CLAIMS

1. A computerized method comprising:

(a) performing a deterministic rule-based search, in a plurality of stored documents, for Personally Identifiable Information (PII) data-items; and

(b) responsive to the deterministic rule-based search indicating that a particular document contains a PII data-item,:

(b1) extracting a textual snippet from said particular document, wherein the textual snippet surrounds said PII data-item,

(b2) adding an entirety of said particular document to a Machine Language classification dataset, that is utilized for training a Machine Learning (ML) engine to classify documents as either containing PII or non-containing PII, and

(b3) adding said textual snippet, and not the entirety of said particular document, to a second dataset utilized for training a Large Language Model (LLM) to find PII data-items in documents for performing Named Entity Recognition (NER) in said documents.

2. The computerized method of claim 1, further comprising:

(c1) performing tokenization of the textual snippet;

(c2) constructing labels for tokens generated by the tokenization of step (c1);

(c3) aligning labels created in step (c2), with sub-word tokens;

(c4) fitting said LLM.

3. The computerized method of claim 2, further comprising:

continuously monitoring a data repository of a Protected Entity for newly-created documents;

automatically performing Machine Language classification of said newly-created document, to determine via Machine Learning whether said newly-created document contains PII or does not contain PII;

if, and only if, the Machine Learning classification indicates that said newly-created document contains PII, then: (I) performing LLM analysis of said newly-created document for finding labels for NER processing, and (II) performing NER processing of said newly-created document based on LLM-determined labels.

4. The computerized method of claim 2, further comprising:

continuously monitoring a data repository of a Protected Entity for recently-modified documents;

automatically performing Machine Language classification of said recently-modified document, to determine via Machine Learning whether said recently-modified document contains PII or does not contain PII;

if, and only if, the Machine Learning classification indicates that said recently-modified document contains PII, then: (I) performing LLM analysis of said recently-modified document for finding labels for NER processing, and (II) performing NER processing of said recently-modified document based on LLM-determined labels.

5. The computerized method of claim 1, comprising:

protecting a document repository of a Protected Entity from leakage of PII,

by performing said deterministic rule-based search for PII exclusively on-premise at said document repository of said Protected Entity and without uploading documents from said document repository of said Protected Entity to a remote server,

and by training said LLM on document snippets extracted locally from documents of said document repository of said Protected Entity, exclusively on-premise at said document repository of said Protected Entity, and without uploading any document snippets or documents from said Protected Entity to a remote server,

and by prompting said LLM to find labels for NER exclusively on-premise at said document repository of said Protected Entity and without uploading any document snippets or documents or from said Protected Entity to a remote server.

6. The computerized method of claim 5,

wherein said LLM is updated on a remote server, that is remote from said document repository of said Protected Entity, based on federated LLM-based analysis results that were generated locally on-premise at said document repository of said Protected Entity and at a plurality of other, separate, document repositories of said Protected Entity and that are collected by and merged at said remote server.

7. The computerized method of claim 1,

wherein the deterministic rule-based search comprises at least:

a first automated rule-based PII search utilizing one or more Regular Expression (RegEX) tools;

a second automated rule-based PII search utilizing one or more particular dictionary files.

8. The computerized method of claim 1, further comprising:

generating an initial estimation that a particular document contains a particular type of PII data-item that is a name of an individual;

performing a negation process that determines that said initial estimation is incorrect even though said particular document indeed contains said name of said individual.

9. The computerized method of claim 1, further comprising:

generating an initial estimation that a particular document contains a particular type of PII data-item, based on at least one of: (i) Regular Expression (RegEx) tools, (ii) pre-defined dictionary files, (iii) Machine Language classification of an entirety of said particular document;

performing a negation process that determines that said initial estimation is incorrect even though said particular document indeed contains a string that, by itself, appears to be a PII data-item.

10. The computerized method of claim 9,

wherein the negation process is based on a feedback loop in which a Protected Entity that controls said data repository provides an indication that the initial estimation is incorrect;

wherein, based on said indication, at least one of the following operations is performed:

(i) modifying one or more parameters of a Machine Language document classification unit that classifies documents as either containing PII or not-containing PII;

(ii) modifying one or more parameters of said LLM.

11. The computerized method of claim 9,

wherein the negation process is based on a feedback loop in which a Protected Entity that controls said data repository provides an indication that the initial estimation is incorrect;

based on said indication, automatically configuring or training said LLM to negate similar further estimations of PII detections that are actually not PII.

12. The computerized method of claim 1,

wherein said LLM comprises a plurality of LLMs;

wherein a first LLM of said plurality of LLMs is trained specifically on, or augmented specifically with, extracted document snippets of said data repository of a particular Protected Entity;

wherein a second LLM of said plurality of LLMs is trained specifically on, or augmented specifically with, entirety of documents of said data repository of said particular Protected Entity.

13. The computerized method of claim 1,

wherein said LLM comprises a plurality of LLMs;

wherein a first LLM of said plurality of LLMs is trained specifically on, or augmented specifically with, extracted document snippets of a particular Protected Entity;

wherein a second LLM of said plurality of LLMs is trained specifically on, or augmented specifically with, entirety of documents of said particular Protected Entity;

wherein at least one of the first LLM and the second LLM is further trained on, or augmented with, a body-of-knowledge that includes at least: a dataset having names and data of customers and clients of said Protected Entity.

14. The computerized method of claim 1,

wherein said LLM comprises a plurality of LLMs;

wherein a first LLM of said plurality of LLMs is trained specifically on, or augmented specifically with, extracted document snippets of a particular Protected Entity;

wherein a second LLM of said plurality of LLMs is trained specifically on, or augmented specifically with, entirety of documents of said particular Protected Entity;

wherein the method comprises selective chaining of LLMs, wherein one of said plurality of LLMs is activated to perform LLM-based processing of a particular document only if another one of said plurality of LLMs has already performed LLM-based processing of said particular document and has already generated a particular type of output.

15. The computerized method of claim 1,

wherein said LLM, that is configured to find PII data-items in documents for NER, is specifically augmented with document snippets that include customer names that match a dataset of names of current customers of a Protected Entity whose documents are in the data repository that is searched for PII.

16. The computerized method of claim 1,

wherein said LLM, that is configured to find PII data-items in documents for NER, is specifically augmented with document snippets that include financial data-items that match a dataset of financial records of a Protected Entity whose documents are in the data repository that is searched for PII.

17. The computerized method of claim 1,

wherein said LLM, that is configured to find PII data-items in documents for NER, is specifically trained on, or augmented with, (i) content of documents of a Protected Entity whose documents are in the data repository that is searched for PII, and also (ii) meta-data of said documents.

18. The computerized method of claim 1,

further comprising:

performing a Named Entity Recognition (NER) process to identify PII data-items in documents of a particular Protected Entity, based on NER tags that were established via an LLM-based processing of at least one of: (i) entirety of documents of said particular Protected Entity, (ii) extracted document snippets that were extracted from documents of said particular Protected Entity;

wherein said LLM-based processing utilizes at least one LLM that was specifically trained on, or augmented with, a body-of-knowledge that includes at least one of: (I) a dataset

of current customers of said Protected Entity, (II) financial records of said Protected Entity, (III) legal records of said Protected Entity.

19. A non-transitory storage medium having stored thereon instructions that, when executed by a machine, cause the machine to perform a method comprising:

(a) performing a deterministic rule-based search, in a plurality of stored documents, for Personally Identifiable Information (PII) data-items; and

(b) responsive to the deterministic rule-based search indicating that a particular document is more likely than not to contain a PII data-items,

(b1) extracting a textual snippet from said particular document, wherein the textual snippets surrounds said PII data-item, and

(b2) adding said textual snippet and said particular document to one or more training datasets utilized for training a Large Language Model (LLM) configured to find PII data-items in documents for Named Entity Recognition (NER) in said documents.

20. A system comprising:

one or more hardware processors, configured to execute code;

associated with one or more memory units, configured to store data;

wherein the one or more hardware processors are configured to perform an automated process comprising:

(a) performing a deterministic rule-based search, in a plurality of stored documents, for Personally Identifiable Information (PII) data-items; and

(b) responsive to the deterministic rule-based search indicating that a particular document is more likely than not to contain a PII data-items,

(b1) extracting a textual snippet from said particular document, wherein the textual snippets surrounds said PII data-item, and

(b2) adding said textual snippet and said particular document to one or more training datasets utilized for training a Large Language Model (LLM) configured to find PII data-items in documents for Named Entity Recognition (NER) in said documents.