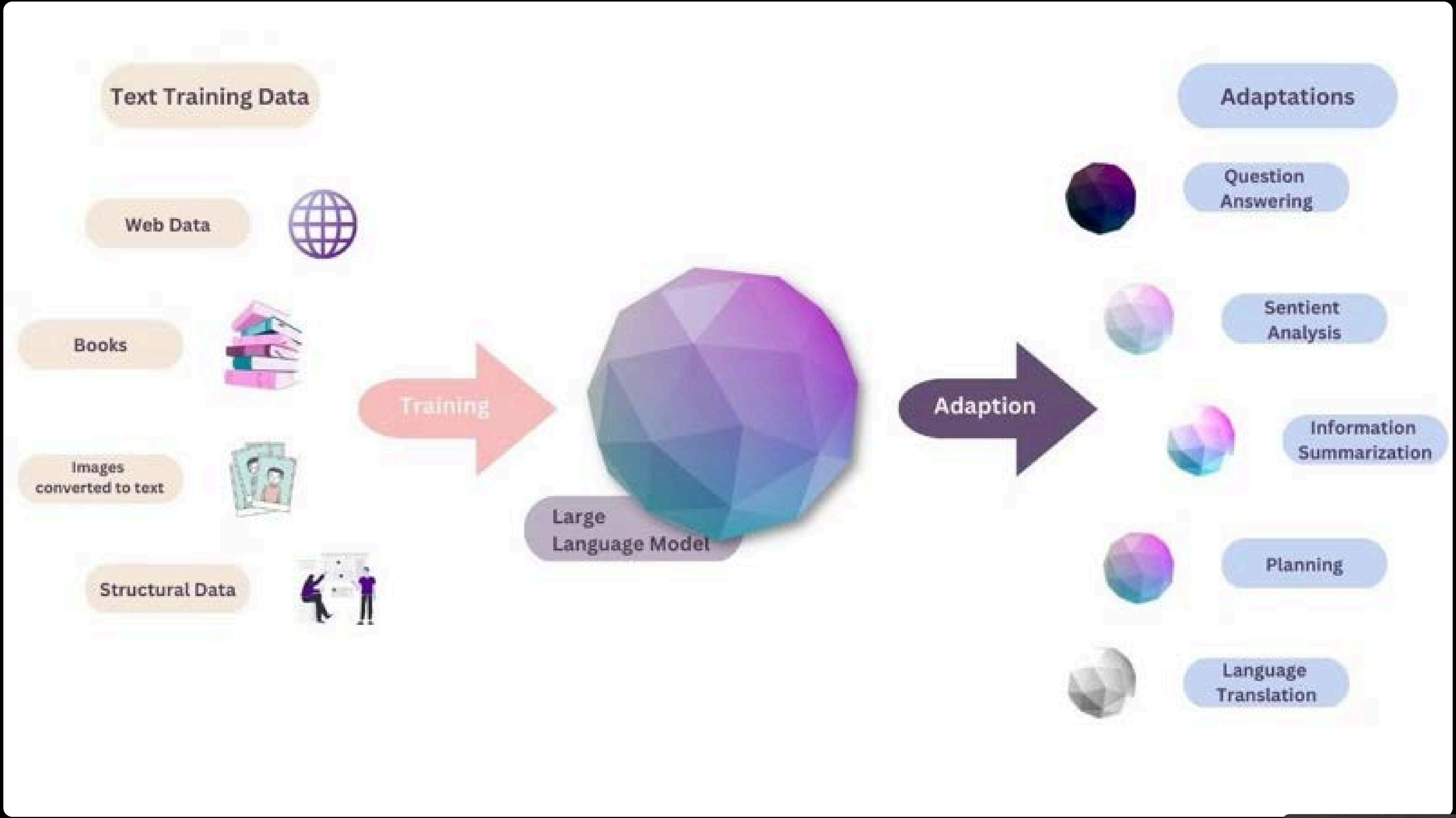
The Overlooked GDPR Compliance Challenges of Large Language Models

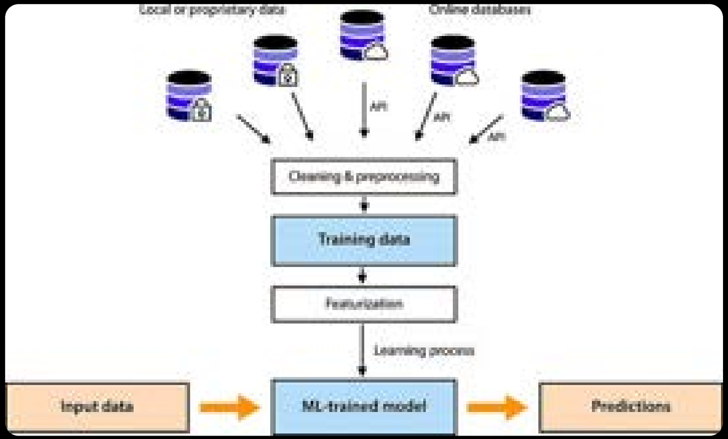
by Antony Hibbert

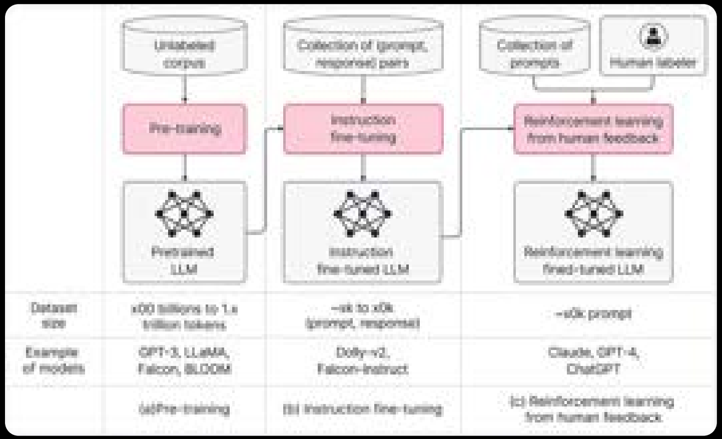
Large Language Models (LLMs) and their usage



Made with Gamma

Training of LLMs

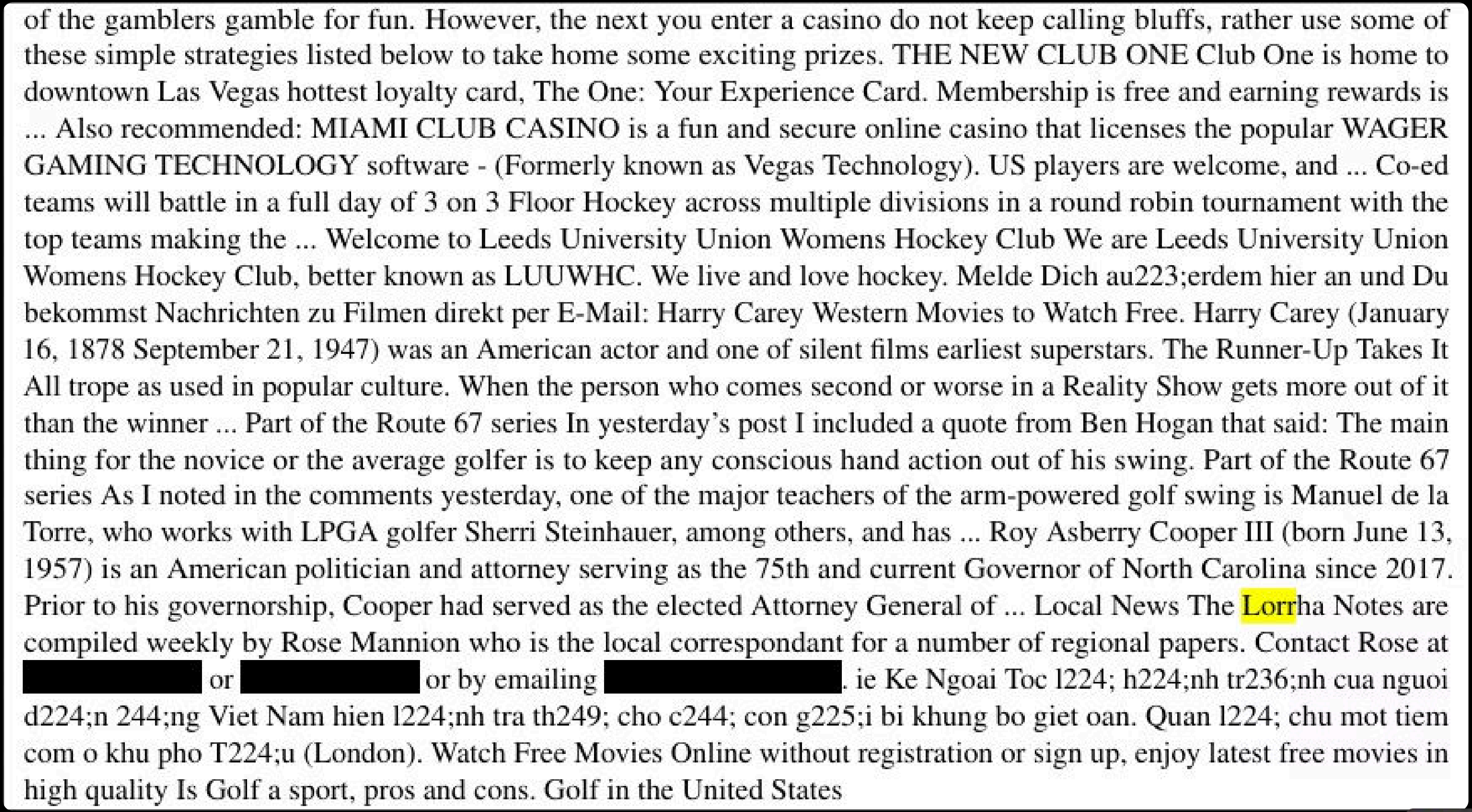



Unlabeled corpus
Collection of (prompt, response) pairs
Collection of prompts
Human labeler
Pre-training
Instruction fine-tuning
Reinforcement learning from human feedback
Pretrained LLM
Instruction fine-tuned LLM
Reinforcement learning fined-tuned LLM

Dataset size	x00 billions to 1.x trillion tokens	~xk to x0k (prompt, response)	~x0k prompt
Example of models	GPT-3, LLaMA, Falcon, BLOOM	Dolly-v2, Falcon-instruct	Claude, GPT-4, ChatGPT
	(a)Pre-training	(b) Instruction fine-tuning	(c) Reinforcement learning from human feedback

Memorization of Personal Data in LLMs

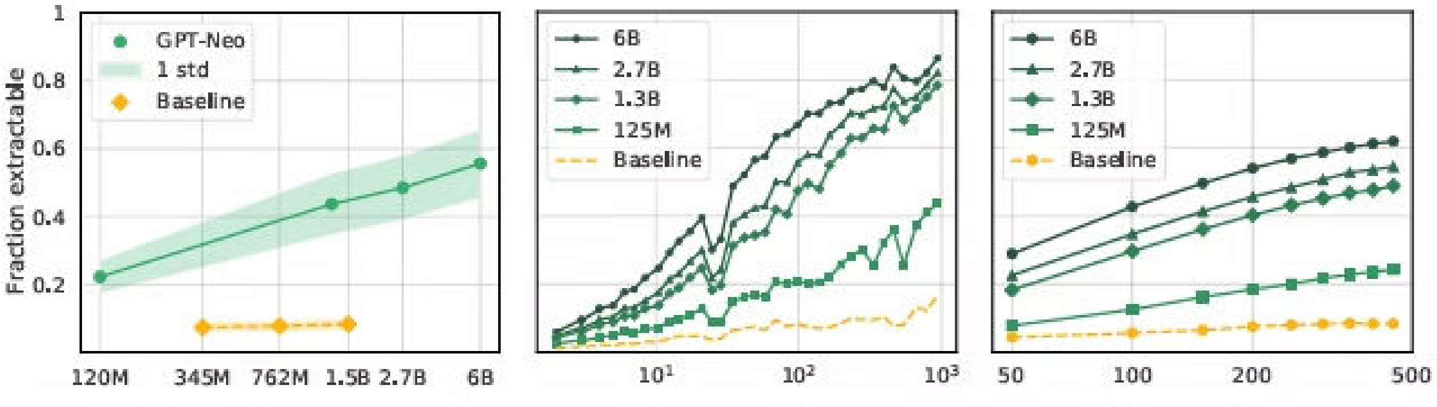
- Large Language Models (LLMs) show substantial evidence of memorizing personal data.



of the gamblers gamble for fun. However, the next you enter a casino do not keep calling bluffs, rather use some of these simple strategies listed below to take home some exciting prizes. THE NEW CLUB ONE Club One is home to downtown Las Vegas hottest loyalty card, The One: Your Experience Card. Membership is free and earning rewards is ... Also recommended: MIAMI CLUB CASINO is a fun and secure online casino that licenses the popular WAGER GAMING TECHNOLOGY software - (Formerly known as Vegas Technology). US players are welcome, and ... Co-ed teams will battle in a full day of 3 on 3 Floor Hockey across multiple divisions in a round robin tournament with the top teams making the ... Welcome to Leeds University Union Womens Hockey Club We are Leeds University Union Womens Hockey Club, better known as LUUWHC. We live and love hockey. Melde Dich au223;erdem hier an und Du bekommst Nachrichten zu Filmen direkt per E-Mail: Harry Carey Western Movies to Watch Free. Harry Carey (January 16, 1878 September 21, 1947) was an American actor and one of silent films earliest superstars. The Runner-Up Takes It All trope as used in popular culture. When the person who comes second or worse in a Reality Show gets more out of it than the winner ... Part of the Route 67 series In yesterday's post I included a quote from Ben Hogan that said: The main thing for the novice or the average golfer is to keep any conscious hand action out of his swing. Part of the Route 67 series As I noted in the comments yesterday, one of the major teachers of the arm-powered golf swing is Manuel de la Torre, who works with LPGA golfer Sherri Steinhauer, among others, and has ... Roy Asberry Cooper III (born June 13, 1957) is an American politician and attorney serving as the 75th and current Governor of North Carolina since 2017. Prior to his governorship, Cooper had served as the elected Attorney General of ... Local News The Lorrha Notes are compiled weekly by Rose Mannion who is the local correspondant for a number of regional papers. Contact Rose at [redacted] or [redacted] or by emailing [redacted]. ie Ke Ngoai Toc l224; h224;nh tr236;nh cua nguoi d224;n 244;ng Viet Nam hien l224;nh tra th249; cho c244; con g225;i bi khung bo giet oan. Quan l224; chu mot tiem com o khu pho T224;u (London). Watch Free Movies Online without registration or sign up, enjoy latest free movies in high quality Is Golf a sport, pros and cons. Golf in the United States

When does Memorization happen?

- As LLMs increase in size, their tendency to memorize personal data increases
- Memorization is influenced by the diversity of training data
- More context aids extraction
- Data exposed later to the model more likely to be memorized rather than generalized



(a) Model scale (b) Data repetition (c) Context size

Memorization as a Desirable Feature

Advantages of Memorization

- Accuracy of Answers
- Facts
- Quotes

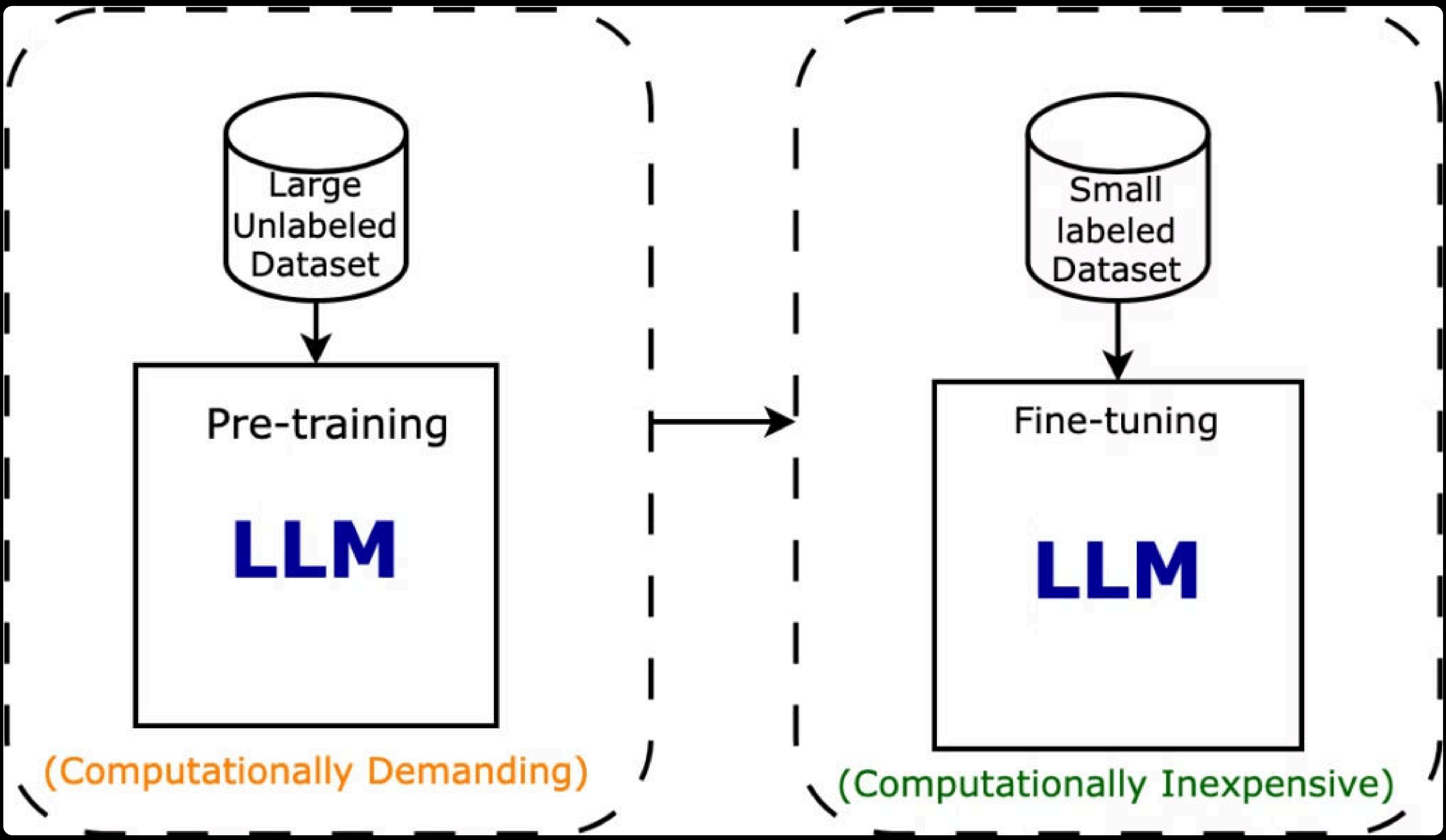
Disadvantages of Memorization

- Breaches of Privacy
- Breaches of Intellectual Property

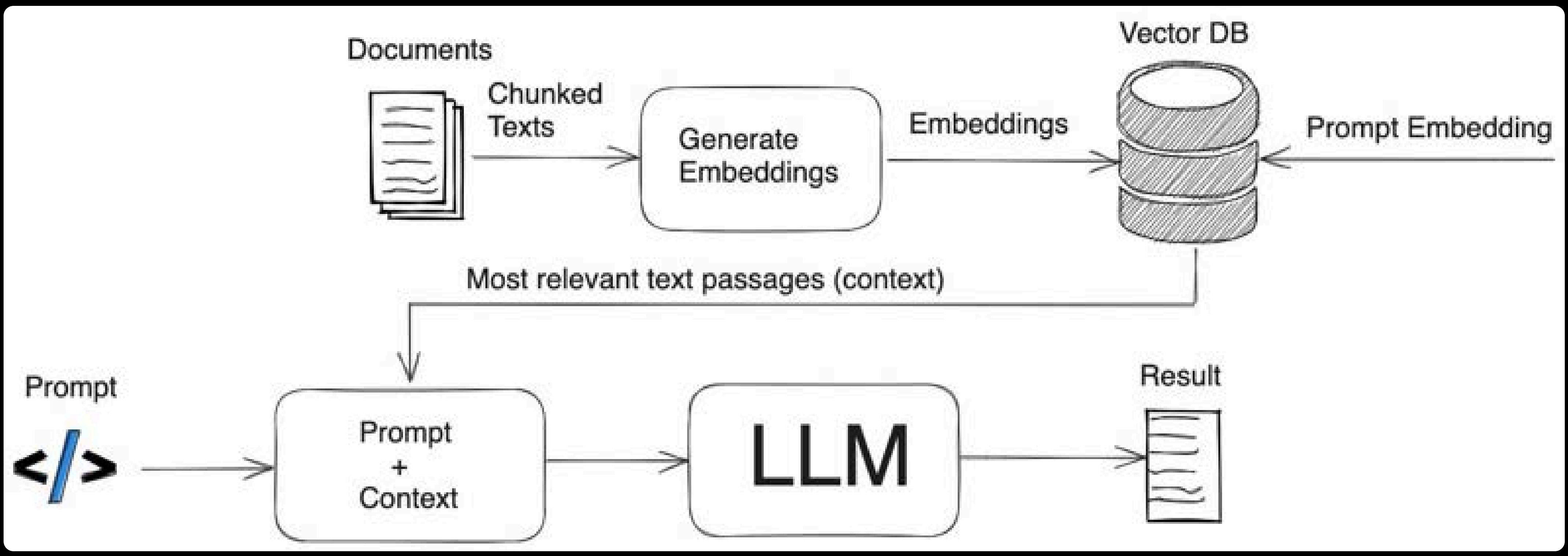
User Contributions to Memorization

- Deploying organizations can inadvertently contribute to memorization through further
  - fine-tuning
  - employing RAG
  - custom instructions and cross-session memorization

Fine-tuning of LLMs



Retrieval Augmented Generation (RAG)



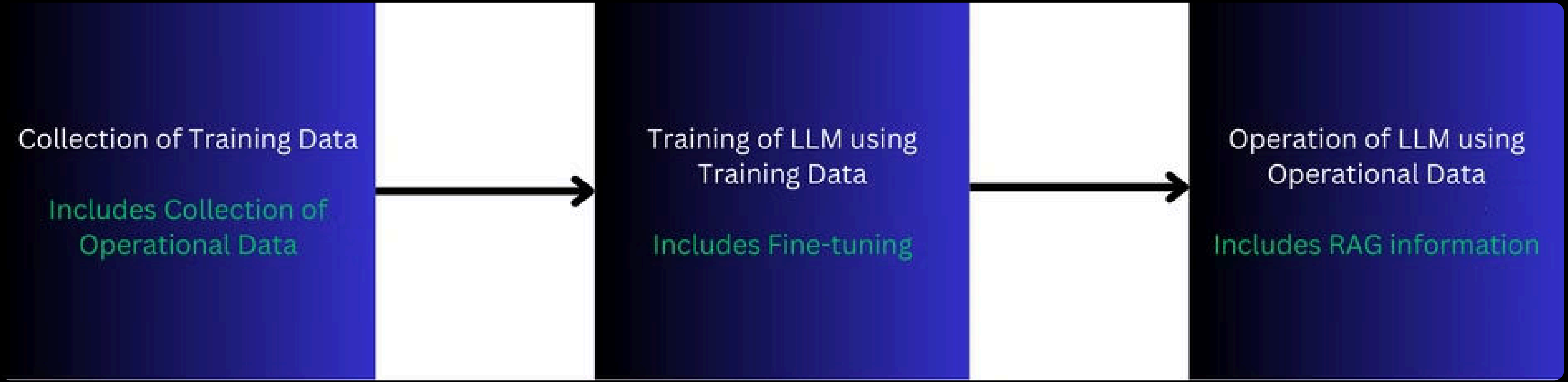
Does not add data to the LLM itself, but their inherent data can still be extracted from the system in an attack.

Custom Instructions and Cross-session memorization may well use a similar process of embeddings to provide more context.

Processing of Personal Data

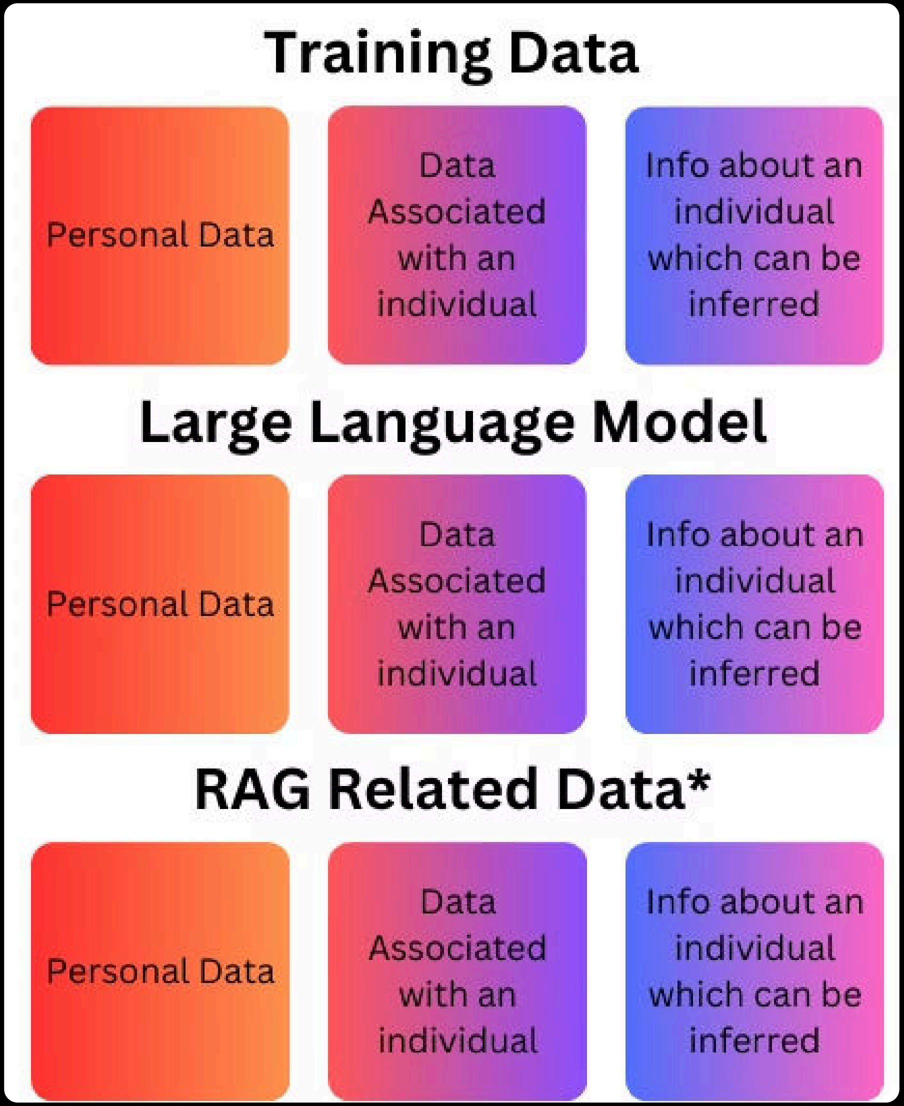
AI Act looks at AI Systems; GDPR looks at processing of personal data

Distinct phases of processing from GDPR perspective:


Collection of Training Data
Includes Collection of Operational Data
Training of LLM using Training Data
Includes Fine-tuning
Operation of LLM using Operational Data
Includes RAG information

We focus here on the training and operational processing.

Personal Data Occurrences



Training Data

Personal Data	Data Associated with an individual	Info about an individual which can be inferred

Large Language Model

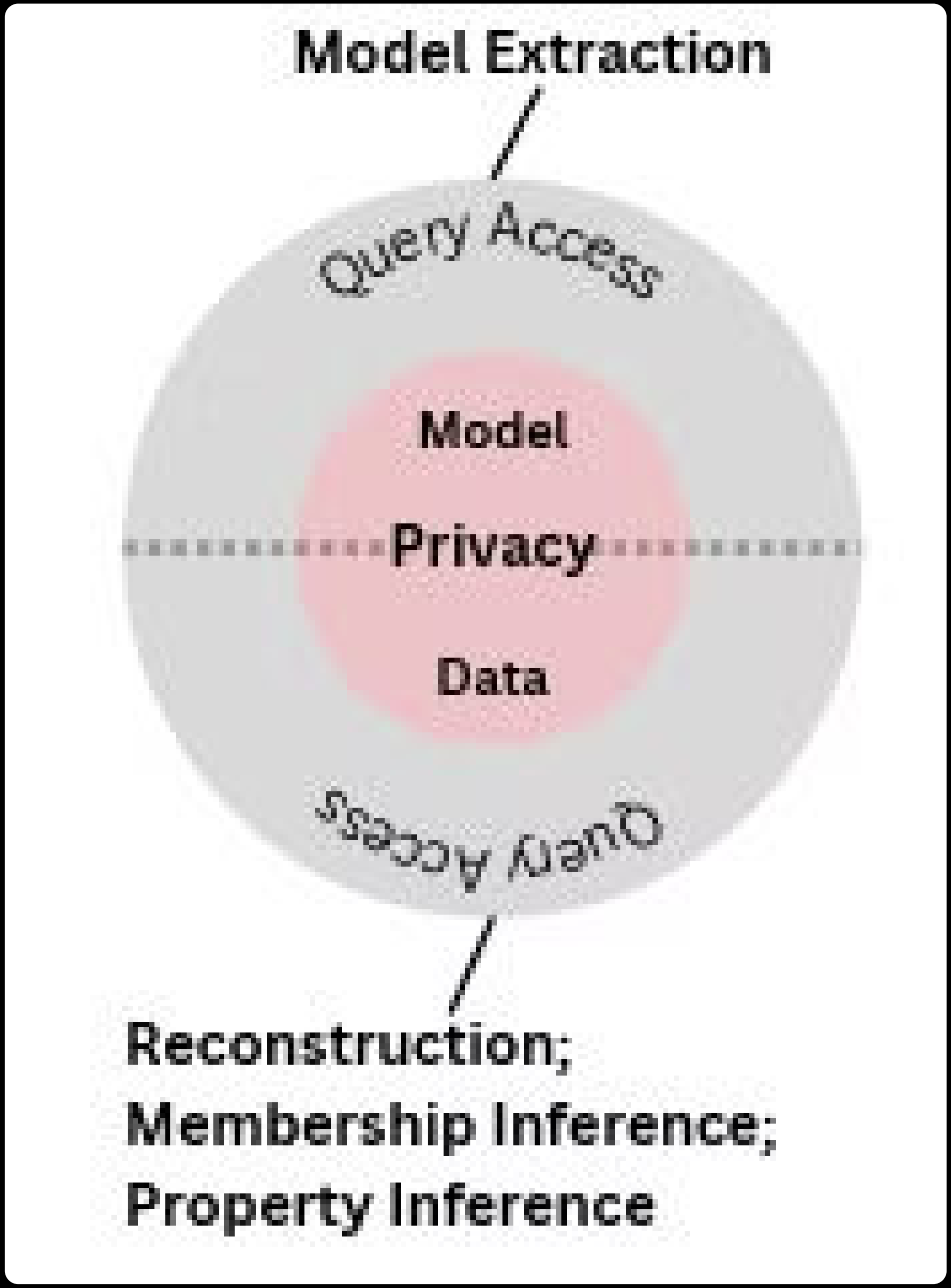
Personal Data	Data Associated with an individual	Info about an individual which can be inferred

RAG Related Data*

Personal Data	Data Associated with an individual	Info about an individual which can be inferred

Inherent Risks of Data Extraction

- Existing methods pose real risks of extracting personal data from LLMs.



- These risks may increase with advancements in extraction technologies and may decrease with defensive techniques

GDPR Compliance and Personal Data in LLMs

Anonymization

- GDPR's Recital 26 is about anonymization
    - The principles of data protection should not apply to anonymous information
    - To determine whether a natural person is still identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly.
- A risk-based approach to re-identification is emerging in the EU.
    - **The definition of 'anonymization' is changing in the EU: Here’s what that means (iapp.org)**

Anonymization Guidelines and LLMs

- Anonymization legal guidelines could be applied to re-identification of personal data in LLMs
- Legal Tests and guidelines for a) isolating individuals, b) linking records, and c) inferring information
  - wp216_en.pdf (europa.eu)

When may Re-identification happen with LLMs?

Test 1

it is still possible to single out an individual

Test 2

It is still possible to link records relating to an individual

Test 3

Information concerning an individual can still be inferred

Other Recognised Factors

Recital 26 of the GDPR includes other objective factors:

1. costs of
2. the amount of time required for identification,
3. taking into consideration the available technology at the time of the processing and technological developments.

Guidance here suggests to consider all relevant contexts: data nature, control and security measures, sample size, public information availability, and data release to third parties (limited vs. unlimited).

Applying Risk Based Legal Tests to Re-identification of Personal Data in LLMs and Key Mitigations

When may Re-Identification happen with LLMs?

Test 1

it is still possible to single out an individual

Depending on the LLM, its size and training data, an LLM may memorize personal data and it is possible to isolate records which identify an individual in the data set.

More likely, observed in studies

Test 2

It is still possible to link records relating to an individual

This association in LLMs is weaker than memorization, but still occurs

This threshold may well be satisfied for some LLMs in practice if a known individual can be linked to information in the LLM.

Also observed in studies

Test 3

Information concerning an individual can still be inferred

Possibly also for unstructured text in LLMs

e.g. a portion of the text refers to the custody of the kids, which may, in the context, allow to infer the marital status

Also observed, but difficult to systematically identify inferences

Considerations in Applying the Legal Tests

- Extraction difficulty varies: enhanced by data processing (e.g., deduplication, obfuscation) but eased by cheap, unauthorized methods.
- The consideration of technological advances conclusions as to feasibility of extraction difficult to assess.
- Single Resolution Board v EDPS: Re-identification risk must consider the potential external access, not just creators.
- Assumptions about users' data, resources, and expertise are unclear; but expect attackers with high capabilities.

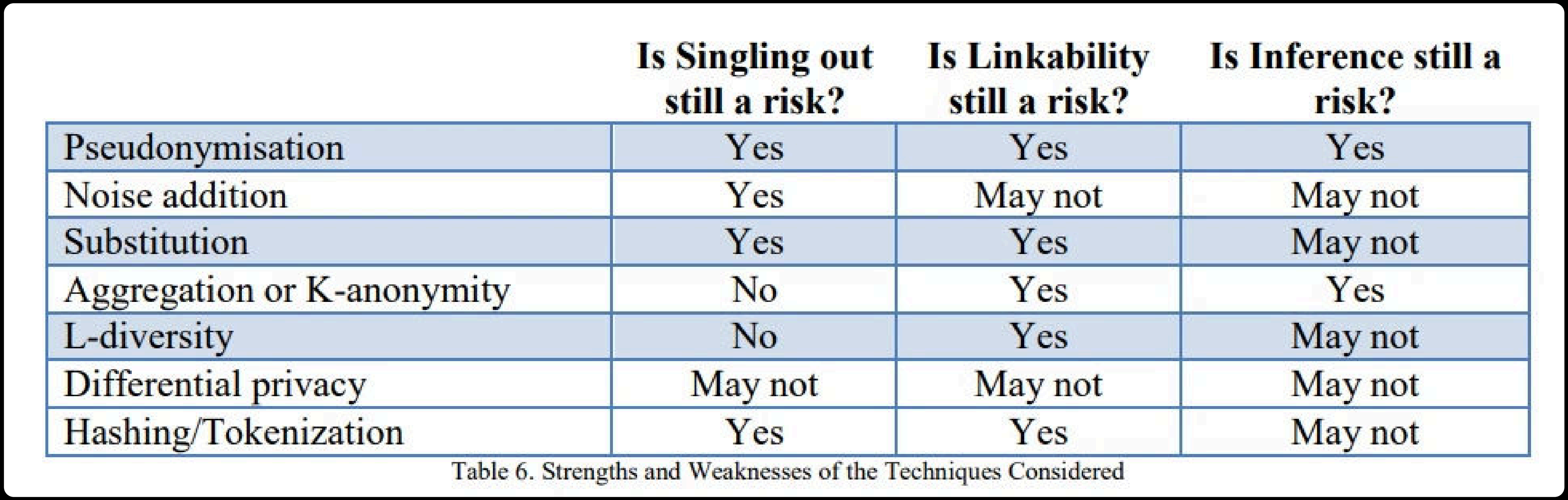
A study showed $200 could enable comprehensive personal data extraction.

Ultimately, organizations must evaluate, directly or through their suppliers, if a Large Language Model (LLM) processes personal data by memorization, linkage, or inference and the feasibility of extraction.

The Role of Pseudonymization and Security and in the Application of these Tests

Pseudonymization

For anonymization, pseudonymization and obfuscation of personal data do seem to be measures that would make re-identification less likely in practice and are considered as part of the tests.



	Is Singling out still a risk?	Is Linkability still a risk?	Is Inference still a risk?
Pseudonymisation	Yes	Yes	Yes
Noise addition	Yes	May not	May not
Substitution	Yes	Yes	May not
Aggregation or K-anonymity	No	Yes	Yes
L-diversity	No	Yes	May not
Differential privacy	May not	May not	May not
Hashing/Tokenization	Yes	Yes	May not

Table 6. Strengths and Weaknesses of the Techniques Considered

Practical Limitations

However, pseudonymization measures are limited because:

1. personal data is not easily identifiable in unstructured text data
2. inference requires additional context to determine if the information should be redacted.

Still need to see then whether personal data can be re-identified per legal tests.

Note that these are techniques applied to data. However similar techniques can be applied during model learning e.g. noise addition during learning, differential privacy stochastic gradient descent.

Security

As per anonymization, looking at all the context means including security in determining the residual risk. For LLMs, this should logically include, in assessing the risks of re-identification, two security strategies:

1. Security Measures against Extraction Attacks

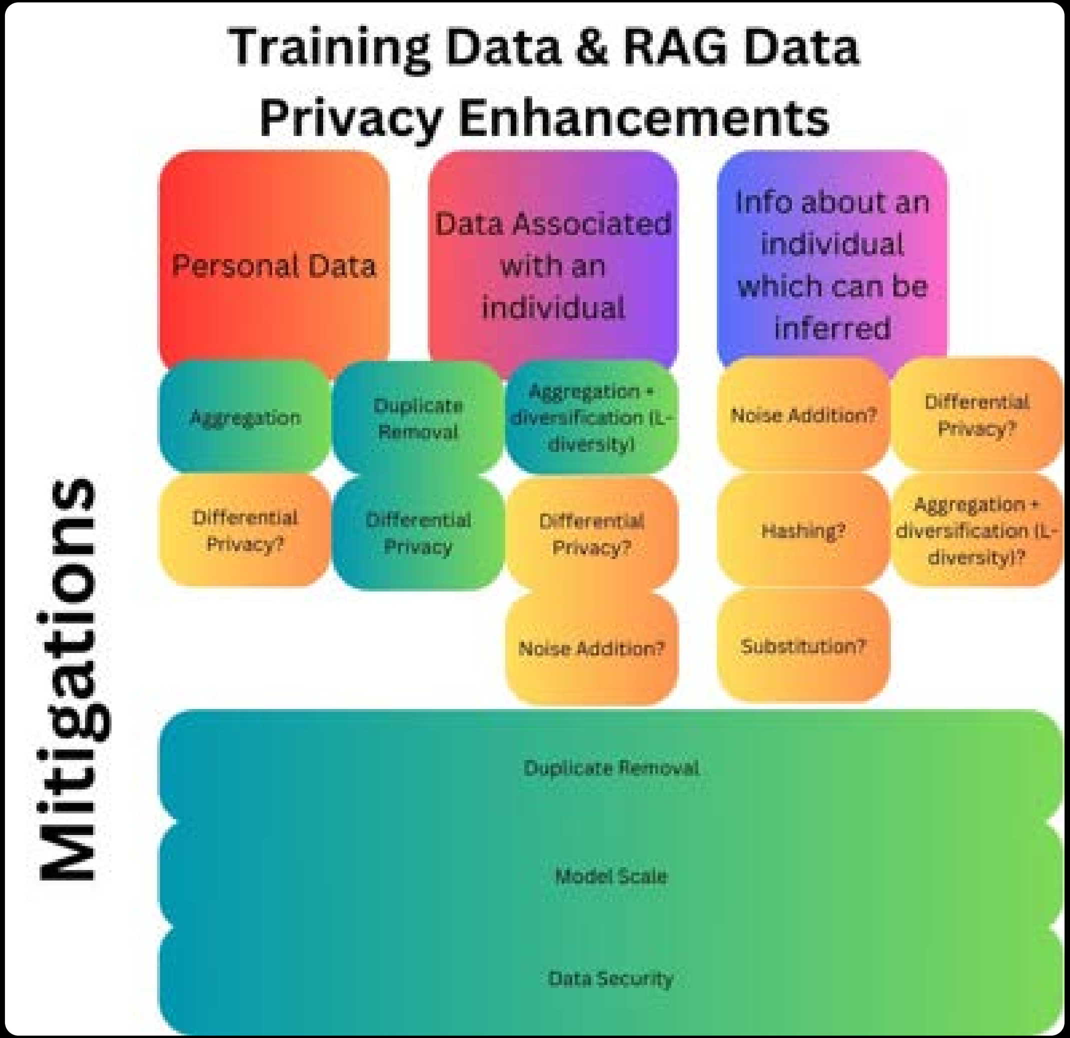
Other mitigation techniques against model extraction, such as limiting user queries to the model, detecting suspicious queries to the model, or creating more robust architectures to prevent side channel attacks exist in the literature.

However, these techniques can be circumvented by motivated and well-resourced attackers and should be used with caution.

2. Security Measures where there has been any Pseudonymization of Private Input/Training Data

E.g. RAG data, whether the tokenization list is kept secure. Note that RAG involves embeddings or transforming of the data into a vector space which does not generalize the personal data. However before usage, the RAG data could be obfuscated, tokenized etc.

Implementing privacy techniques and security measures on both the model and training data can mitigate risks of re-identification.



Training Data & RAG Data Privacy Enhancements

Mitigations

Personal Data

- Aggregation
- Differential Privacy?
- Duplicate Removal
- Differential Privacy

Data Associated with an individual

- Aggregation + diversification (L-diversity)
- Differential Privacy?
- Noise Addition?

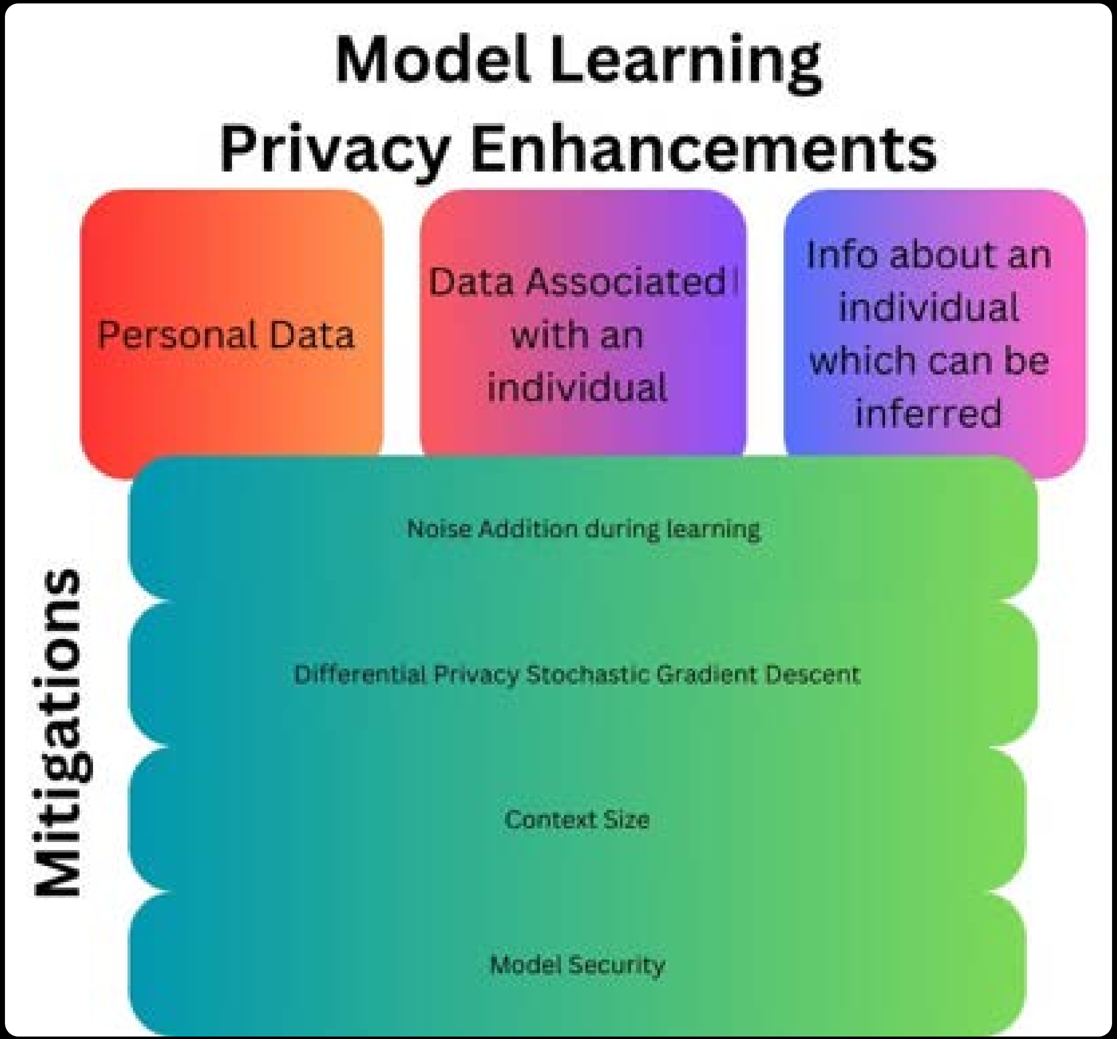
Info about an individual which can be inferred

- Noise Addition?
- Hashing?
- Substitution?
- Differential Privacy?
- Aggregation + diversification (L-diversity)?

Duplicate Removal

Model Scale

Data Security


Model Learning
Privacy Enhancements
Personal Data
Data Associated with an individual
Info about an individual which can be inferred
Mitigations
Noise Addition during learning
Differential Privacy Stochastic Gradient Descent
Context Size
Model Security

In conclusion, measures can be taken to limit the risks of re-identification and hence, depending on the model and measures taken, may serve to make the risks or re-identification of personal data from LLMs low.

A comprehensive risk assessment which includes the risk of re-identification seems necessary.

GDPR Compliance Consequences

- If there is personal data per the re-identification tests, compliance issues with GDPR arise for the Model, not just the processing of training data or operational data

So if there are Re-identifiable personal data, the further considerations under GDPR include:

GDPR Requirement	Processing of Training Data and Operational Data	Processing of Model Data
Legal Basis	Consent or legitimate interest are problematic in many instances	Also problematic , like copyright
Individual Rights	Individuals have rights over their data, including access, correction, and deletion	Difficult to isolate and modify individual data points
Minimisation	Only the minimum amount of personal data necessary for the purpose (of training a model), problematic	Minimisation of e.g. memorization in the model - Linked to purpose for which personal data is processed in use of the model, problematic
Transparency	Organizations must be clear and open with individuals about how their personal data is used and processed, problematic	Also problematic in terms of model usage

cont.

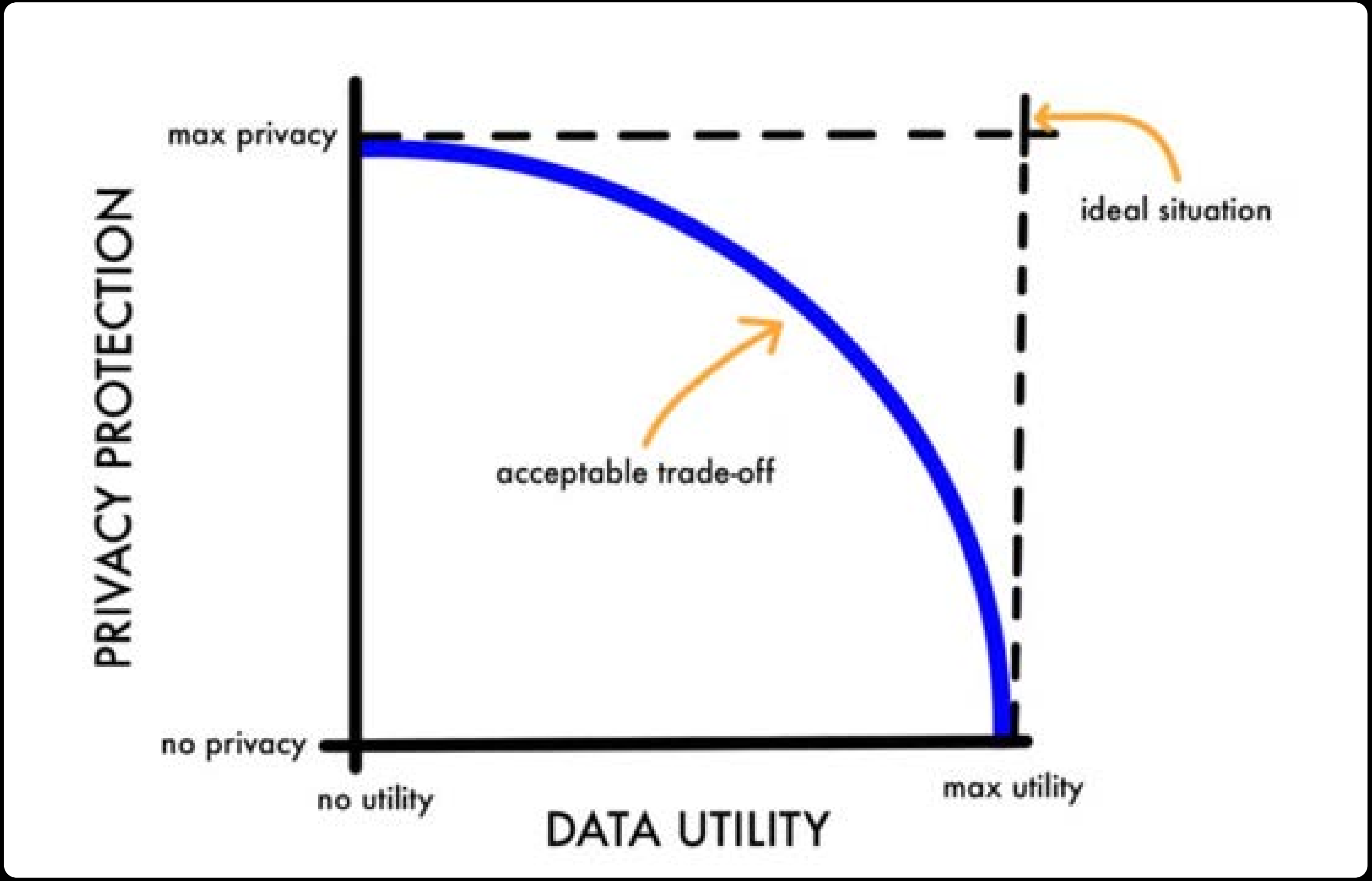
GDPR Requirement	Processing of Training Data and Operational Data	Processing of Model Data
Data Protection by Design and Default	Organizations must integrate data protection measures into their processing activities and systems from the start, ensuring that by default only necessary data is processed.	This requirement is focused on model development, not usage
Data Security	Must be securely stored and protected	Novel attacks possible, directly on the model or via the training data
International Transfers	Transfers of personal data outside the European Economic Area must ensure an adequate level of data protection and comply with strict conditions for transfer.	May also be subject to adequacy requirements e.g. for transfers or usage in other countries

Compliance Obligations for LLM Deployers

- Organizations which are influencing how and why the LLMs are used are then (joint) data controllers. Influence in particular on:
  - the use of personal data in the LLM for specific purposes
  - the use of personal data in operational data, further training data and RAG data
- Joint responsibility does not equate to equal responsibility. GDPR requires joint controllers to transparently define and the essence of the arrangement shall be made available to the data subject.
- Each controller has a duty ensure that the data is not further processed in a manner incompatible with the purposes for which they were originally collected by the controller sharing the data.
  - Nigh impossible to communicate the essence of the arrangement to all possible data subjects. Need for reasonableness, especially for pseudonymized data where there already is a curtailment of data subject rights.

Performance vs. Pseudonymization Trade-off vs. Fundamental Capabilities of LLMs

- Techniques like differential privacy may reduce memorization risks without impairing LLM performance.
- However, there's a delicate balance between data protection and LLM utility.



Note memorization can also hurt model performance

A Fundamental Capability of LLMs

- **"It's 'impossible' to create ChatGPT without copyrighted content" - OpenAI**
- **Is it impossible to for ChatGPT to work well without memorization?**
- "For certain types of tasks it is desirable that the model remembers verbatim text. E.g., reviews of a book can greatly benefit from verbatim quotes, and likewise news articles about political speeches."
- "Model performance and alignment. In order to answer questions about the world, LLMs need to memorize facts about the world. Even when facts are retrieved from external sources the LLM needs to have enough world knowledge to make sense of those facts. Likewise, there is information that is structurally indistinguishable from PII, but meant for public use and useful for the model to memorize. Examples for this are the phone numbers of emergency services or the support email address of a company."

Pseudonymizing the Training and Operational Data Before Training the Model

- Article 4(5) GDPR introduces pseudonymization as the:
- processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

Does not seem to apply to LLMs (no separation and security of raw personal data).

Nevertheless if LLMs are e.g. trained on non-public data, then:

Controllers are exempt from complying with Articles 15–20 of the GDPR, which cover data subject rights like data rectification. However other obligations under the GDPR, (see 8 above) are not curtailed

but any not pseudonymized data identifying, inferring or associated personal data still needs to be rectified upon request

What about Pseudonymization Techniques in the Model Training itself?

Not clear. If there is no additional information is kept, then this situation doesn't fit Article 4(5) which talks only about separation.

Note that the data in the model is still regarded as personal data!!

Take Aways:

- LLMs memorize!
- Organizations must evaluate, directly or through their suppliers, if a Large Language Model (LLM) processes personal data by memorization, linkage, or inference.
- Following a risk-based approach akin to legal standards for anonymization, without adequate safeguards, LLMs pose re-identification risks, making GDPR applicable not only to the processing of training data but also to the LLM itself.
- Implementing privacy techniques and security measures on both the model and training data can mitigate risks but might compromise LLMs' key functionality of memorization.
- Pseudonymization, as defined by GDPR, requires data separation, a criterion not met by public LLMs, though it may be applicable for specific models, potentially easing some GDPR obligations towards data subjects. Will likely need to be reconsidered as a concept, not least with respect to pseudonymization of models themselves.

Tony Hibbert

AI Governance Expert

Simplifying AI Governance

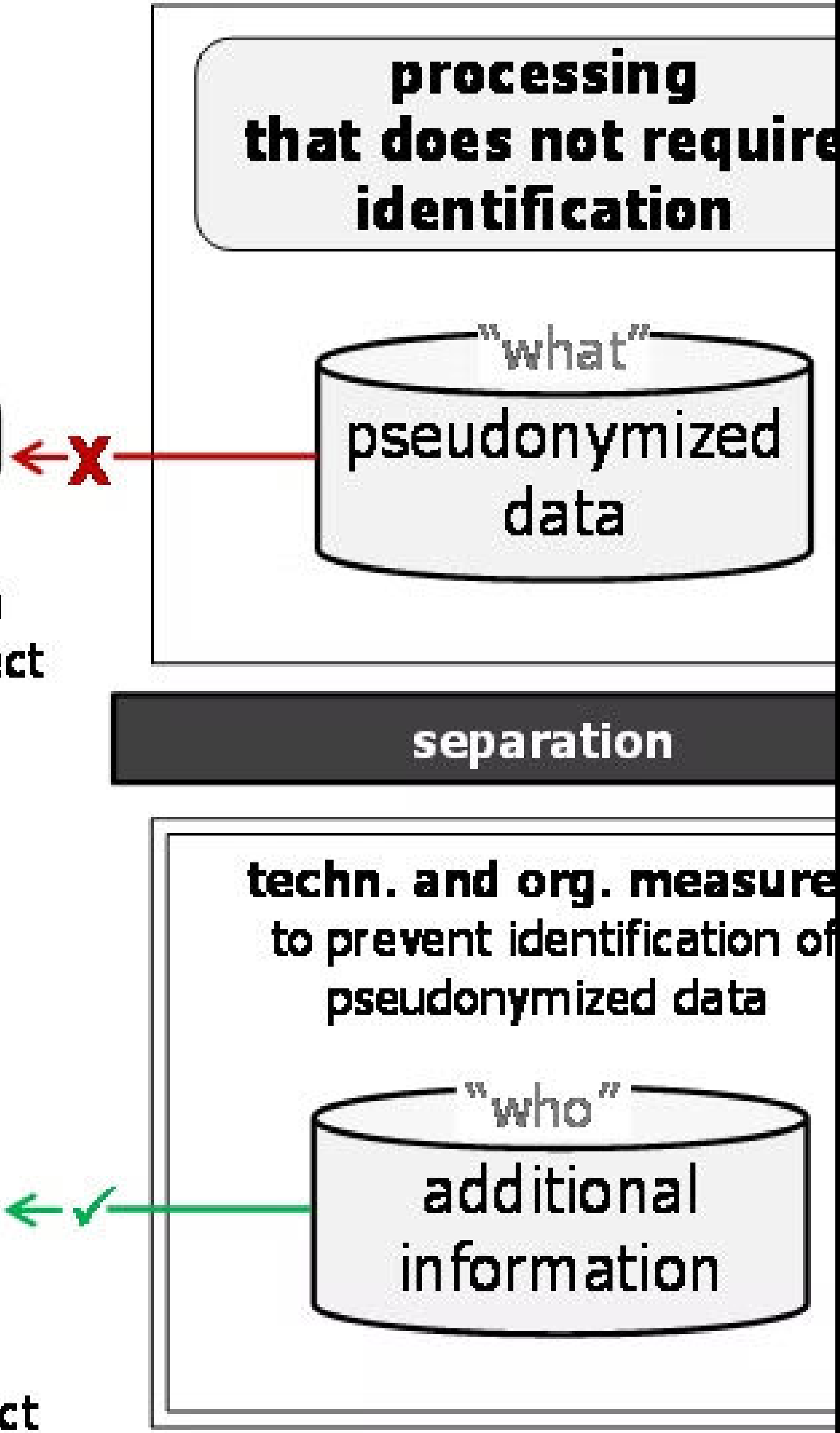
antony_hibbert@outlook.com

https://www.linkedin.com/in/antonyhibbert-ai-governance-expert/

Antony Hibbert

Experienced AI Governance Expert | Navigating the path to responsible AI for fin...

Extra Slides



Pseudonymization Challenges

Pseudonymization is a method under the GDPR:

- Article 4(5) GDPR introduces pseudonymization as the:

"processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

Applying Article 4(5) to LLMs:

- Training data must be kept separate.
- Data from the Internet used in raw form without separation → not considered pseudonymized under GDPR.
- ChatGPT's use of internet data isn't kept separate, risking re-identification.
- Therefore, it doesn't meet the criteria for pseudonymized personal data.

The (Limited) Benefits of Pseudonymization under the GDPR

GDPR Requirement	Processing of Model Data	Prior Pseudonymization of Training or Operational Data
Legal Basis	Also problematic , like copyright	Requirement remains Article 6(4)(e) may help allow for the processing of pseudonymized data for uses beyond the purpose for which the data was originally collected.
Individual Rights	Difficult to isolate and modify individual data points	*Controllers do not need to provide data subjects with access, rectification, erasure or data portability if they can no longer identify a data subject.* but any not pseudonymized data identifying, inferring or associated personal data still needs to be rectified upon request
Minimisation	Minimisation of e.g. memorization in the model - Linked to purpose for which personal data is processed in use of the model, problematic	Does not negate the requirement to limit the amount of data collected. Compliance with data minimization is achievable but challenging for LLMs.
Transparency	Also problematic in terms of model usage	Transparency still problematic Should include the fact that their data will be anonymized

cont.

GDPR Requirement	Processing of Model Data	Prior Pseudonymization of Training or Operational Data
Data Protection by Design and Default	The requirement is focused on model development, not usage	Pseudonymisation is consistent with data protection by design But difficult on the scale of massive unstructured data sets, so depends on training data and measures actually taken
Data Security	Novel attacks possible, also via training data	Improves security of personal data
International Transfers	May also be subject to adequacy requirements	May assist transfers in terms of supplementary measures