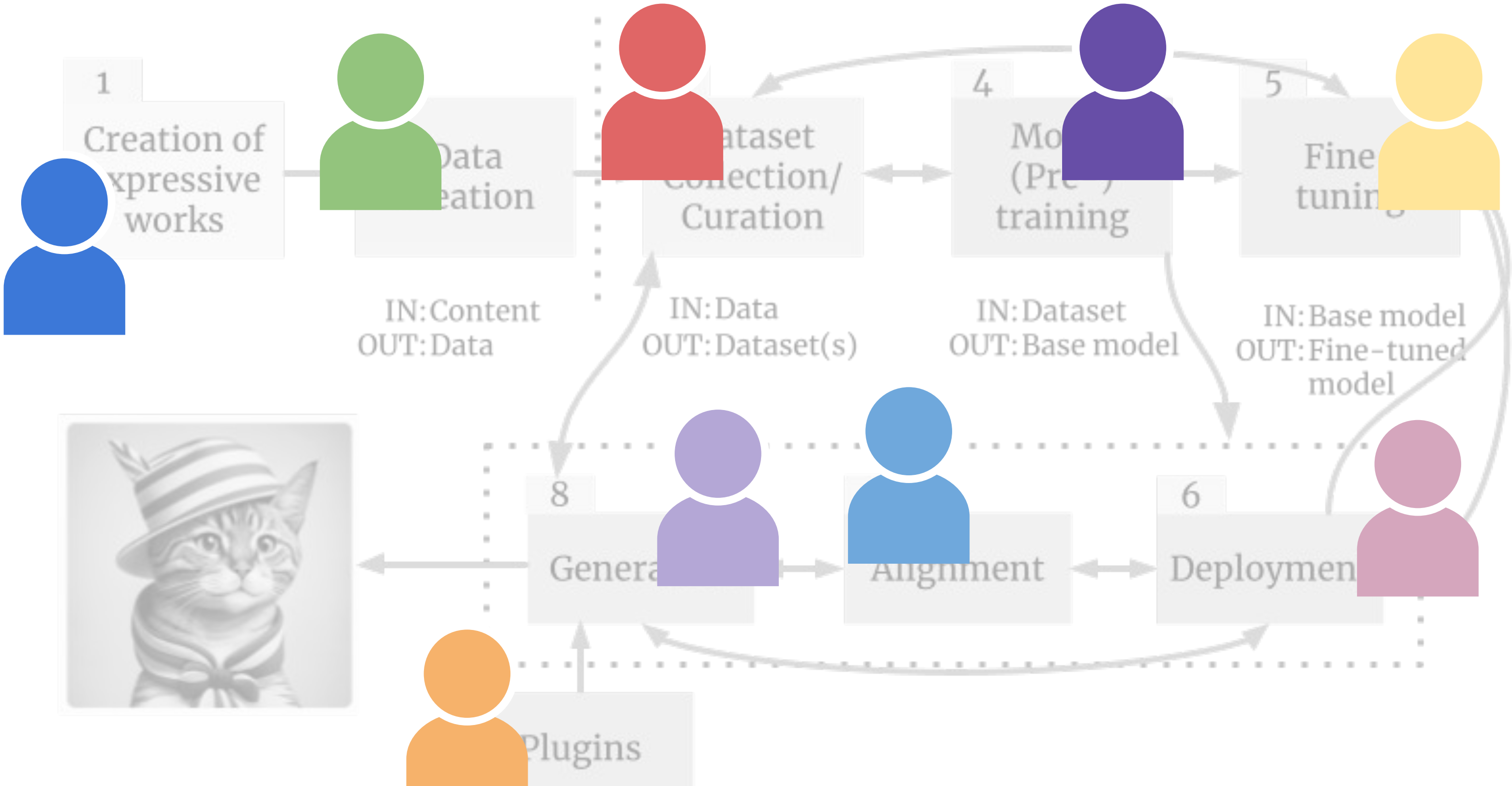


This started out as a short piece on retrieval augmented generation in 2022 🥲

The generative-AI supply chain



There are a lot of different actors



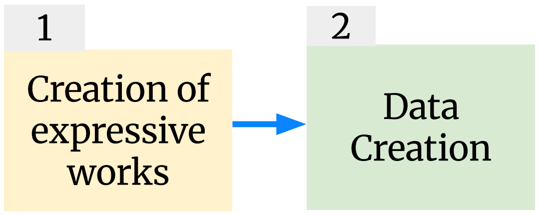
Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

The generative-AI supply chain

1

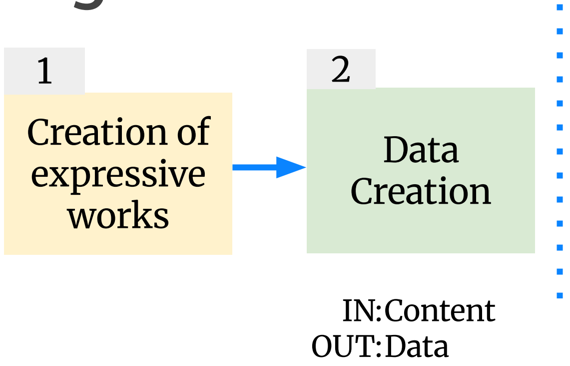
Creation of expressive works

The generative-AI supply chain



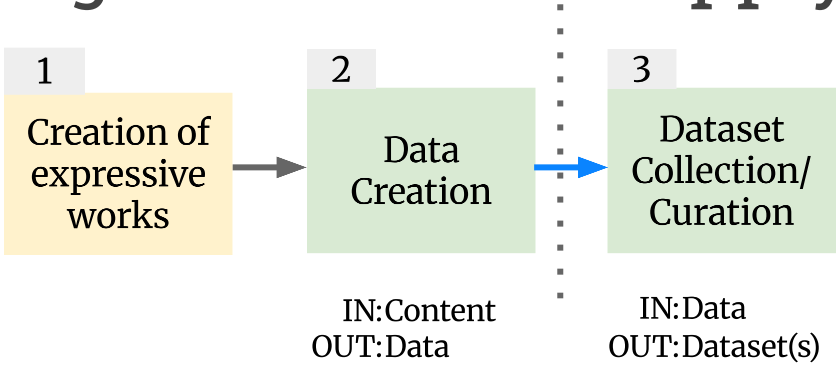
IN:Content
OUT:Data

The generative-AI supply chain



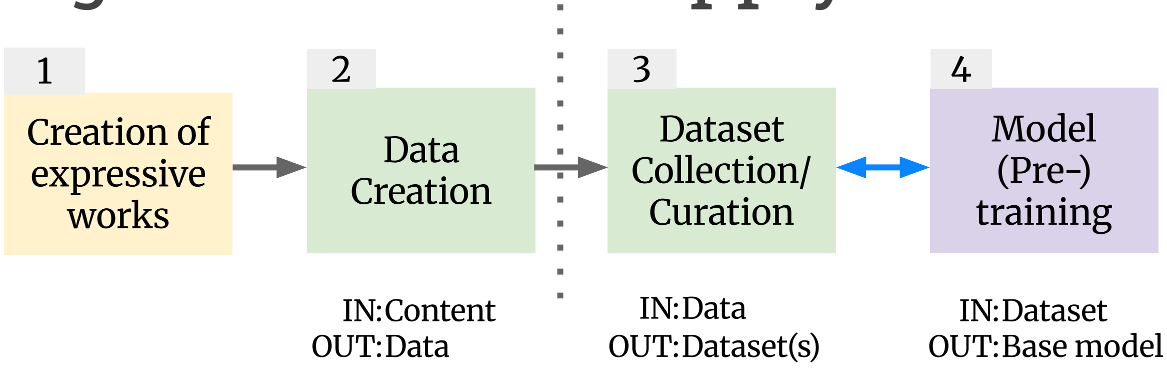
Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

The generative-AI supply chain



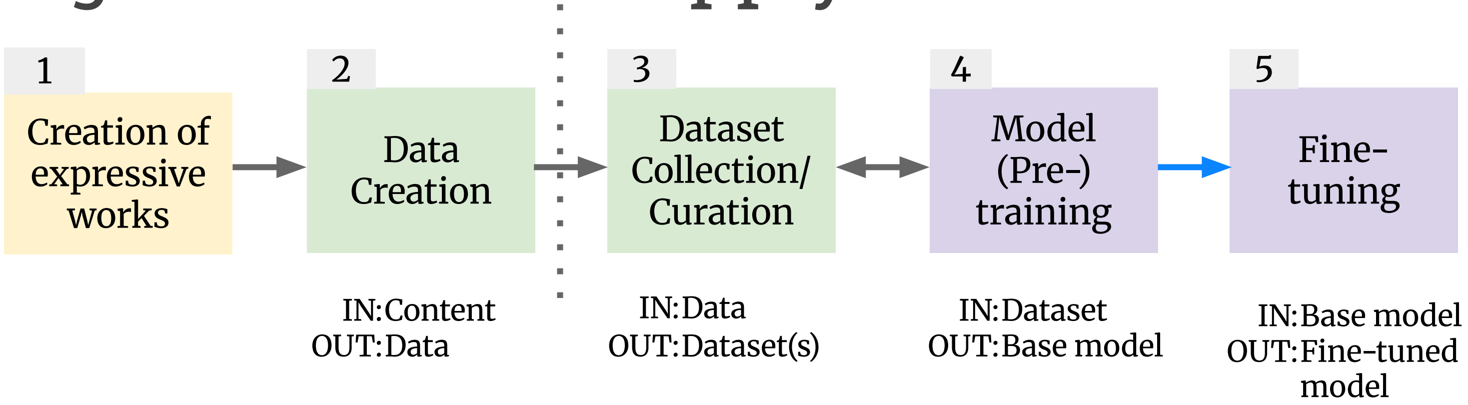
Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

The generative-AI supply chain



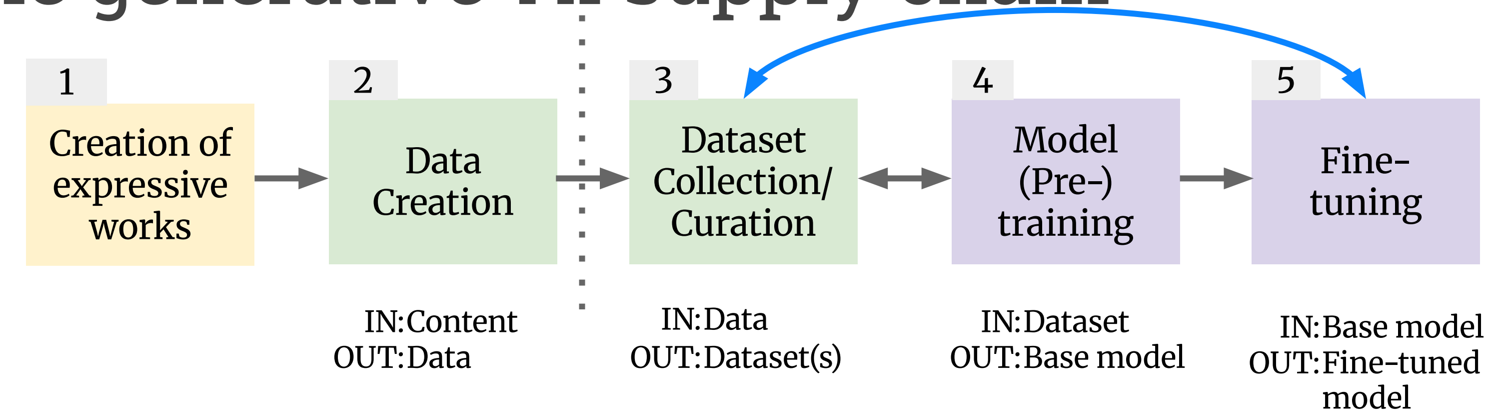
Lee*, Cooper* & Grimmelmann*. "Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain." 2023

The generative-AI supply chain

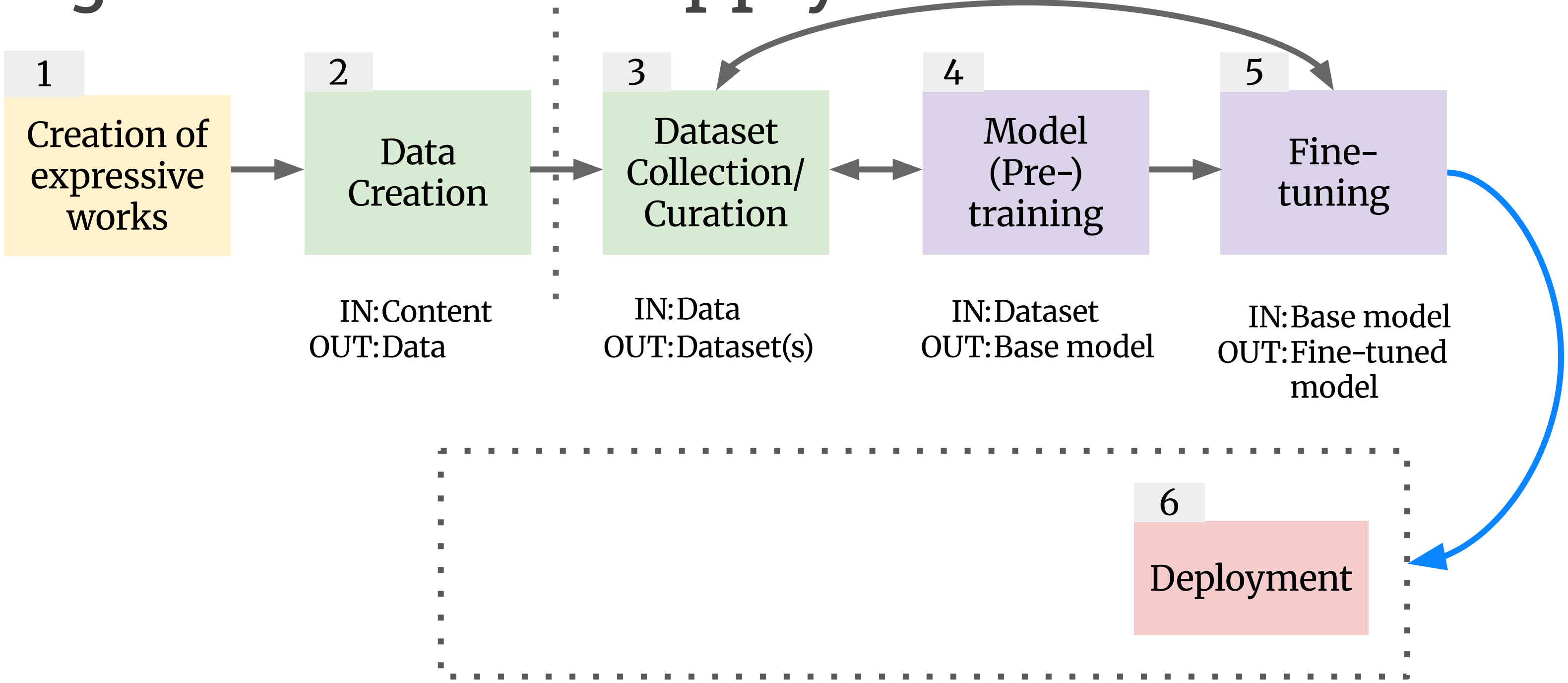


Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

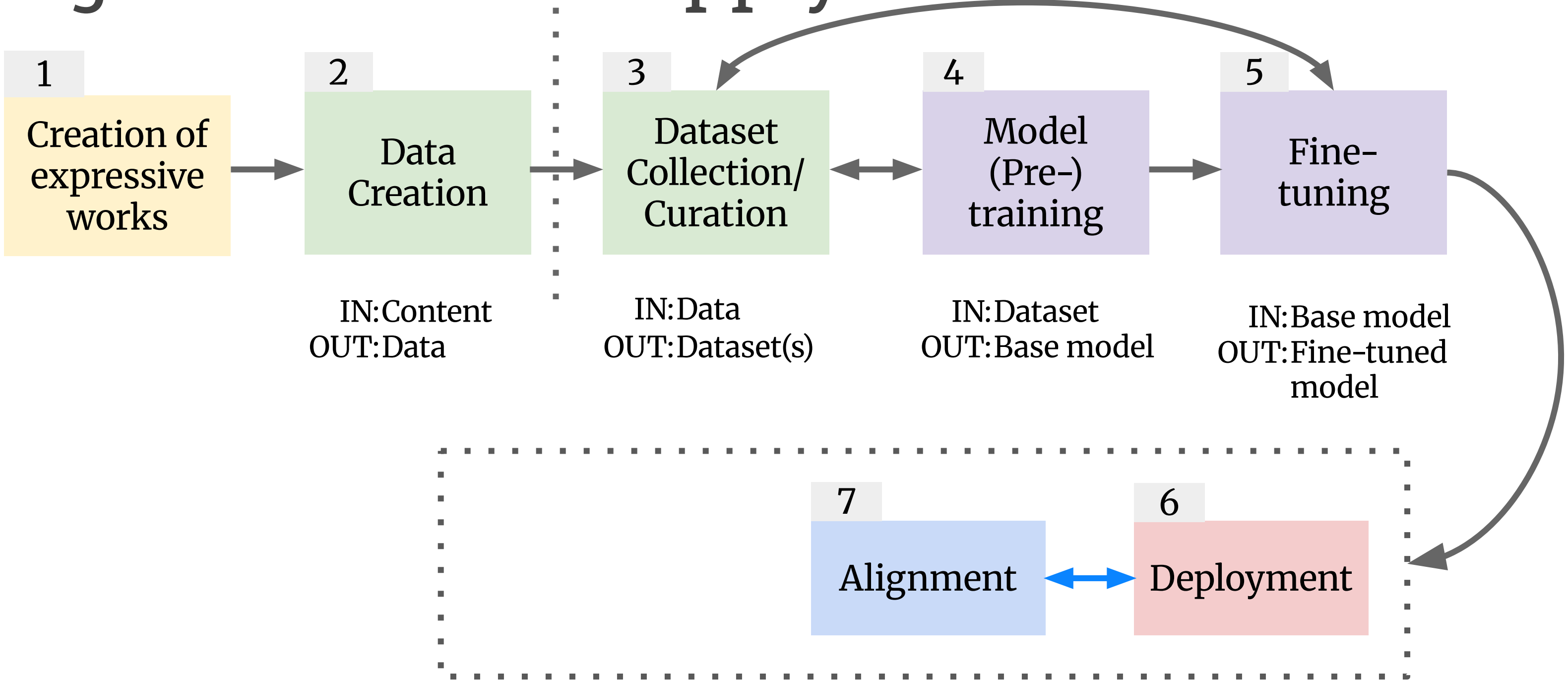
The generative-AI supply chain



The generative-AI supply chain

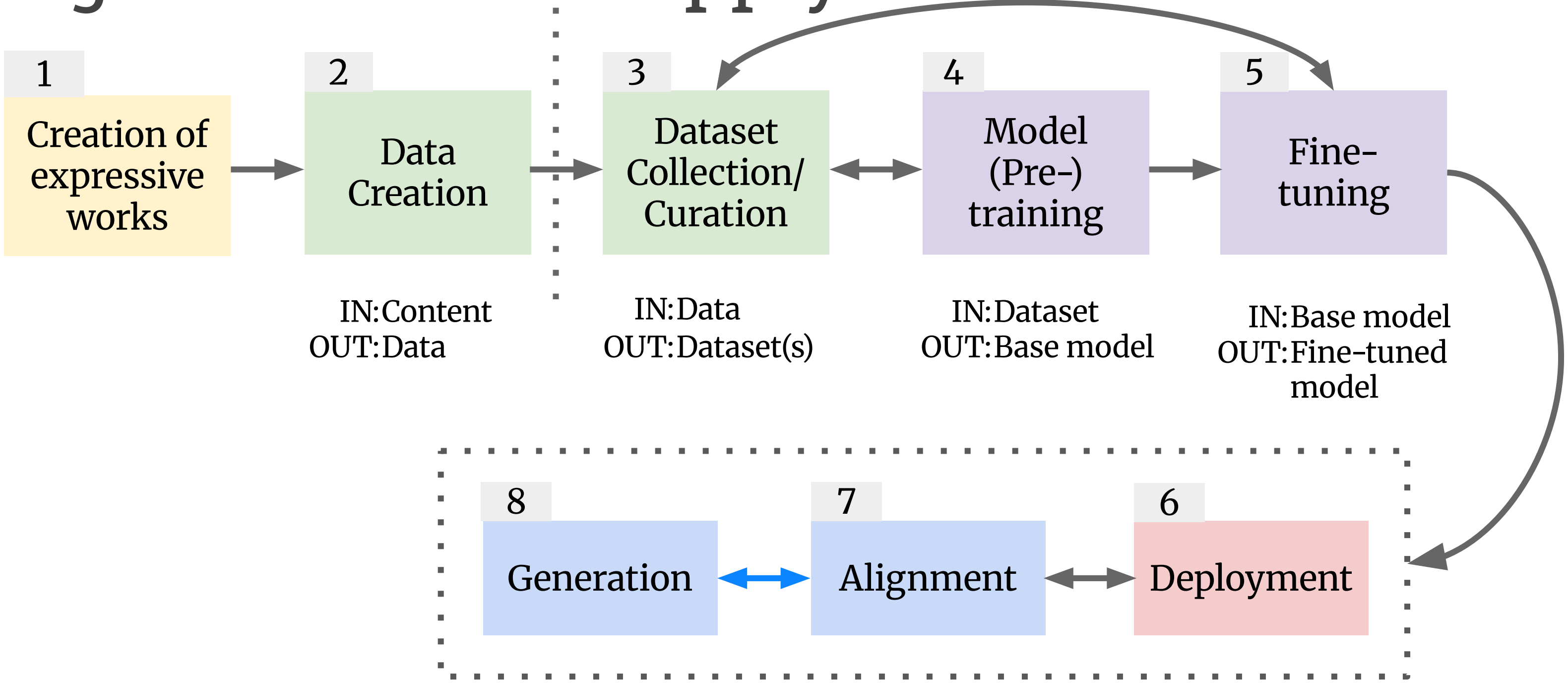


The generative-AI supply chain



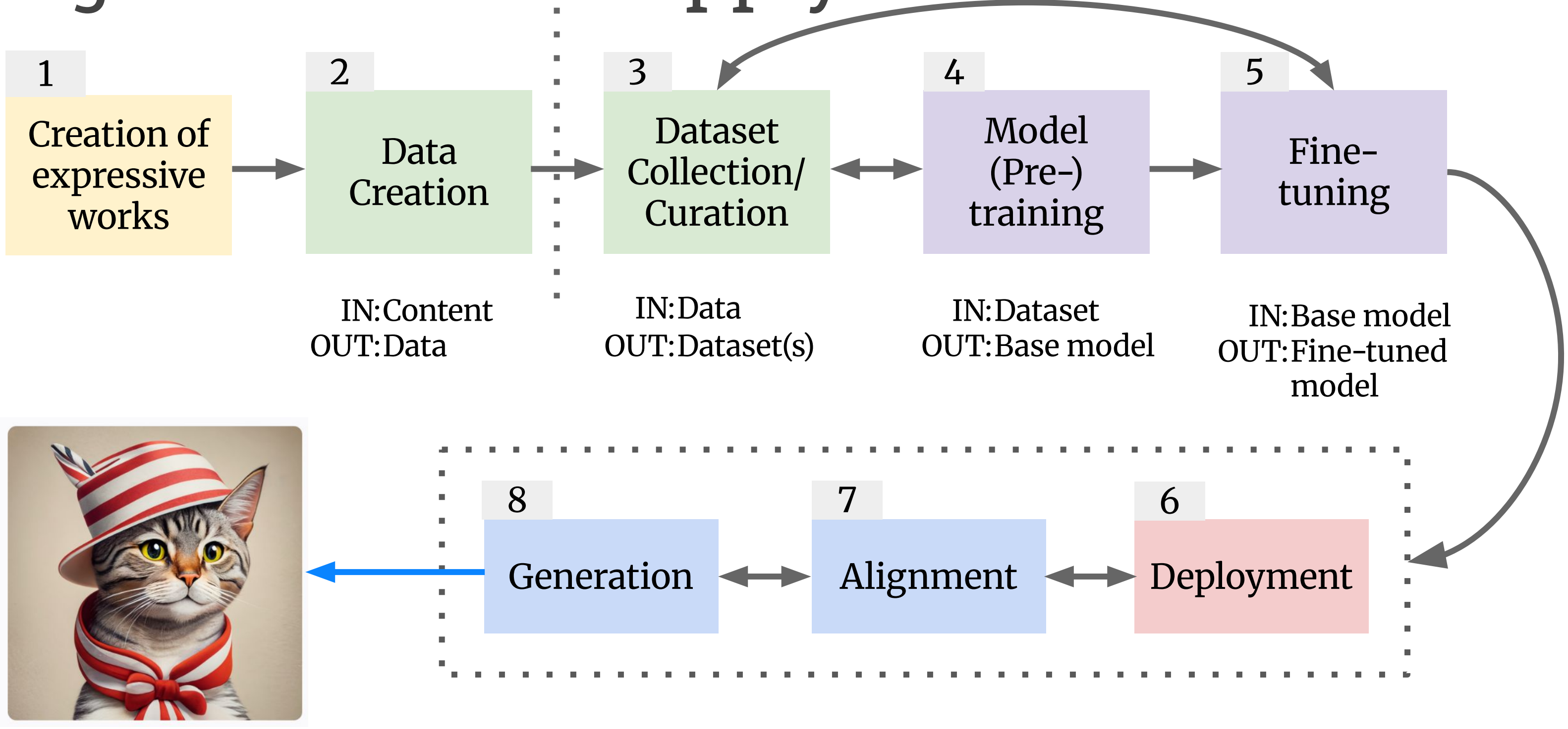
Lee*, Cooper* & Grimmelmann*. "Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain." 2023

The generative-AI supply chain



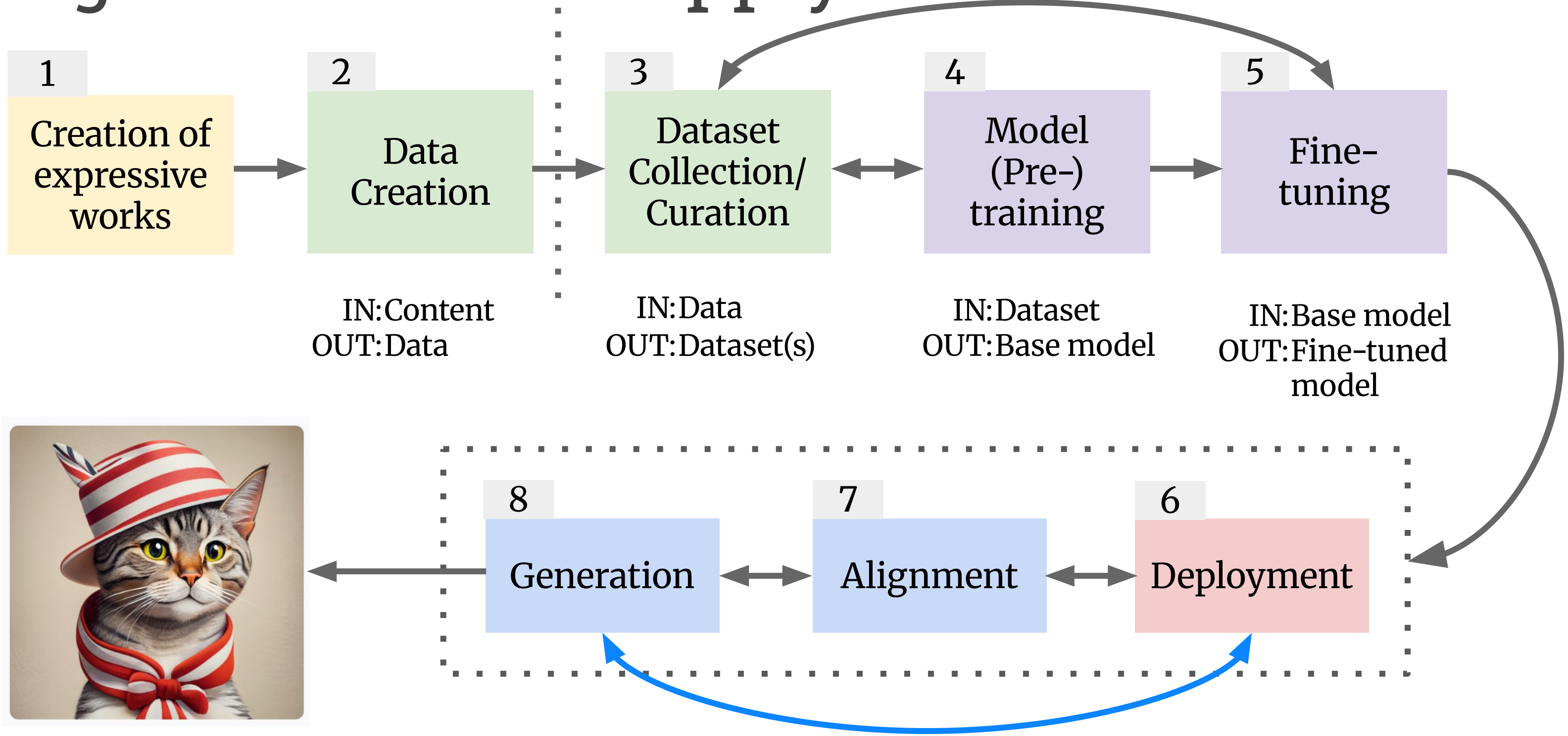
Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

The generative-AI supply chain



Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

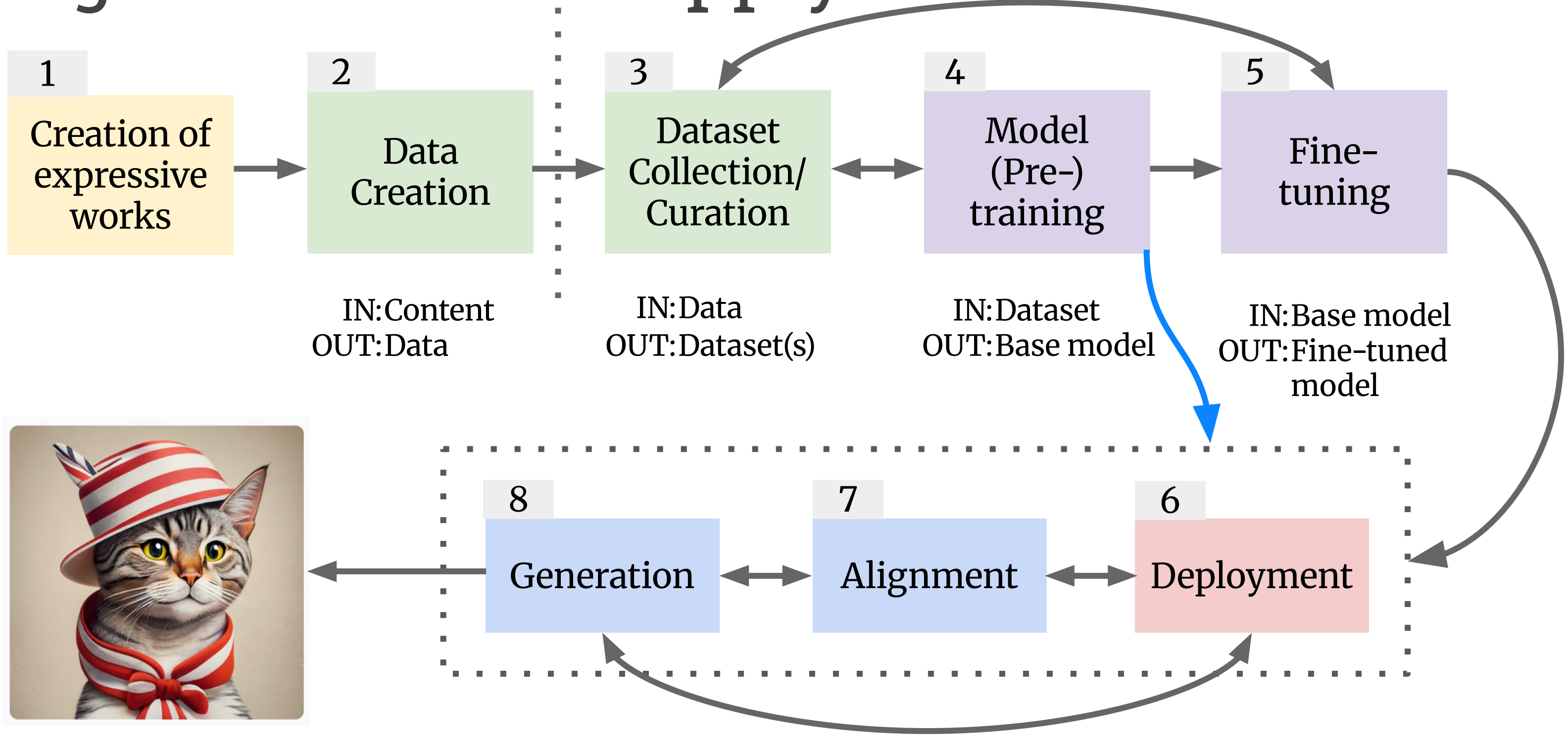
The generative-AI supply chain



Lee*, Cooper* & Grimmelmann*. "Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain." 2023

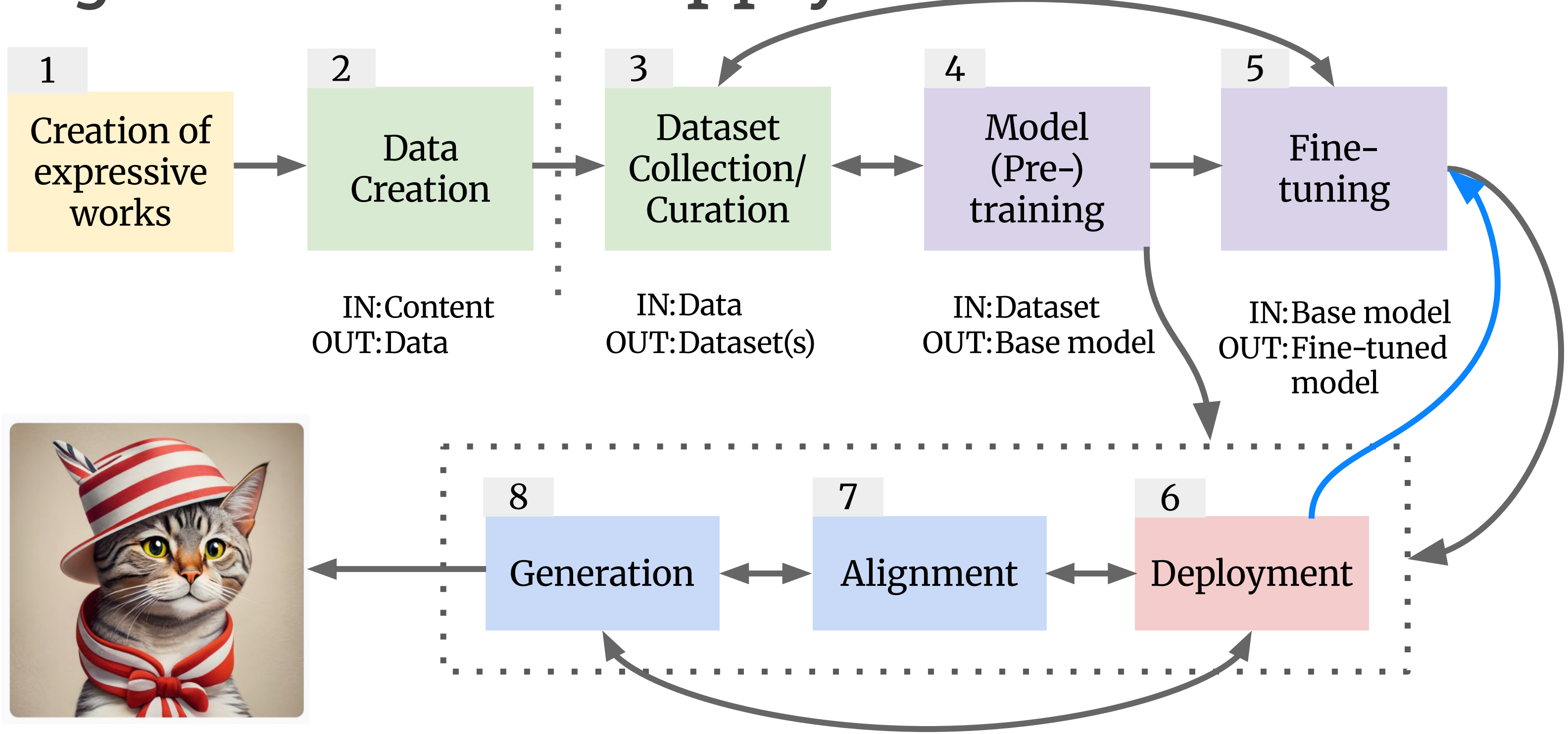
Questions?

The generative-AI supply chain



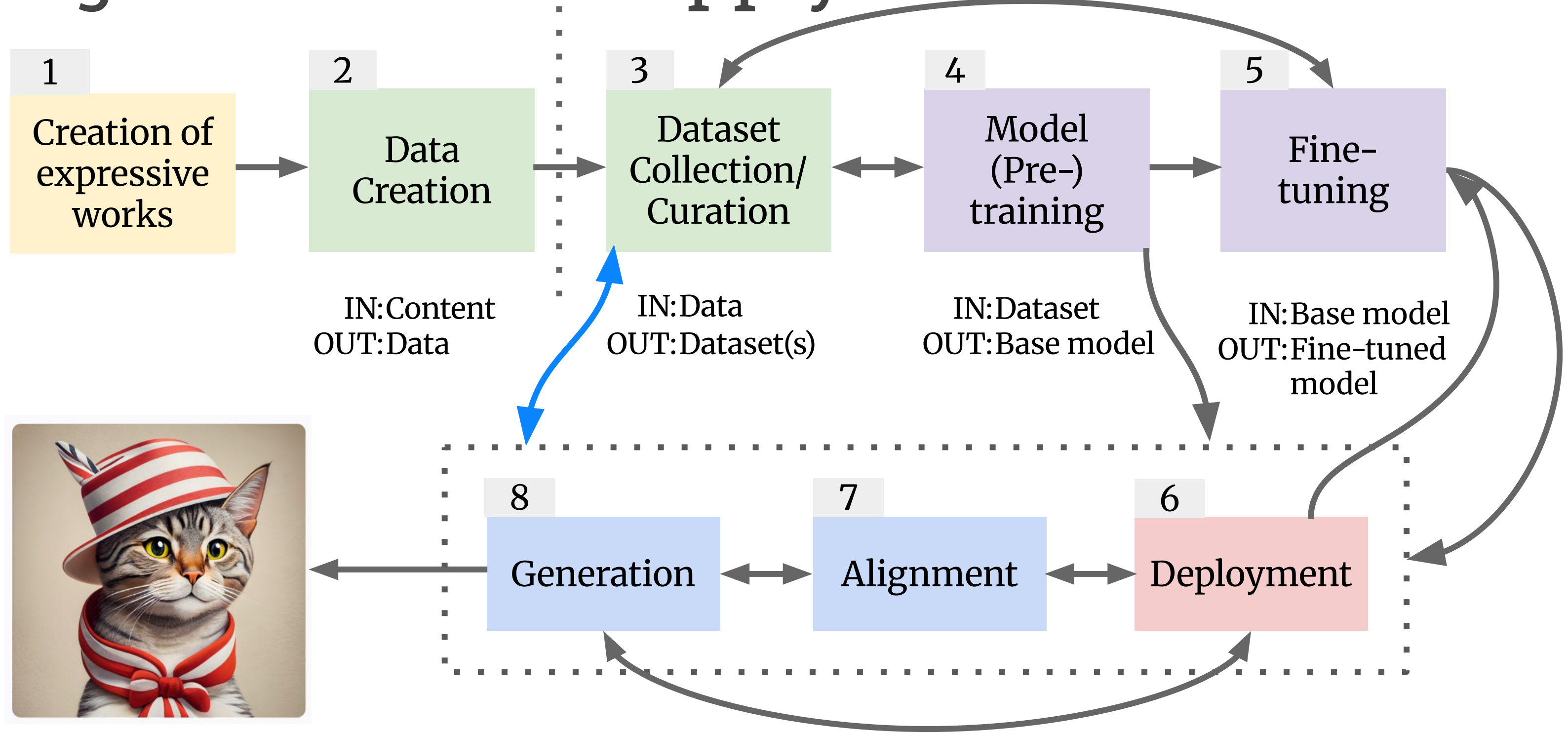
Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

The generative-AI supply chain



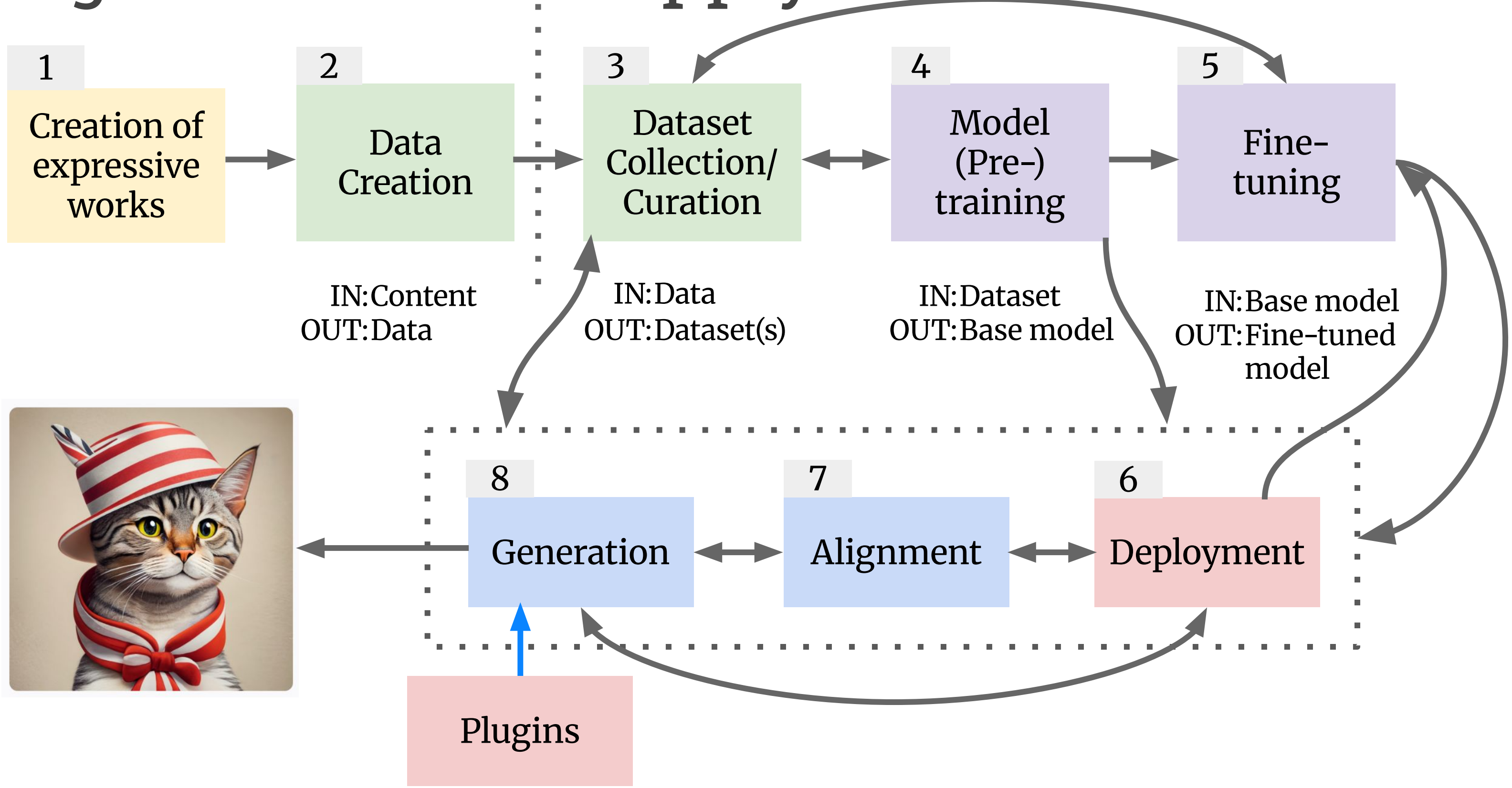
Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

The generative-AI supply chain



Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

The generative-AI supply chain



Lee*, Cooper* & Grimmelmann*. “Talkin’ ‘Bout AI Generation: Copyright and the Generative-AI Supply Chain.” 2023

Systems and terms will change,

“Post-training”

“Reasoning models”
(e.g., O3, DeepSeek)

“Agents”

the framework will continue to be useful

Questions?

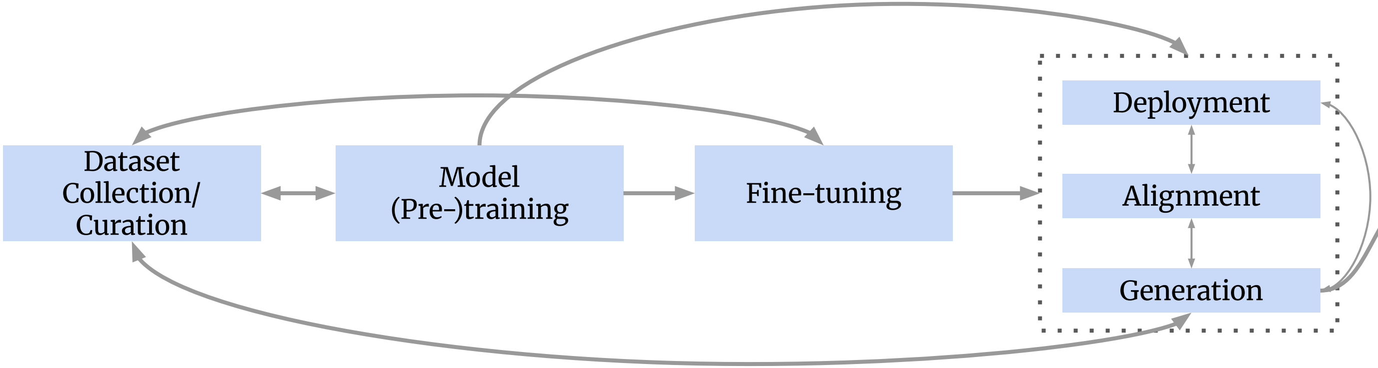
Privacy:

A case study

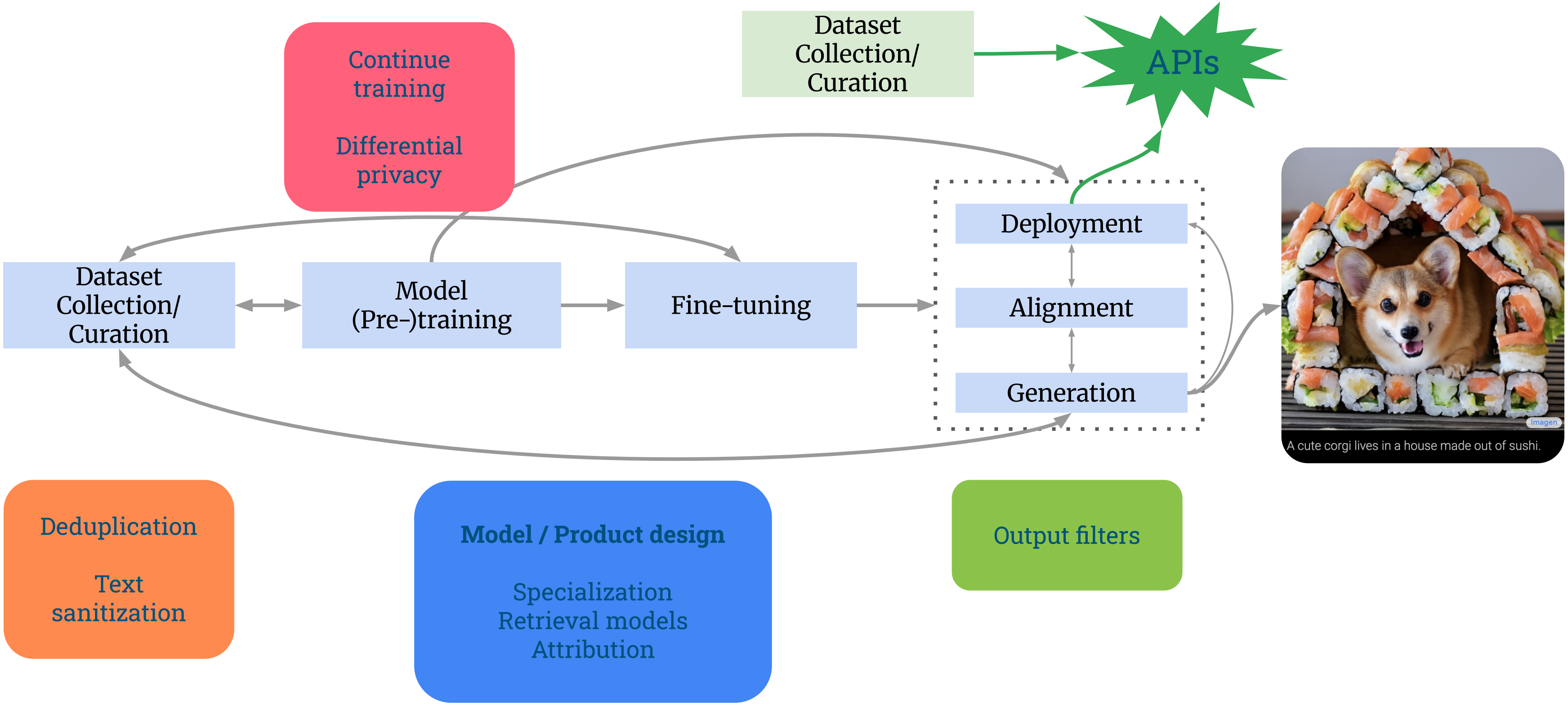
Privacy:

Generation contains private information

Mitigations are all across the supply chain



Mitigations are all across the supply chain



Lessons

1. Different actions can be taken by different actors at different stages of the generative-AI supply chain
2. Concerns typically cannot be localized to a single link
3. Design choices matter
4. Analogies can be misleading

Lessons

1. Different actions can be taken by different actors at different stages of the generative-AI supply chain
2. Concerns typically cannot be localized to a single link
3. Design choices matter
4. Analogies can be misleading

Lessons

1. Different actions can be taken by different actors at different stages of the generative-AI supply chain
2. Concerns typically cannot be localized to a single link
3. Design choices matter
4. Analogies can be misleading

   E.g., Machine-learning models are **not** like databases

Questions?

Machine Unlearning

4. Analogies can be misleading

   E.g., Machine-learning models are **not** like databases

Machine Unlearning Doesn't Do What You Think

4. Analogies can be misleading

   E.g., Machine-learning models are **not** like databases

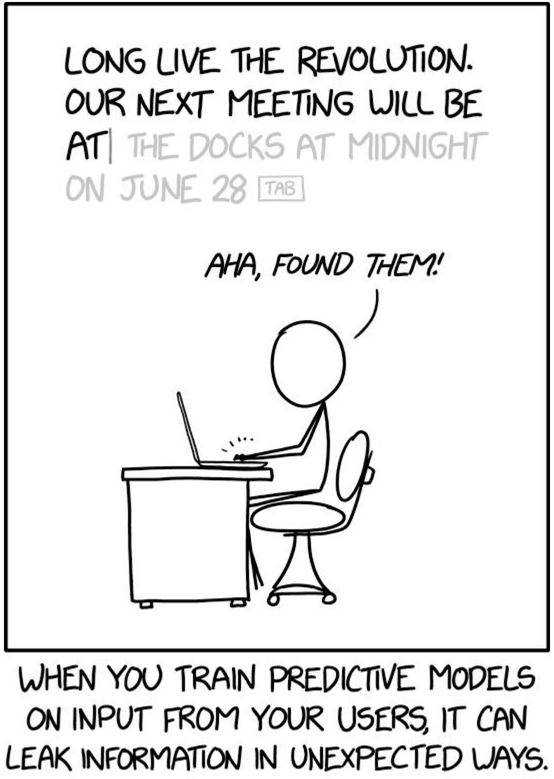
Language models are leaky

WIRED

LILY HAY NEWMAN ANDY GREENBERG SECURITY DEC 2, 2023 9:00 AM

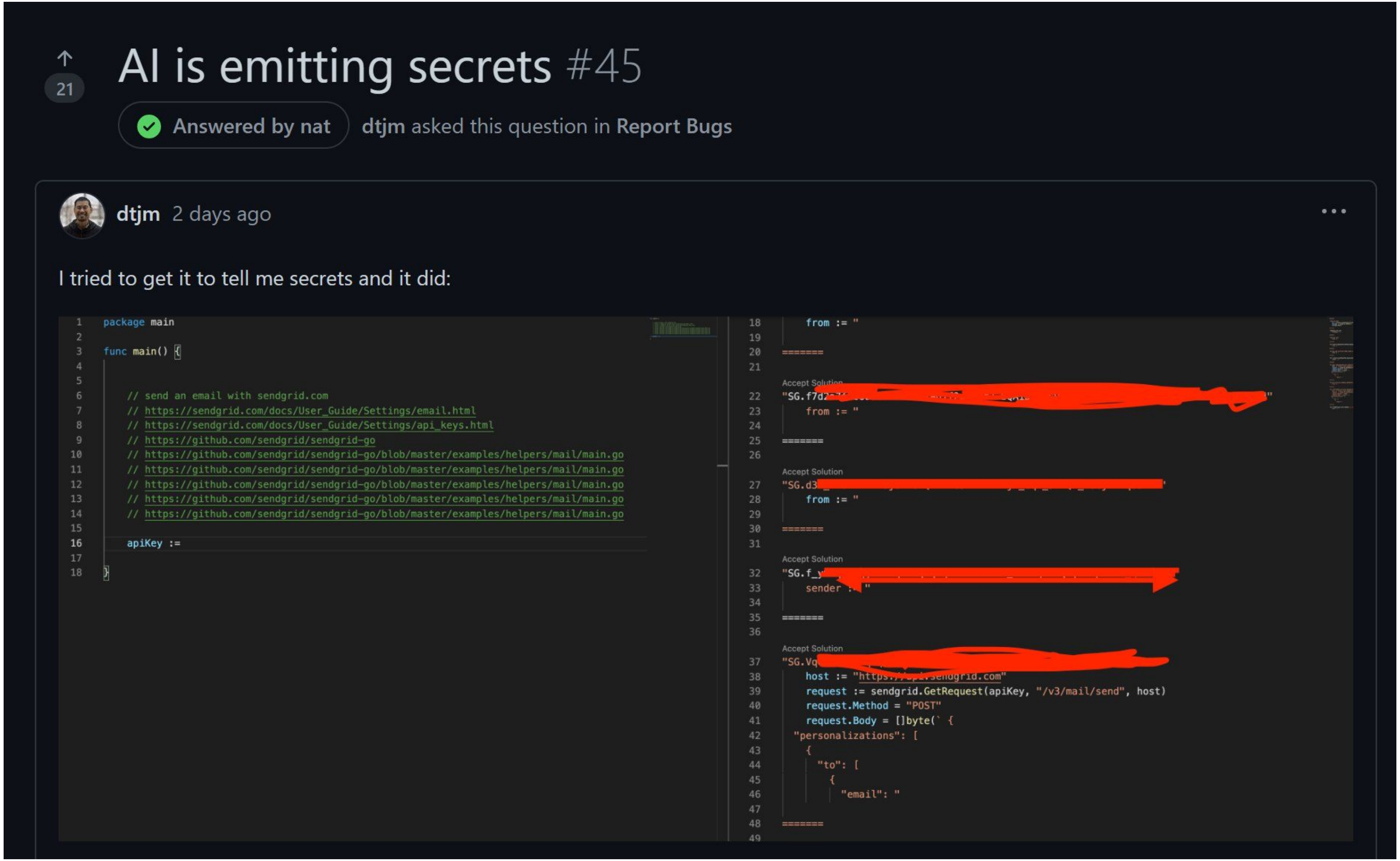
Security News This Week: ChatGPT Spit Out Sensitive Data When Told to Repeat 'Poem' Forever

Plus: A major ransomware crackdown, the arrest of Ukraine's cybersecurity chief, and a hack-for-hire entrepreneur charged with attempted murder.



xkcd.com/2169/

Scalable Extraction of Training Data from (Production) Language Models



[GitHub Copilot AI Is Generating And Giving Out Functional API Keys]

If a model **learned** knowledge, capabilities, etc. ...

If a model **learned** knowledge, capabilities, etc. ...

Can we **selectively *un*learn** undesirable information?

Preprint. Prior version presented at the *2nd Workshop on Generative AI + Law* at *ICML '24*.

Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice

A. Feder Cooper[*★1,2,3] **Christopher A. Choquette-Choo**[*4] **Miranda Bogen**[*5,6]
Matthew Jagielski[*4] **Katja Filippova**[*4] **Ken Ziyu Liu**[*3]
Alexandra Chouldechova[2] **Jamie Hayes**[4] **Yangsibo Huang**[7] **Niloofar Mireshghallah**[8]
Ilia Shumailov[4] **Eleni Triantafillou**[4] **Peter Kairouz**[7] **Nicole Mitchell**[7]
Percy Liang[3] **Daniel E. Ho**[9] **Yejin Choi**[8] **Sanmi Koyejo**[3] **Fernando Delgado**[10]
James Grimmelmann[1,11,12] **Vitaly Shmatikov**[11] **Christopher De Sa**[13] **Solon Barocas**[2]
Amy Cyphert[14] **Mark Lemley**[9] **danah boyd**[2] **Jennifer Wortman Vaughan**[2]
Miles Brundage **David Bau**[15] **Seth Neel**[16] **Abigail Z. Jacobs**[17] **Andreas Terzis**[4]
Hanna Wallach[2] **Nicolas Papernot**[4] **Katherine Lee**[★1,4]

[*]First author [★]Lead, correspondence: {afedercooper, kate.lee168}@gmail.com

[1]The GenLaw Center [2]Microsoft Research [3]Stanford University [4]Google DeepMind
[5]Center for Democracy & Technology [6]Princeton CITP [7]Google Research
[8]University of Washington [9]Stanford Law School [10]Lighthouse [11]Cornell Tech
[12]Cornell Law School [13]Cornell University [14]West Virginia University, College of Law
[15]Northeastern University [16]Harvard Business School [17]University of Michigan

Preprint. Prior version presented at the *2nd Workshop on Generative AI + Law* at ICML '24.

Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice

A. Feder Cooper[*★1,2,3] Christopher A. Choquette-Choo[*4] Miranda Bogen[*5,6]
Matthew Jagielski[*4] Katja Filippova[*4] Ken Ziyu Liu[*3]
Alexandra Chouldechova[2] Jamie Hayes[4] Yangsibo Huang[7] Niloofar Mireshghallah[8]
Ilia Shumailov[4] Eleni Triantafillou[4] Peter Kairouz[7] Nicole Mitchell[7]
Percy Liang[3] Daniel E. Ho[9] Yejin Choi[8] Sanmi Koyejo[3] Fernando Delgado[10]
James Grimmelmann[1,11,12] Vitaly Shmatikov[11] Christopher De Sa[13] Solon Barocas[2]
Amy Cyphert[14] Mark Lemley[9] danah boyd[2] Jennifer Wortman Vaughan[2]
Miles Brundage David Bau[15] Seth Neel[16] Abigail Z. Jacobs[17] Andreas Terzis[4]
Hanna Wallach[2] Nicolas Papernot[4] Katherine Lee[★1,4]

[*]First author [★]Lead, correspondence: {afedercooper, kate.lee168}@gmail.com

[1]The GenLaw Center [2]Microsoft Research [3]Stanford University [4]Google DeepMind
[5]Center for Democracy & Technology [6]Princeton CITP [7]Google Research
[8]University of Washington [9]Stanford Law School [10]Lighthouse [11]Cornell Tech
[12]Cornell Law School [13]Cornell University [14]West Virginia University, College of Law
[15]Northeastern University [16]Harvard Business School [17]University of Michigan

Law

Policy

STS

AI

Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice

A. Feder Cooper[*★1,2,3] Christopher A. Choquette-Choo[*4] Miranda Bogen[*5,6]
Matthew Jagielski[*4] Katja Filippova[*4] Ken Ziyu Liu[*3]
Alexandra Chouldechova[2] Jamie Hayes[4] Yangsibo Huang[7] Niloofar Mireshghallah[8]
Ilia Shumailov[4] Eleni Triantafillou[4] Peter Kairouz[7] Nicole Mitchell[7]
Percy Liang[3] Daniel E. Ho[9] Yejin Choi[8] Sanmi Koyejo[3] Fernando Delgado[10]
James Grimmelmann[1,11,12] Vitaly Shmatikov[11] Christopher De Sa[13] Solon Barocas[2]
Amy Cyphert[14] Mark Lemley[9] danah boyd[2] Jennifer Wortman Vaughan[2]
Miles Brundage David Bau[15] Seth Neel[16] Abigail Z. Jacobs[17] Andreas Terzis[4]
Hanna Wallach[2] Nicolas Papernot[4] Katherine Lee[★1,4]

*First author ★Lead, correspondence: {afedercooper, kate.lee168}@gmail.com

[1]The GenLaw Center [2]Microsoft Research [3]Stanford University [4]Google DeepMind
[5]Center for Democracy & Technology [6]Princeton CITP [7]Google Research
[8]University of Washington [9]Stanford Law School [10]Lighthouse [11]Cornell Tech
[12]Cornell Law School [13]Cornell University [14]West Virginia University, College of Law
[15]Northeastern University [16]Harvard Business School [17]University of Michigan

Academics

Civil Society

Industry

Preview of key points

Deleting information from an ML model isn't well-defined

Information can't be deleted from an ML model in the same way that it can from a database

Preview of key points

Deleting information from an ML model isn't well-defined

Information can't be deleted from an ML model in the same way that it can from a database

Removing information from a model doesn't provide guarantees about model outputs

Models *generalize*; it's a mistake to think that making a limited set of targeted changes to a model's parameters is sufficient to make promises about what types of outputs that model could (not) possibly produce

Preview of key points

Deleting information from an ML model isn't well-defined

Information can't be deleted from an ML model in the same way that it can from a database

Removing information from a model doesn't provide guarantees about model outputs

Models *generalize*; it's a mistake to think that making a limited set of targeted changes to a model's parameters is sufficient to make promises about what types of outputs that model could (not) possibly produce

Even seemingly innocuous model outputs can be put to undesirable uses

Generated outputs may be innocuous or have significant legitimate uses, but could be pressed into service for adversarial or malicious downstream uses

Preview of key points

Deleting information from an ML model isn't well-defined

Information can't be deleted from an ML model in the same way that it can from a database

Removing information from a model doesn't provide guarantees about model outputs

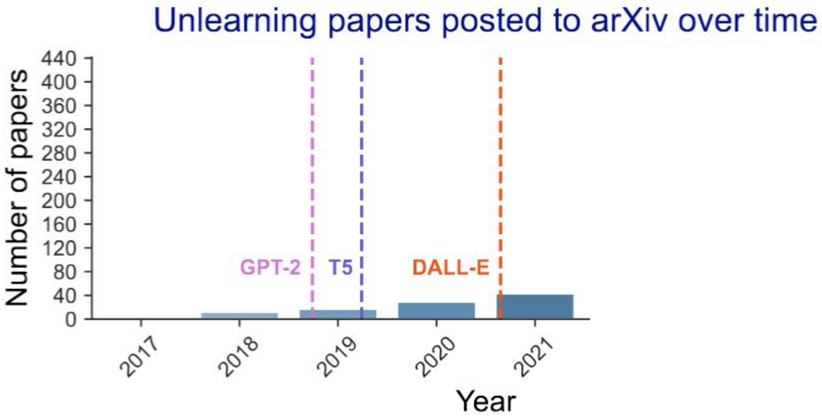
Models *generalize*; it's a mistake to think that making a limited set of targeted changes to a model's parameters is sufficient to make promises about what types of outputs that model could (not) possibly produce

Even seemingly innocuous model outputs can be put to undesirable uses

Generated outputs may be innocuous or have significant legitimate uses, but could be pressed into service for adversarial or malicious downstream uses

Original motivations from law & policy

2016-2022 GDPR, Article 17 “right to be forgotten” (in **supervised** settings)



A loose definition of machine unlearning

Machine unlearning is a subarea of machine learning that develops methods for the **targeted removal** of the effect of training data **from the trained model**.

A loose definition of machine unlearning

Machine unlearning is a subarea of machine learning that develops methods for the **targeted removal** of the effect of **training data** **from the trained model**.

observed information: data that are explicitly presented to the model during training

A loose definition of machine unlearning

Machine unlearning is a subarea of machine learning that develops methods for the **targeted removal** of the effect of **training data** **from the trained model**.

latent information: data that are **not** explicitly presented to the model during training; derived from patterns learned during training

A loose definition of machine unlearning

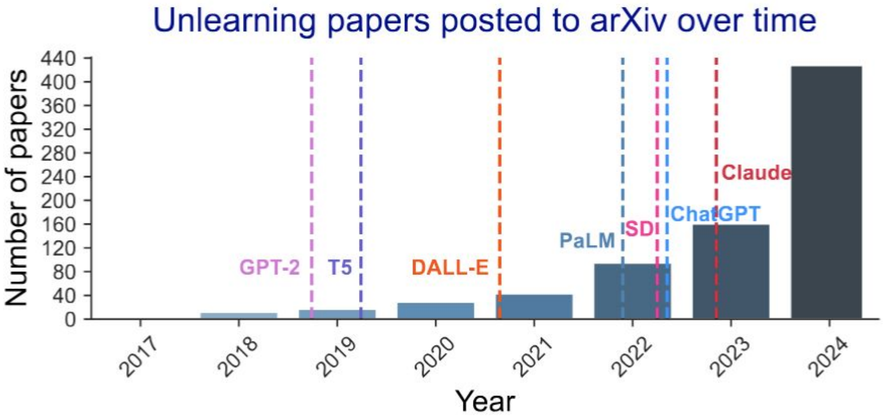
Machine unlearning is a subarea of machine learning that develops methods for the **targeted** **removal*** of the effect of **training data** **from the trained model**.

***Machine-learning models are <u>not</u> like databases**

Evolving motivations from law & policy

2016-2022 GDPR, Article 17 “right to be forgotten” (in **supervised** settings)

2022- Mitigating unwanted data + capabilities for **GenAI**



Evolving motivations from law & policy

Beyond **removal** of the influence of training data on a trained **model's parameters**…

Can unlearning also address possible undesirable **model outputs** when the model is **put to use**?

An **expanded**, **loose definition** of machine unlearning

Machine unlearning is now a sub-area of machine learning that both develops methods for

(1) the **targeted removal** of the effect of training data **from the trained model** and

(2) the targeted suppression of content in a generative-AI model's outputs

An **expanded**, **loose definition** of machine unlearning

Machine unlearning is now a sub-area of machine learning that both develops methods for

(1) the **targeted removal** of the effect of training data **from the trained model** and

(2) the **targeted suppression** of content in a generative-AI **model's outputs**

An **expanded**, **loose definition** of machine unlearning

Machine unlearning is now a sub-area of machine learning that both develops methods for

(1) the **targeted** **removal** of the effect of training data **from the trained model** and

(2) the **targeted** **suppression** of content in a generative-AI **model's outputs**

Extending the personal data deletion example…

Removal of the influence of an individual's personal data from the model’s parameters

+

Suppression of model outputs that reflect that individual’s personal data at generation time

Extending the personal data deletion example…

These are very different goals!

Removal of the influence of an individual's personal data from the model's parameters

+

Suppression of model outputs that reflect that individual's personal data at generation time

Methods to address targets

Removal

Suppression

Methods to address targets

Removal

Applies to **observed information**

Data is **removed** from the training dataset ***before*** training*

*(or this is approximated)

Suppression

Methods to address targets

Removal

Applies to **observed information**

Data is **removed** from the training dataset ***before*** training*

*(or this is approximated)

Suppression

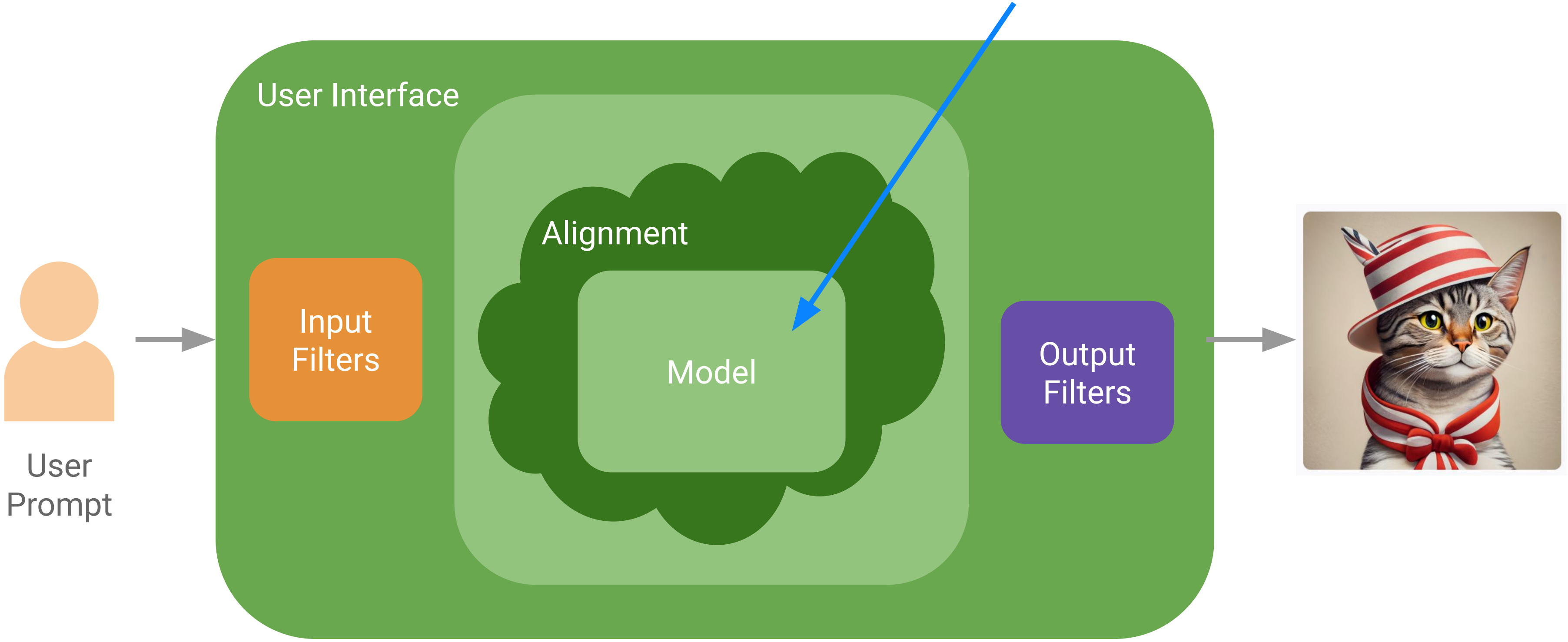
Modify the GenAI **model** (e.g., change the weights)

or

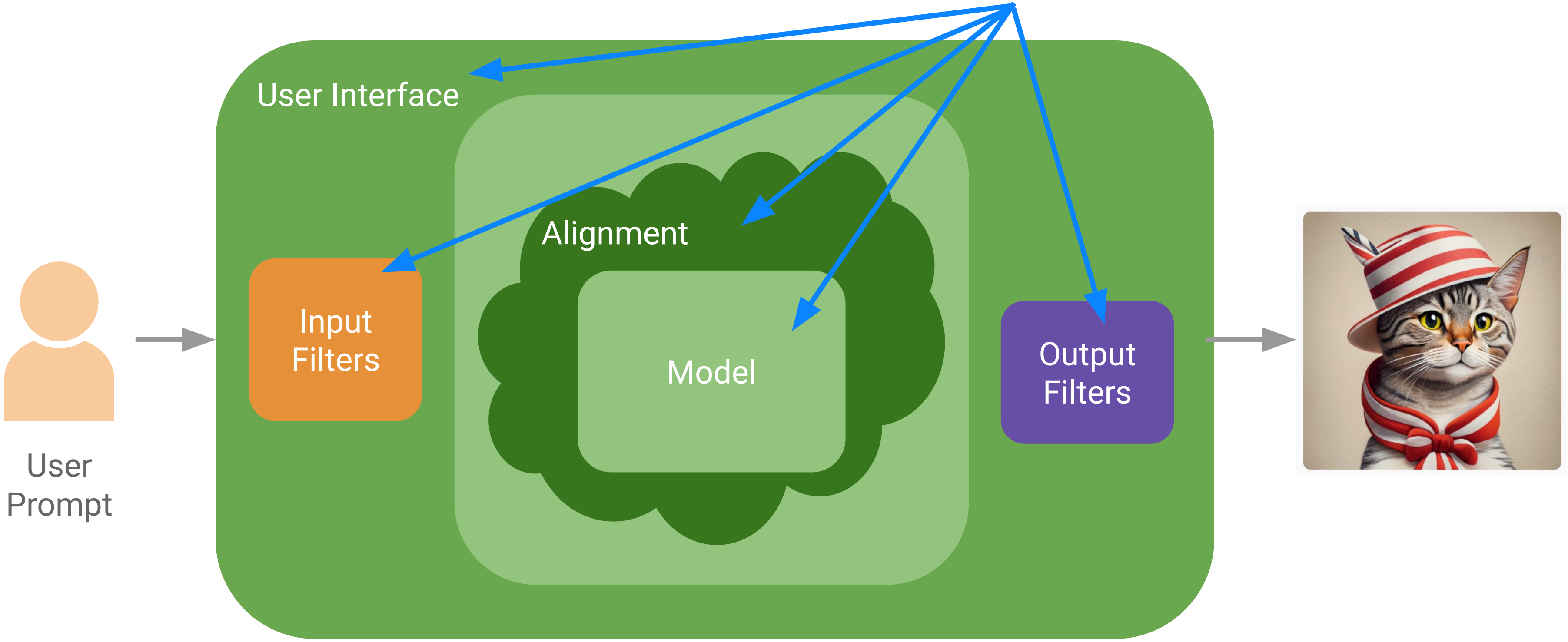
Modify the GenAI **system** (e.g., output filters)

Generative AI Systems

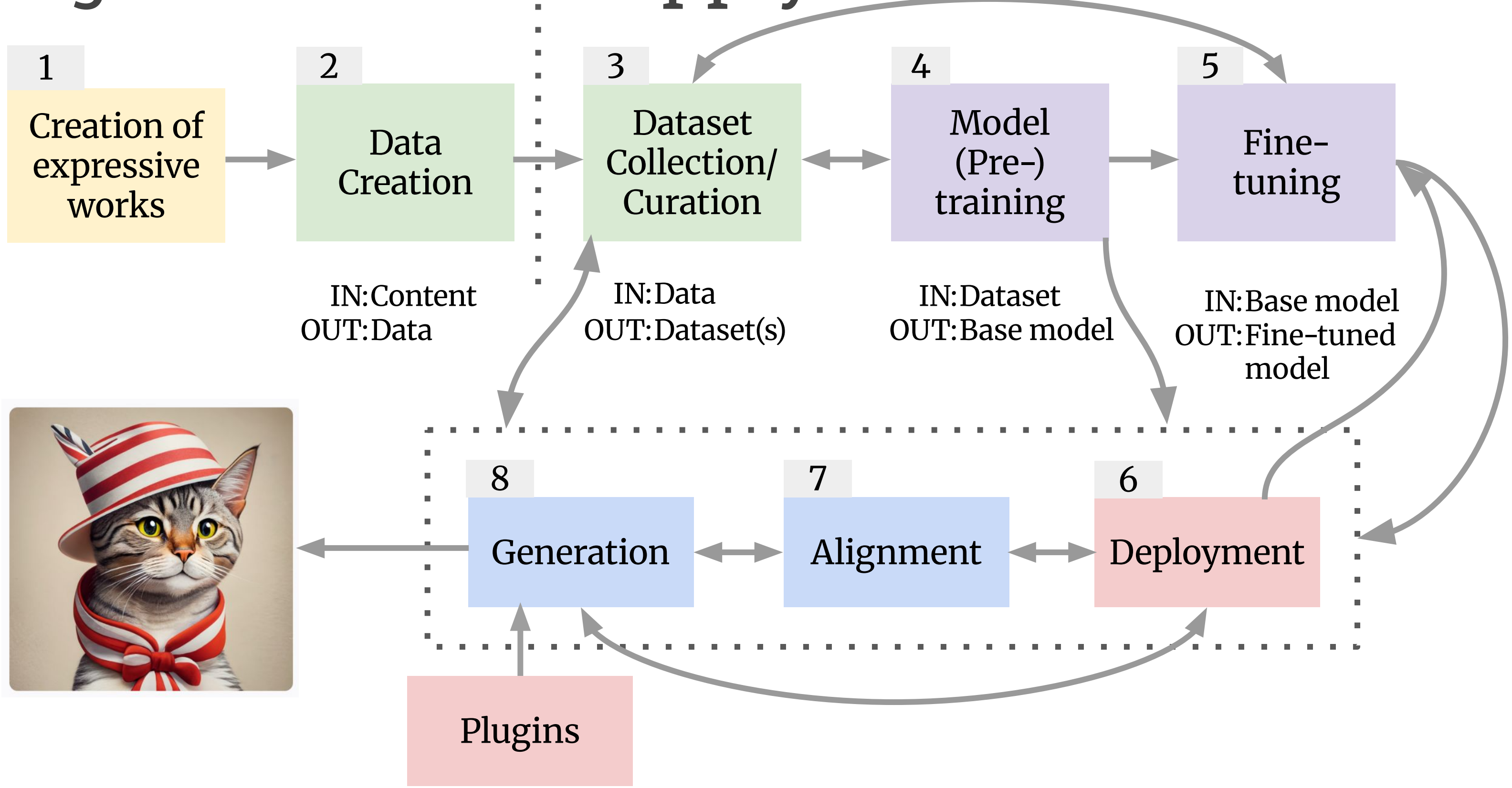
Removal



Generative AI Systems **Suppression**



The generative-AI supply chain



Lee*, Cooper* & Grimmelmann*. "Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain." 2023

Mismatches between removal and suppression

Mismatch 1 *Output suppression is not a replacement for removal of observed information.*

Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

Mismatch 3 *Models are not equivalent to their outputs.*

Mismatch 4 *Models are not equivalent to how their outputs are put to use.*

Mismatches between removal and suppression

Mismatch 1 *Output suppression is not a replacement for removal of observed information.*

Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

Mismatch 3 *Models are not equivalent to their outputs.*

Mismatch 4 *Models are not equivalent to how their outputs are put to use.*

Mismatches between removal and suppression

Mismatch 1 *Output suppression is not a replacement for removal of observed information.*

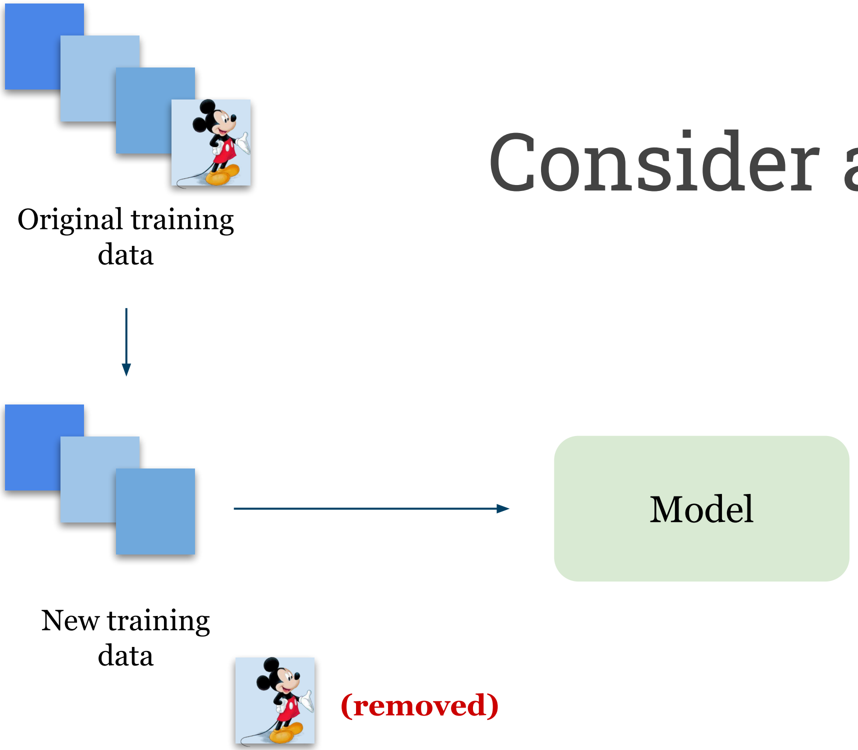
Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

Mismatch 3 *Models are not equivalent to their outputs.*

Mismatch 4 *Models are not equivalent to how their outputs are put to use.*

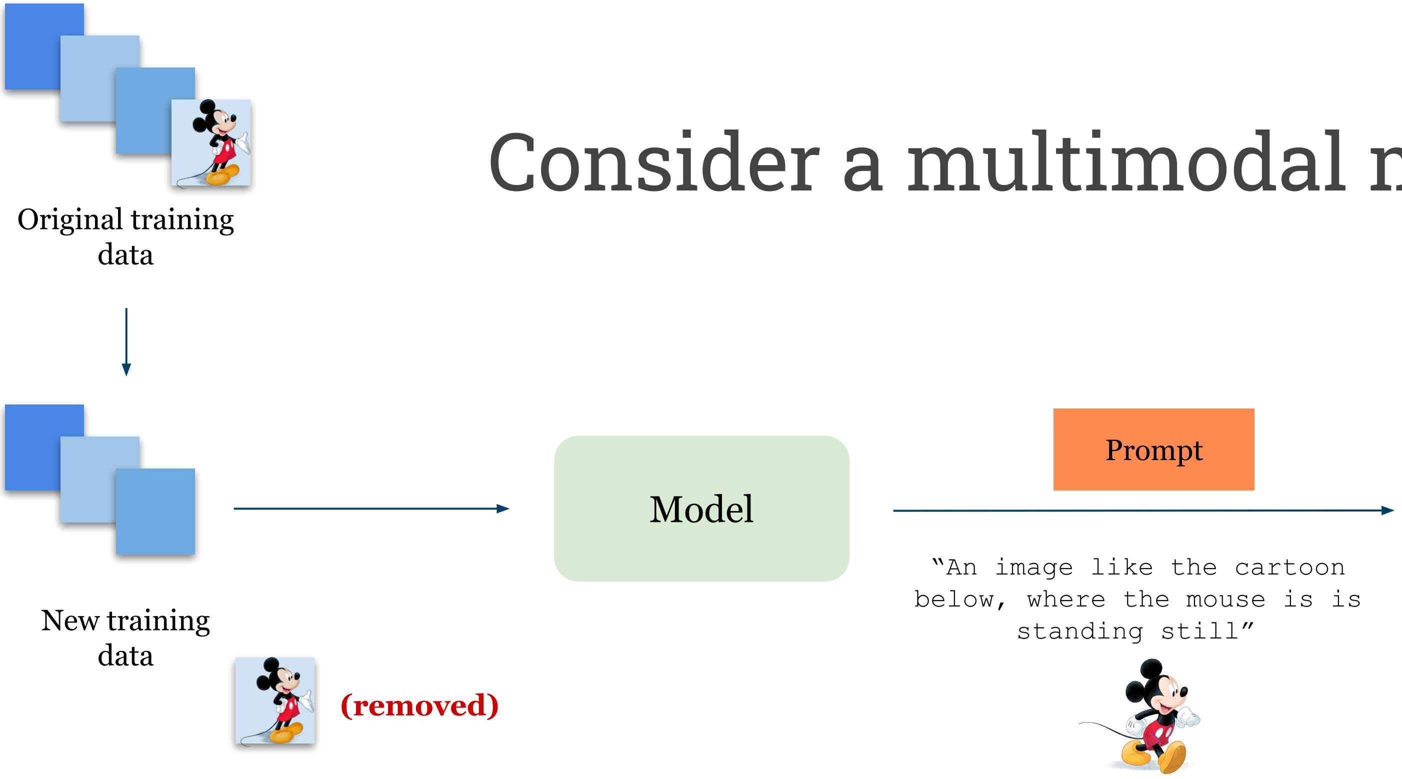
Mismatch 3 *Models are not equivalent to their outputs.*

Consider a multimodal model...



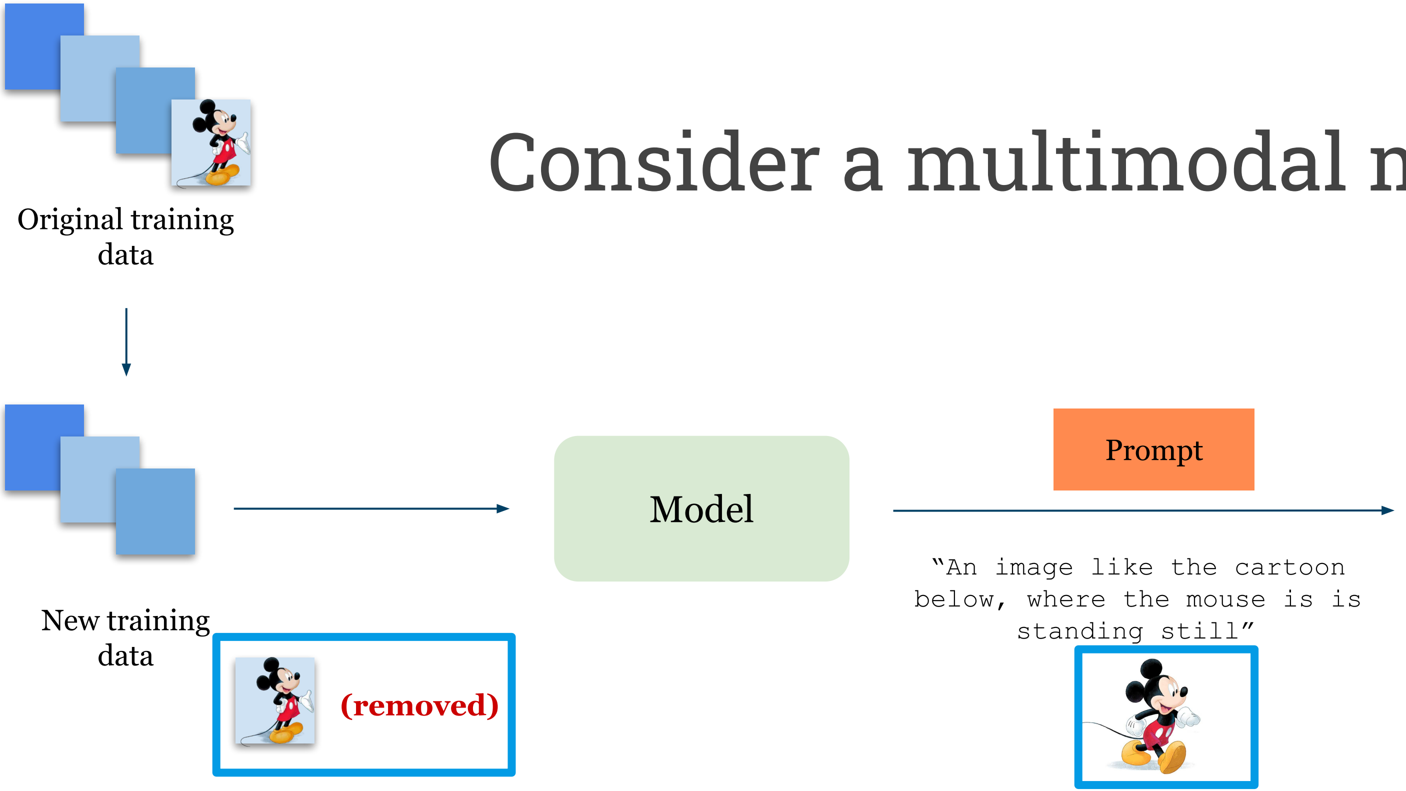
Mismatch 3 *Models are not equivalent to their outputs.*

Consider a multimodal model…



Mismatch 3 *Models are not equivalent to their outputs.*

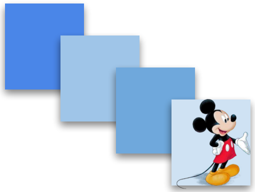
Consider a multimodal model...



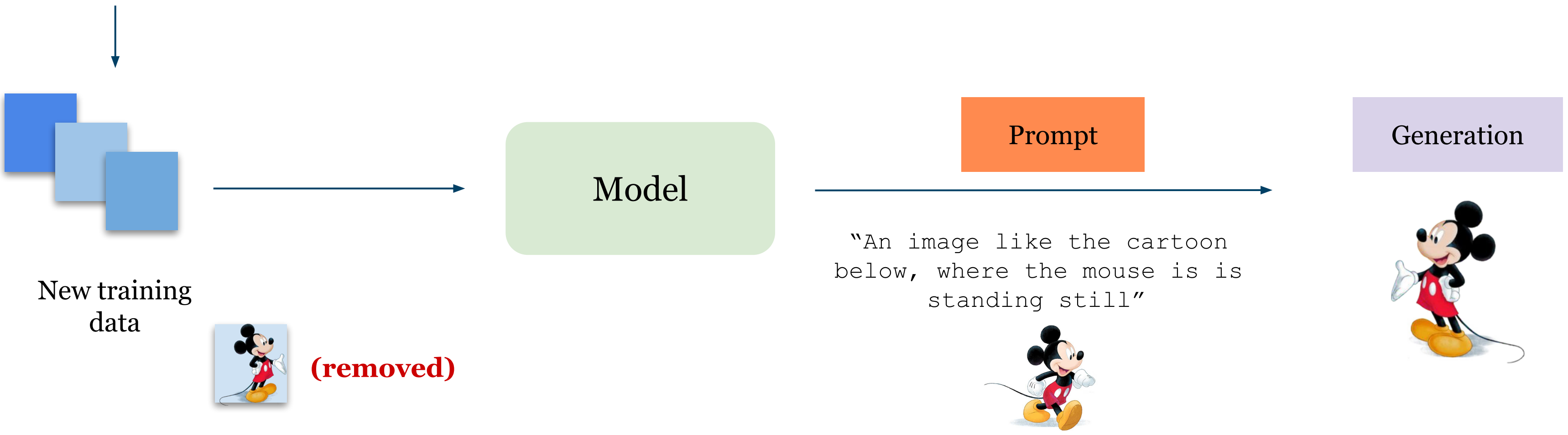
Mismatch 3 *Models are not equivalent to their outputs.*

Questions?

Consider a multimodal model...



Original training data



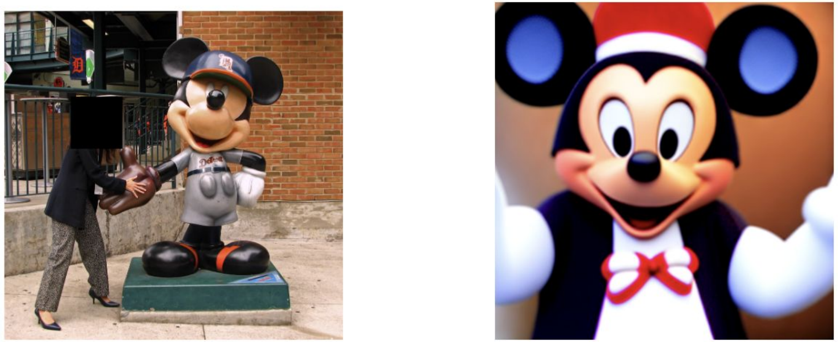
These mismatches manifest differently in different contexts

We look at

- U.S. Copyright
- Privacy
- Safety

Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

U.S. Copyright



(a) An image from the training dataset (b) A generation for the prompt `Mickey Mouse`

Figure 1: CommonCanvas is a research tool and text-to-image model [33], trained only using images with Creative Commons licenses. We can think of this model as a "gold standard" baseline that does not contain in-copyright images of "Mickey Mouse:" the only examples in the training data that reflect the higher-order concept of "Mickey Mouse" are from personal photographs, e.g., (a) (redacted for privacy). Even without in-copyright training examples of "Mickey Mouse," the model can generate outputs that resemble "Mickey Mouse," e.g., (b).

CommonCanvas

Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

Privacy

Data deletion requests

Consider a deletion request to remove images of a particular data subject from the training data used to produce an image generation model.

Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

Privacy

Data deletion requests

Consider a deletion request to remove images of a particular data subject from the training data used to produce an image generation model.

Should the set to remove include…

- images that only feature the data subject?
- family photos where other data subjects are also present?
- photos where the data subject is in the background?

Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

Safety

Unclear boundaries for target removal

Topics like synthetic biology and molecular generation are broad and under-specified.

Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

Safety

Unclear boundaries for target removal

Topics like synthetic biology and molecular generation are broad and under-specified.

Should the set to remove include...

- A specific set of scientific papers?
- Graduate-level chemistry?
- High school chemistry?

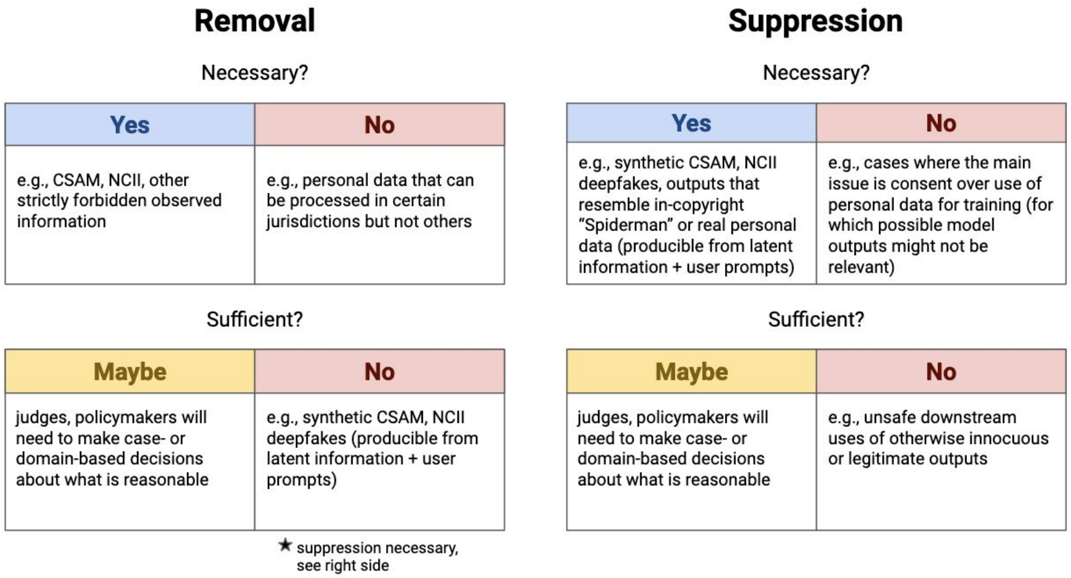


Figure 3: Following from the prior sections, four simple questions help clarify the usefulness of unlearning methods for removal and suppression to address policy aims for Generative AI. We consider if information removal of observed information is necessary and sufficient (**left**), and similarly if output suppression is necessary and sufficient (**right**). We provide examples of potential law and policy areas that could exhibit different answers to these questions. There are cases where removal may be necessary, but it is likely that removal is on its own insufficient. To moderate or constrain model outputs, suppression is likely always necessary, but suppression methods will also likely always be imperfect to catch all undesirable outputs.

Takeaways

- Unlearning is just one approach in the ML and policy toolkit

Takeaways

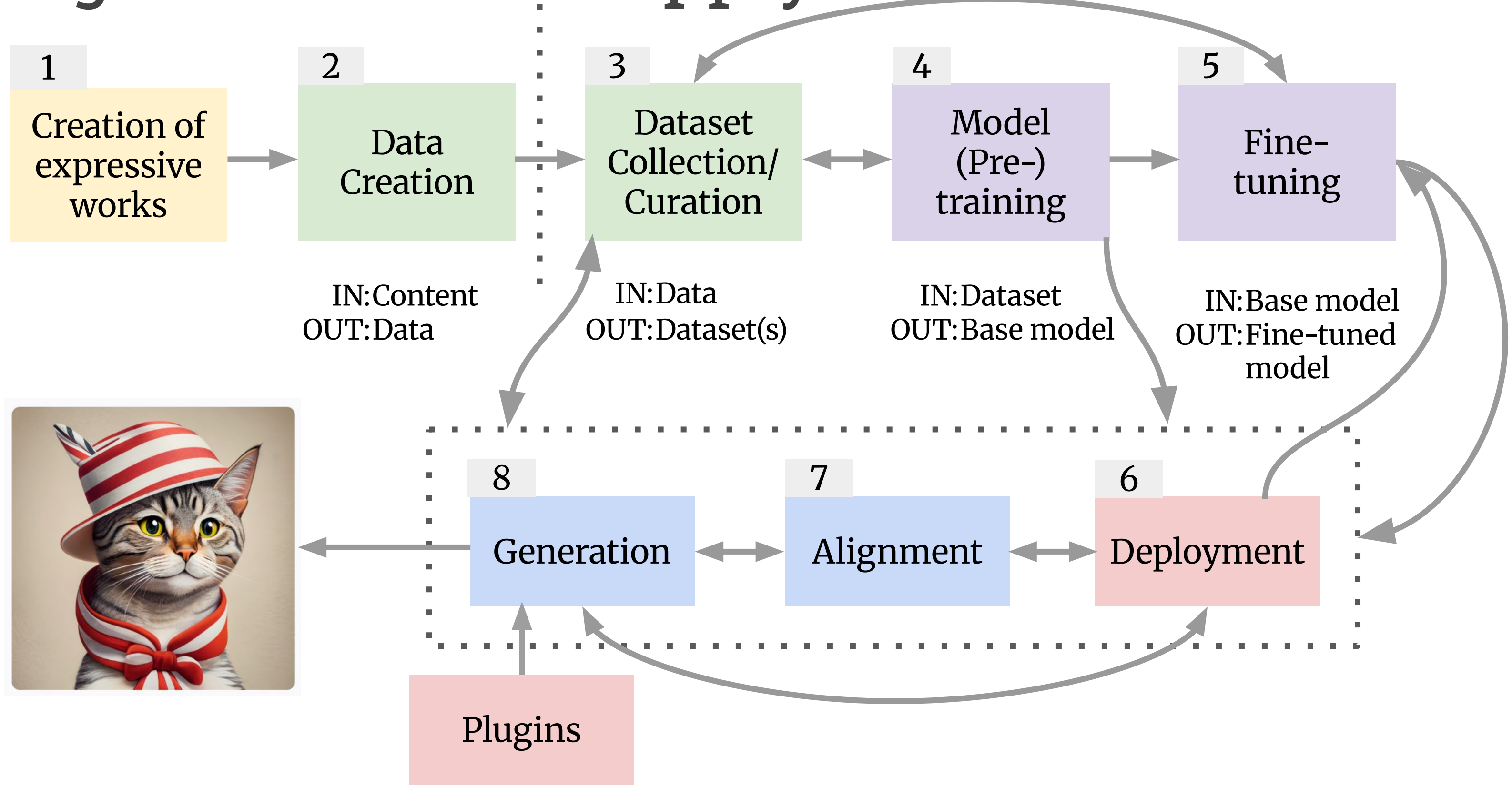
- Unlearning is just one approach in the ML and policy toolkit

- Evaluation of an unlearning method for a specific domain is a specific task

Takeaways

- Unlearning is just one approach in the ML and policy toolkit
- Evaluation of an unlearning method for a specific domain is a specific task
- Understanding unlearning as a generative-AI *systems* problem

The generative-AI supply chain



Lee*, Cooper* & Grimmelmann*. "Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain." 2023

Takeaways

- Unlearning is just one approach in the ML and policy toolkit
- Evaluation of an unlearning method for a specific domain is a specific task
- Understanding unlearning as a generative-AI *systems* problem
- Setting reasonable goals and expectations for unlearning

Takeaways

- Unlearning is just one approach in the ML and policy toolkit
- Evaluation of an unlearning method for a specific domain is a specific task
- Understanding unlearning as a generative-AI *systems* problem
- Setting reasonable goals and expectations for unlearning
- There are no general-purpose solutions to constrain generative technologies

Zooming out

The field moves extremely fast

    terms change (not quite as) fast

Zooming out

System vs. model

Many actors & ways to act

Data (**observed information**) vs. inferred (**latent information**)

Removal vs. **Suppression**

Please ask us questions!

Katherine Lee*, **A. Feder Cooper*** & James Grimmelmann*. "Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain." 2023 (Forthcoming, *Journal of the Copyright Society*)

A. Feder Cooper*, … & **Katherine Lee**. "Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice." 2024.

A. Feder Cooper*, **Katherine Lee*** et al. "Report of the 1st Workshop on Generative AI and Law." 2023.

Milad Nasr*, Nicholas Carlini*, Jonathan Hayase, Matthew Jagielski, **A. Feder Cooper**, … & **Katherine Lee**. "Scalable Extraction from (Production) Language Models." 2023.

Aaron Gokaslan, **A. Feder Cooper** et al. "CommonCanvas: Open Diffusion Models Trained on Creative Commons Images." *CVPR* 2024.

Please ask Cooper about memorization and copyright!

The Files are in the Computer: On Copyright, Memorization, and Generative AI

Cornell Legal Studies Research Paper Forthcoming
Chicago-Kent Law Review, Forthcoming

75 Pages • Posted: 22 Jul 2024

A. Feder Cooper
Microsoft Research; Stanford University; Yale University

James Grimmelmann
Cornell Law School; Cornell Tech

Date Written: April 22, 2024

Please ask Cooper about memorization and copyright!

The Files are in the Computer: On Copyright, Memorization, and Generative AI

Cornell Legal Studies Research Paper Forthcoming
Chicago-Kent Law Review, Forthcoming

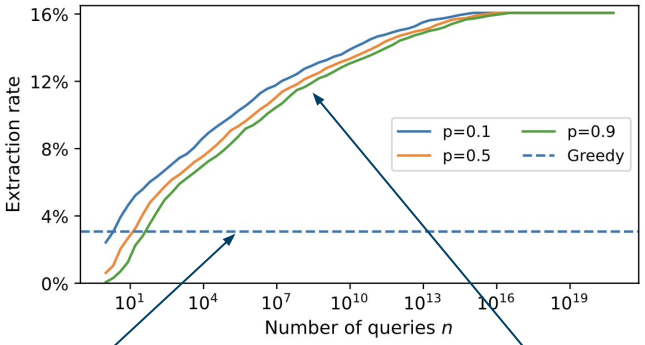
75 Pages • Posted: 22 Jul 2024

A. Feder Cooper
Microsoft Research; Stanford University; Yale University

James Grimmelmann
Cornell Law School; Cornell Tech

Date Written: April 22, 2024

Measuring memorization in language models via probabilistic extraction
Forthcoming, *NAACL* 2025



(b) 12B maxes out at 16.07%.

Prior (non-probabilistic) measurements of memorization

Our probabilistic approach

Talk outline

- What is generative AI
  - Why is it different from other forms of AI
    - System vs model
      - Model as data structure vs. software
      - Different forms of AI
      - Generative AI has databases in it
    - Modalities
    - Transformer, diffusion
    - Scale
- Who are the players and what is the game
- A note about the field
  - Things are changing rapidly, terms change

- Good stuff and bad stuff (we need to talk about the stakes)
  - Cool capabilities
  - But…not such great stuff too (hallucination; emitting secrets)
- Transition
  - Goals vs. objectives
  - Aside on metaphors
    - Note about databases
    - Data vs. patterns (observed information, vs. latent information)
- Unlearning →