Fine-Tuning Large Language Models for Private Document Retrieval

Frank Sommers
Docusure, Inc
USA

Alisa Kongthon
King Mongkut's University of Technology Thonburi
Thailand

Sarawoot Kongyoung
National Electronics and Computer Technology Center
Thailand

NECTEC
June 6, 2024
Bangkok, Thailand

Motivation

There is a large gap between SOTA Document Intelligence (DI) research and practice.

We think the main reason for that gap is due to the presence of sensitive, private documents at most organizations:

SOTA results and techniques, by their very nature, are reported on **public** datasets.

On the other hand, real-world organizations must handle **sensitive, private documents**.

We'd like to share some experience and practical advice on how to bridge this gap.

"Think this is bad? You should see the inside of my head."

What You'll Learn

1. **Document Intelligence (DI)**
    - What it is
    - The current state of the art
    - Data-centric AI

2. **Privacy-Aware DI**
    - Neural network training on documents containing sensitive, private data
    - A framework for document privacy

3. **Fine-tuning DI Models**
    - A framework for fine-tuning pre-trained transformer models
    - Privacy implications

4. **Hands-on fine-tuning practicum**
    - **LayoutLMv3:** SOTA document intelligence model
    - **idefics2**: SOTA multimodal model with strong document task capabilities

About the Presenters

Frank Sommers is co-founder of Docusure, Inc, and founder of Autospaces, two US-based companies focused on document and workflow automation. He has more than 25 years of experience in document management and workflow automation in the financial services industry. His interests include functional programming, parallel and distributed systems, privacy-preserving data management, and training and deploying neural network-based system for production environments. He is most recently co-author of *Programming in Scala, 5th Edition*, a book dedicated to the Scala programming language.

Alisa Kongthon

Sarawoot Kongyoung

A Practical Problem with Broad Application

Document:

"Written, printed, or electronic piece of information that provides evidence, proof, or support a claim, idea, or decision" (*)

- **Written documents**: reports, letters, memos
- **Printed documents**: books, articles, brochures
- **Electronic documents**: digital files, emails, web pages

We won't cover:

- **Multimedia documents**: images, audio, video files

(*) http://meta.ai, accessed on 5/6/2024

Documents

- Not only text
- vs "Document" in the NLP sense
- Text + Layout + Visual
- 2-dimensional
- 1 or more pages
- Heterogenous page sizes
- Figures, tables, images, etc
- Rich layout with semantics
- Internal document structure
- Heterogenous languages
- Heterogeneous media (paper, electronic)

vs Structured data:

- JSON, XML
- Software source code
- RDF triples, Semantic Web

Business Documents

"*Document AI, or **Document Intelligence**, is a relatively new research topic that refers to techniques for automatically **reading**, **understanding**, and **analyzing business documents**. Understanding business documents is a very challenging task due to the diversity of layouts and formats, poor quality of scanned document images as well as the complexity of template structures.*"

Lei Cui, Microsoft Research

Early Foundations (1800s to 1990s)

- **Optical character recognition** (OCR)
  - Gustav Tauschek (1929)
  - IBM
  - Ray Kruzweil (1978)
    - Omni-font OCR
- **Document imaging and management**
  - Enabled by inexpensive storage, 1980s
- **Information retrieval**
  - Keyword-based text retrieval
  - Search, categorization, indexing
  - Boolean queries
  - Vector space models

Gustav Tauschek's Reading Machine
Computer History Museum
https://history-computer.com/technology/optical-character-recognition/

Towards Document Understanding (1990s to 2000s)

- **Structured document analysis**
  - Layout and structure
  - Zoning, segmentation
  - Use of HMMs to identify logical document sections
- **Content extraction and classification**
  - Naive Bayes
  - Support Vector Machines
  - Decision trees to classify documents based on content
- **Early NLP integration**
  - Named-entity recognition and semantic indexing

Machine Learning and AI (2000s to 2010s)

- **Advanced NLP**
  - More computing power + larger datasets.
  - Topic modeling (e.g., Latent Dirichlet Allocation, LDA)
  - Deep syntactic parsing
  - Sentiment analysis
- **Deep learning**
  - CNNs and RNNs
  - Transformers applied classification, sentiment analysis, and entity recognition
- **End-to-end learning**
  - Fewer handcrafted features, raw data directly influences model decisions
  - Greater accuracy and utility

SOTA: 2020s to Present

- **BERT and Transformer-based models**
  - Improved understanding of context and relationships in documents
  - Expanding uses:
    - Summarization, translation, document question answering
- **Multimodal document processing**
  - Integration of text, layout, and images
  - Holistic document understanding
- **Integration with business processes**
  - Take business action on document information
  - Integration with Robotic Process Automation (RPA)
  - Model ensembles and agentic systems
- **Open-Source / Open-Weight Models**

Document Tasks

Task		Examples
Document Classification	Categorize documents into predefined classes Multipage-documents	• Email classification (spam, categories) • Business document stipulations • Legal document classification (deposition supporting documents, etc).
Named Entity Recognition (NER)	Identify and classify named entities within a document (person names, business-specific entities)	• Information extraction: Extract entities and provide them as input to downstream business processes) • Content filtering • Content retrieval
Document Summarization	Generate concise summary of longer documents while retaining key information	• Legal document summarization • Academic research assistance • News aggregation
Visual Question Answering from Documents (VQA)	Answer questions posed in natural language by understanding the textual and visual content of documents	• Interactive document analysis • Customer support (e.g, complaint letters) • Information extraction

Document Tasks

Task		Examples
Information Retrieval from Document Collections	Find and retrieve relevant information from large document collections based on natural language queries	• Search engines • Academic research • Legal / corporate decision support • Technical document / customer support
Document Translation	Translate text in documents from one language to another	• International business and law • Internationalization of documentation / customer service • Marketing communications
Document Layout Analysis	Analyze and understand the structural layout of documents, including text, images, and tables	• Automated document formatting • Content reflow for different devices • Accessibility improvements
Table, form, and chart understanding	Extract information from tables, forms, diagrams	• Financial document analysis • Health-care data processing • Survey data analysis
Document Generation	Create new documents based on specified criteria	• Legal document drafting • Personalized marketing communications • Automated report generation

Benchmark Datasets

Dataset	Tasks	Description	Top Models	Metric
Truth Tobacco Industry Documents https://www.industrydocuments.ucsf.edu/tobacco/about/overview/	• Document classification • Many subsets	• 95,690,703 pages in 16,213,203 documents		
IIT-CDIP Complex Document Information Processing https://ir.cs.georgetown.edu/downloads/sigir06cdipcoll_v05-with-authors.pdf	• Document classification • Layout analysis • Forms understanding	• 11 million document images • Can be split into 42M images • Subset of Legacy Tobacco Documents	Used as pretraining E.g., for LayoutLMv3	
RVL-CDIP https://adamharley.com/rvl-cdip/	• Document classification • Multi-Modal Classification • Layout Analysis	• Subset of IIT-CDIP Test Collection • 320,000 grayscale training, 40,000 validation, 40,000 test • Scanned documents, • 16 classes (letter, resume, email, form, etc) • 25K images per class	EAML Cross-Modal DocFormerBASE LayoutLMV3Large LiLT[EN-R]BASE	***Accuracy*** 97.70% 97.05% 96.17% 95.93% 95.68%

Benchmark Datasets

Dataset	Tasks	Description	Top Models	Metric
FUNSD Form Understanding in Noisy Scanned Documents https://guillaumejaume.github.io/FUNSD/	• Entity labeling • For relation extraction	• 199 fully annotated scanned forms in	LayoutMask ERNIE-Layoutlarge LayoutMask (base) GeoLayoutLM LayoutLMv3Large GeoLayoutLM LayoutLMv3Large TPP BROS LayoutLMv2Large	***F1*** 93.20 93.12 92.91 92.86 92.08 89.45 80.35 79.20 77.01 70.57
FUNSD-r	Named Entity Recognition	• 199 docs, entities in 3 categories	TPP LayoutLMv3 LayoutMask	***F1*** 80.40 78.77 77.10

Benchmark Datasets

Dataset	Tasks	Description	Top Models	Metric
CORD v2 Consolidated Receipt Dataset for Post-OCR Parsing	Key information extraction	• 11,000 Indonesian receipts from restaurants and shops	GeoLayoutLM LayoutLMv3Large LayoutMask (large) LILT LayoutLMv2Large	***F1*** 97.97 97.46 97.19 96.99 96.01
SROIE	Key information extraction, OCR	• 1000 scanned receipt images and annotations	LayoutLMv2 large LayoutLMv2BASE	***F1*** 97.8 96.25
CORD-r	Named Entity Recognition	• 999 doc samples, labeled entities in 30 categories	TPP LayoutLMv3 LayoutMask	***F1*** 91.85 89.34 81.84
DocVQA	Visual Question Answering	• 50,000 questions defined on 12,000 document images	Human SMoLA-PaLI-X Specialist SMoLA-PaLI-X Generalist Qwen-VL-Plus	***ANLS*** 0.981 0.908 0.906 0.9024

Benchmark Datasets

Dataset	Tasks	Description	Top Models	Metric
OCR-VQA	Visual Question Answering	• 207,572 images and question answer pairs • Document content and OCR transcriptions		
TEXT-VQA	Visual Question Answering	• 28,408 images from OpenImages • 45,336 questions, • 453,360 ground truth answers		
MM-Vet	Multimodal, integrated capabilities	• Recognition, OCR, Knowledge, Spatial awareness, Math	GPT-4V GPT-4V-Turbo Qwen-VL-Max Gemini Pro Vision GLM4 Vision InternVL 1.5 Claude 3 Opus LLaVA-NeXT-34B	***GPT-4 score*** 67.7 67.6 66.6 64.3 63.9 62.8 58.1 57.4

Document Models

- **Text-Only Modality**
  - BERT or BERT-derivatives (RoBERTa, DistilBERT, etc)
  - Sentence Transformers
  - TfidfTransformer (Scikit Learn)
  - All LLM tools, e.g., GPT-4, LLama, etc.
- **Text + Layout + Vision**
  - Text features using e.g., BERT
  - Layout: Bounding boxes from OCR process
  - Page image
  - E.g., LayoutLM family
- **Vision-Only**
  - Image encoder, text (e.g., JSON) decoder
  - Multi-modal models
  - E.g., GTP-4o, Gemma, idefics2

Document Model Performance

Model	Modality	FUNSD	CORD	RVL-CDIP	DOC-VQA
BERT large (Devlin, et al., 2018)	T	65.6	90.3	89.9	67.5
DiT large (Li at al., 2022)	V			92.7	
Donut (Kim et al., 2022)	V		91.6	95.3	72.1
mPLUG-DocOwl (Ye at al., 2023)	V				62.2
StructuralLM large (Li at al., 2020)	T+L	85.1		96.2	83.9
LayoutLM large (Xu et al., 2020)	T+L	77.9		91.9	
UniDoc (Gu et al., 2021)	V+T+L	87.9	96.9	95.1	
LAMBERT (Garncarek et al. 2021)	T+L	96.1			
TILT large (Powalski et al., 2021)	V+T+L		96.3	95.5	87.1
LayoutLMv2 (Xu et al., 2021)	V+T+L	84.2	96.0	95.6	78.8

Document Model Performance

Model	Modality	FUNSD	CORD	RVL-CDIP	DOC-VQA
LayoutLMv3 large (Huang at al., 2022)	V+T+L	92.1	97.5	95.9	83.4
UDOP (Tang at al., 2023)	V+T+L	91.6	97.6	96.0	84.7
LayoutLLM (Fujitake, 2024)	V+T+L	95.3	98.6	98.8	86.9
GeoLayoutLM (Luo, et al., 2023)	V+T	89.45			

vs. Humans

Preliminary study (Docusure, 2024):

- US financial services company
- Corpus of 17m+ business documents
- 600+ users over 6 years
- Human-only document classification
- **> 11% of documents misclassified**
- 89% human vs 95% $\text{LayoutLMv3}_{\text{LARGE}}$

FBI Fingerprint Archive
Washington, D.C., USA, 1944

Data-Centric AI: Bring Your Own Documents

- SOTA models already produce outstanding results on common tasks
- Benchmark datasets vs your dataset
- How good are these models on your own documents?
- How can you use them on your own dataset?
- **Data-centric vs model-centric**
  - Hold models constant, improve the data

Demo

GPT-4o for Zero-Shot Classification via Public User Interface

Regulatory Compliance

Building with the OpenAI API Platform

The OpenAI Platform allows you to build entirely custom applications. As the developer of your application, you are responsible for designing and implementing how your users interact with our technology. To make this easier, we've shared our Safety best practices, and offer tools like our Moderation Endpoint and customizable system messages.

We recognize that our API introduces new capabilities with scalable impact, so we have service-specific policies that apply to all use of our APIs in addition to our Universal Policies:

1. Don't compromise the privacy of others, including:
    a. Collecting, processing, disclosing, inferring or generating personal data without complying with applicable legal requirements
    b. Using biometric systems for identification or assessment, including facial recognition
    c. Facilitating spyware, communications surveillance, or unauthorized monitoring of individuals

Bring Your Private Documents: Financial Services

AZ-103-ARB 6/30/2020

Retail Installment Contract and Security Agreement

Seller Name and Address	Buyer(s) Name(s) and Address(es)		Summary
CACTUS AUTO SALES, LLC. 9823 N. CAVE CREEK RD. PHOENIX, AZ 85020-1723	LUIS G SANTOS 9502 N 15TH AVENUE 134 PHOENIX, AZ 85021	DENISSE Y GUTIERREZ 9502 N 15TH AVENUE 134 PHOENIX, AZ 85021	No. 3351955H Date 12/29/2022

☐ Business, commercial or agricultural purpose Contract.

Truth-In-Lending Disclosure

Annual Percentage Rate	Finance Charge	Amount Financed	Total of Payments	Total Sale Price
The cost of your credit as a yearly rate.	The dollar amount the credit will cost you.	The amount of credit provided to you or on your behalf.	The amount you will have paid when you have made all scheduled payments.	The total cost of your purchase on credit, including your down payment of
25.99 %	$ 10,652.48	$ 15,090.94	$ 25743.42	$ 4700.00 $ 30443.42

Payment Schedule. Your payment schedule is:

No. of Payments	Amount of Payments	When Payments are Due
53	$ 476.73	Starting 01/28/2023 and every 28th of the month thereafter
1	$ 476.73	Estimated final payment on 06/28/2027
N/A	$ N/A	N/A

Security. You are giving us a security interest in the Property purchased.

Late Charge. If all or any portion of a payment is not paid within 10 days of its due date, you will be charged a late charge of 5% of the unpaid amount of the payment due.

Prepayment. If you pay off this Contract early, you will not have to pay a penalty.

Contract Provisions. You can see the terms of this Contract for any additional information about nonpayment, default, any required repayment before the scheduled date, and prepayment refunds and penalties.

Arizona Used Motor Vehicle Warranty

This section applies to used vehicles only. **The Seller hereby warrants that this Vehicle will be fit for the ordinary purposes for which the Vehicle is used for 15 days or 500 miles after delivery, whichever is earlier, except with regard to particular defects disclosed on the first page of this agreement. You (the purchaser) will have to pay $25.00 for each of the first two repairs if the warranty is violated.**

Attention Purchaser. Sign here only if the dealer told you that this Vehicle has the following problem(s) and that you agree to buy the Vehicle on those terms:

1. _____
2. _____
3. _____

By: _____ Date

By: _____ Date

By: _____ Date

The within instrument or agreement is pledged as collateral to Wells Fargo Bank, N.A., its successors and assigns.

LGSH
DYGT

Retail Motor Vehicle Credit Application

☒ Credit Sale ☐ Lease Application Number: Date: 03/04/2024

Creditor Name and Address: Lobel Financial Corp 7535 E. Hampden Ave., Suite 520 Denver, CO 80231-4844

TYPE OF CREDIT REQUESTED:

☐ Business ☒ Individual ☐ Joint—We intend to apply for joint credit (initials): _____ _____

The words "you" and "your" refer to each person or business submitting this application. The words "we", "us" and "our" refer to the seller and the financial companies to which your application is submitted. The words "married" and "spouse" include registered domestic partners or civil union where applicable.

IMPORTANT APPLICANT INFORMATION: Federal law requires financial companies to obtain sufficient information to verify your identity. You may be asked several questions and to provide one or more forms of identification to fulfill this requirement. In some instances we may use outside sources to confirm the information.

Complete **JOINT APPLICANT'S** section only if application is for joint credit.

(A) Applicant	Applicant Information	(B) Joint Applicant or Other Party
FULL NAME (First, Middle, Last): GUSTAVO GUADAL SALMERON		FULL NAME (First, Middle, Last):
GOV'T ID TYPE: Driver license	GOV'T ID NO.: 170563729 / GOV'T ID ISSUED BY: CO	GOV'T ID TYPE / GOV'T ID NO. / GOV'T ID ISSUED BY
GOV'T ID ISSUE DATE: N/A	GOV'T ID EXP. DATE: N/A / DATE OF BIRTH: 02/21/2000	GOV'T ID ISSUE DATE / GOV'T ID EXP. DATE / DATE OF BIRTH
SOC. SEC. NO.: 210-02-2305	PRIMARY PHONE ☐ CELL / SECOND PHONE: 970-539-6964 ☒ CELL	SOC. SEC. NO. / PRIMARY PHONE ☐ CELL / SECOND PHONE ☐ CELL
EMAIL ADDRESS		EMAIL ADDRESS
STREET ADDRESS: 1126 26TH AVE / APT#: 204 / HOW LONG? 1 YEARS 4 MONTHS		STREET ADDRESS / APT# / HOW LONG? ___ YEARS ___ MONTHS
CITY: GREELEY / STATE: CO / ZIP: 80634		CITY / STATE / ZIP
MAILING ADDRESS (if different from Street Address) / APT# / MONTHLY RENT OR MORTGAGE PAYMENT		MAILING ADDRESS (if different from Street Address) / APT# / MONTHLY RENT OR MORTGAGE PAYMENT
CITY / STATE / ZIP		CITY / STATE / ZIP
RESIDENTIAL STATUS ☐ OWN ☐ RENT ☐ WITH RELATIVES ☐ WITH FRIENDS ☐ OTHER / LANDLORD/MORTGAGE		RESIDENTIAL STATUS ☐ OWN ☐ RENT ☐ WITH RELATIVES ☐ WITH FRIENDS ☐ OTHER / LANDLORD/MORTGAGE
LANDLORD PHONE / PREVIOUS ADDRESS (if less than 2 yrs at current address): 1126 26 TH AVE / APT# 204		LANDLORD PHONE / PREVIOUS ADDRESS (if less than 2 yrs at current address) / APT#
CITY: GREELEY / STATE: CO / ZIP: 80634 / HOW LONG? ___ YRS ___ MOS		CITY / STATE / ZIP / HOW LONG? ___ YRS ___ MOS
CURRENT EMPLOYER: MASTER CLEANIG / GROSS MONTHLY SALARY: 2000.00		CURRENT EMPLOYER / GROSS MONTHLY SALARY
CURRENT EMPLOYER'S ADDRESS / CITY / STATE		CURRENT EMPLOYER'S ADDRESS / CITY / STATE
ZIP / WORK PHONE: 303-753-6084 / HOW LONG? 0 YRS 4 MOS / OCCUPATION/JOB TITLE: General Labor		ZIP / WORK PHONE / HOW LONG? ___ YRS ___ MOS / OCCUPATION/JOB TITLE
PREVIOUS EMPLOYER (if less than 2 yrs at current job) / GROSS MONTHLY SALARY		PREVIOUS EMPLOYER (if less than 2 yrs at current job) / GROSS MONTHLY SALARY
PREVIOUS EMPLOYER'S FULL ADDRESS / PHONE		PREVIOUS EMPLOYER'S FULL ADDRESS / PHONE
SECONDARY EMPLOYER NAME (if applicable) / SECONDARY EMPLOYER ADDRESS		SECONDARY EMPLOYER NAME (if applicable) / SECONDARY EMPLOYER ADDRESS
CITY / STATE / ZIP / GROSS MONTHLY SALARY		CITY / STATE / ZIP / GROSS MONTHLY SALARY
SECONDARY EMPLOYER PHONE / HOW LONG? ___ YRS ___ MOS / OCCUPATION/JOB TITLE		SECONDARY EMPLOYER PHONE / HOW LONG? ___ YRS ___ MOS / OCCUPATION/JOB TITLE

OTHER INCOME NOTE:*

☐ (A) or ☐ (B) GROSS MONTHLY OTHER INCOME	OTHER INCOME SOURCE	☐ (A) or ☐ (B) GROSS MONTHLY OTHER INCOME	OTHER INCOME SOURCE

* Alimony, child support, or separate maintenance incomes do not have to be revealed unless the applicant wishes to have such sources considered as a basis for repayment of the requested credit amount.

REFERENCE	PHONE	REFERENCE	PHONE
ADDRESS	RELATIONSHIP	ADDRESS	RELATIONSHIP
BANK REFERENCE ☐ CHECKING ☐ SAVINGS		BANK REFERENCE ☐ CHECKING ☐ SAVINGS	

AGREEMENT TO FURNISH INSURANCE

(THIS IS USED AS PART OF MOTOR VEHICLE SALES CONTRACT AND SECURITY AGREEMENT)

Sale Date: 2/28/2024 Stock# 48361

CUSTOMER INFORMATION		SELLER INFORMATION
Name(s)	Sesar Nohe Camacho Mayoral	Own a Car of Fresno
Address	170 Club Dr Apt B	5788 N. Blackstone Ave
City, State Zip	Lemoore, CA 93245	Fresno, CA 93710
Phone	(559) 773-9777	(559) 435-3600

The undersigned Buyers(s) agree(s) to provide and furnish Insurance Policy, covering property which is the subject of Security Agreement dated February 28th, 2024 _____.

The Vehicle referred to herein is described as follows:

Year 2021 Make Nissan Model Versa Body Car VIN# 3N1CN8EV4ML890337

Miles 40517

Such insurance policy must be delivered to the Seller within 7 days from the date hereof, and if Seller does not receive such Policy by the time stated. Seller may (but is not required to) procure insurance of the kind and type agreed to be furnished under the terms of the above mentioned Security Agreement.

INSURANCE COMPANY		INSURANCE AGENT	
Name	National General	Name	
Policy No.	2022044704	Address	
Exp. Date	8/28/24	City, State Zip	

LIENHOLDER INFORMATION			
Loss Payee	LOBEL FINANCIAL	Assigned to	LOBEL FINANCIAL
Phone	(714) 995-3333	Phone	(714) 995-3333
Address	Po Box 3000 Anaheim CA 92803		

Forward Binder or Declaration page to:

Fax (559) 435-3620 OR Email financedepartment@ownacar.com

FIRE, THEFT, COMPREHENSIVE – DEDUCTIBLE $ 500 **COLLISION – DEDUCTIBLE $** 500

In the event I fail to furnish a valid insurance policy, or written evidence, from an insurance company for comprehensive and deductible collision insurance coverage, within the time specified from the above date, I hereby agree to pay to Seller or its assignees any earned premium for any policy they may have to place for the above described vehicle. In accordance with repayment procedures established under California Civil Code Section 2962.8.

I / we further agree to assume forthwith any and all responsibility for damage to the property referred to above or resulting from the use, maintenance or operation thereof, and agree to hold Seller and its assignee free of any loss, claim, or liability resulting from any damage to said property or from the use, maintenance or operation thereof.

NOTICE TO BUYER: This agreement does not authorize the ordering of **Public Liability or Property Damage Insurance** Any insurance ordered by the financial institution will cover loss of or damage to the above described vehicle only and will not include **Public Liability or Property Damage Insurance**

WARNING: IT IS YOUR RESPONSIBILITY UNDER CALIFORNIA LAW TO OBTAIN LIABILITY INSURANCE OR BE SUBJECT TO PENALTIES FOR VIOLATING SECTION 16020 OF THE VEHICLE CODE, WHICH MAY INCLUDE LOSS OF LICENSE OR A FINE. THE INSURANCE ACQUIRED BY THE LIENHOLDER DOES NOT PROVIDE LIABILITY COVERAGE AND DOES NOT SATISFY YOUR RESPONSIBILITY UNDER CALIFORNIA LAW

Buyer _____ Date 2/28/2024 Dealer/Representative _____ Date 2/28/2024

Co Buyer _____ Date

ASNSoftware reserves all rights and provides this form without express or implied warranty as to content, fitness or legal compliance ASN-0807014
Copyright © ASNSoftware Inc. All rights reserved.

Bring Your Private Documents: Financial Services

SEARCH FORTUNE SIGN IN Subscribe Now

Home News Tech Finance Leadership Well Recommends Fortune 500

TECH · SAMSUNG

Samsung threatens to fire employees if they leak data to A.I. chatbots like ChatGPT

BY NICHOLAS GORDON
May 2, 2023 at 6:25 PM GMT+7
Updated May 4, 2023 at 4:37 PM GMT+7

Samsung said employees misusing ChatGPT risk "termination," according to an internal memo.
JUNG YEON-JE—AFP/GETTY IMAGES

Public Benchmarks vs (Your) Private Data

Application Area	Private documents
Financial Services	Loan applications, bank statements, tax records
Healthcare	Patient medical records
Legal	Litigation documents, patent applications, client correspondence, court decisions
Government	Permits, petitions, applications, ID cards, tax records
Travel	Hotel registration forms, visa applications
Real Estate	Real estate loan documents, lease documents,
HR	Employee records, contracts
Industrial	Trade secrets, company plans, designs

Jurisdictional Data Partitioning

Data Privacy Laws

- **Ensuring consent**
- Ensuring that personal data is protected in a data transfer
- Most countries have some form of it:
  - **Thailand:**
    - Personal Data Protection Act (PDPA). Consent for data transfer. Can be transferred under specific conditions.
  - **European Union:**
    - General Data Protection Regulation (GDPR)
    - Restricts transfer of personal data outside the EU, unless there are specific mechanisms
  - **United States:**
    - California Consumer Privacy Act (CCPA) and California Privacy Rights Act (CPRA)
      - Restricts data transfer, 3rd-party sharing
    - Health Insurance Portability and Accountability Act (HIIPA)
      - Restricts transfer of health data
  - **Canada**:
    - Personal Information Protection and Electronic Documents Act (PIPEDA)
      - Must obtain consent for data transfer

Data Residency Laws

- **Must keep data within a jurisdiction**
- Full or partial data localization rules in many countries
  - **India:**
    - **Proposed Personal Data Protection Bill (PDPB).** Sensitive data to be stored in India, but can processed abroad under specific conditions.
  - **Indonesia:**
    - **Government Regulation No 71:** Public service data to be stored and managed within country
  - **Vietnam:**
    - **Law on Cybersecurity:** Certain types of personal data be stored within country
  - **South Korea:**
    - **Act on the Promotion of IT Network User and Information Protection:** Data localization especially for financial data
  - **Germany:**
    - **Federal Data Protection Act (BDSG):** Certain sectors can mandate local storage
  - **UAE:**
    - Several free zones have their own localization rules
  - **China:**
    - **Cybersecurity Law, Personal Information Protection Law (PIPL).** Critical infrastructure operators must keep collected data inside country
  - **Etc.**

Local LLM Inferencing

- **Open-source models**
  - Open-source model implementations
  - Pre-trained models: open-weights
  - Commercial-friendly license vs research / non-commercial use
  - GPT-4: Not open-source
  - Llama3: Open-source, commercial-friendly LLM
  - LLava, Idefics2: Open-source, commercial-friend multi-modal models
- **Local inference tools**
  - Ollama
  - Llama.cpp
  - HuggingFace Transformers with API endpoint (FastAPI, etc)

Ollama: Run LLMs Locally

- Mac, Linux, Windows
- Server or laptop / personal computer
- GPU support
- Bundles model weights, config, data into a Modelfile
  - Like "Docker" client for model inferencing
- Out-of-the-box experience
- Support for wide range of models

Demo

Running local idefics2

LLM-Based Automation Trends
with
Private Data

Private Data: Required Documents for a Car Loan

APP#: **3919772**
STATUS: **eContract Ready for Funding**
REMINDER: **No Reminder**
ASSIGNED TO: NO ONE

RECEIVED: **5/22/2024, 9:48 AM**
DECISION: **APPROVED**
AGE: **8d**

DEALER: **DERBY MOTORS CORPORATION**
DLR PHONE: **(956) 592-2286**
VEHICLE: **2016 JEEP RENEGADE**

NOTES | DOCUMENTS | STIPULATIONS | DISCREPANCIES | CONTACTS | TASKS | VEHICLE | ACTIVITY LOG

STIP	DOCUMENT	FROM	VERIFIER REVIEW	MANAGER REVIEW
Approval	Received		Accepted	Accepted
Odometer Statement (262 for CA)	Received	Dealer	Accepted	Accepted
Application for Title	Received	Dealer	Accepted	Accepted
Report of Sale	Received	Dealer	Accepted	Accepted
Book sheet (AT)	Received		Accepted	Accepted
Photos of Car	Received		Accepted	WAIVED
Insurance Dec page or ID card	Received	Signer	Accepted	Accepted
DL or ID (B1)	Received	Signer	Accepted	Accepted
DL or ID (B2)	Received	Co-Signer	Accepted	Accepted
Verifications form (T&E/VOR/VOI)	Received		Accepted	Accepted
Proof of Residence	Received	Signer	Accepted	Accepted
Melissa Data	Received		Accepted	Accepted
References	Received	Signer	Accepted	Accepted
VOE Form	Received		Accepted	Accepted
Employer Listing or Business Licei	Received		Accepted	Accepted
YTD Paystub (W2 income)	Received	Signer	Accepted	Accepted
POI calculation worksheet	Received		Accepted	Accepted
CLARITY REPORT	Received		Accepted	Accepted
SOCIAL SECURITY CARD BYR 2	Received		Accepted	Accepted
texas eligibility	Received		Accepted	Accepted
texas eligibility byr 2	Received		Accepted	Accepted
Social Security Card	Received	Signer	Accepted	Accepted
Proof of Residence B2	Received	Co-Signer	Accepted	Accepted
e-Contract	Received		Accepted	Accepted

Autospaces Apptracker software

Document Management: Car Loan

Document processing:

- Document classification (document type)
- Attribute extraction
- Document comparison (fact resolution)
- Document generation (e.g., loan approval letter)
- Attribute entry into Line-of-business system

Document Management: Car Loan

Document Management: Car Loan

Document Management: Car Loan

Document Intelligence

Human-in-the-loop: Document comparison (fact resolution), Integration with line-of-business system: RPA (robotic process automation)

Generative Process Automation

End-to-End Neural Network-Based Automation

Human in the loop

Loan Application → **Generated Decision**

TX-103-ARB 9/15/2016

Motor Vehicle Retail Installment Contract and Security Agreement

Seller Name and Address
LINARES AUTO CENTER
2019 W. HIGHWAY 83
MCALLEN, TX 78501-7003

Buyer(s) Name(s) and Address(es)
DALIA PAT DAVILA
27783 FM 1561
SAN BENITO, TX 78586

Summary
No. 3914306G
Date 05/21/2024

Retail Motor Vehicle Credit Application

✘ Credit Sale ☐ Lease Application Number: Date: 05/21/2024
Creditor Name and Address: Lobel Financial Corp 7330 San Pedro Avenue, Suite 402 San Antonio, TX 78216-6234
TYPE OF CREDIT REQUESTED:
☐ Business ✘ Individual ☐ Joint—We intend to apply for joint credit (initials): _____

The words "you" and "your" refer to each person or business submitting this application. The words "we", "us" and "our" refer to the seller and the financial companies to which your application is submitted. The words "married" and "spouse" include registered domestic partners or civil union where applicable.
IMPORTANT APPLICANT INFORMATION: Federal law requires financial companies to obtain sufficient information to verify your identity. You may be asked several questions and to provide one or more forms of identification to fulfill this requirement. In some instances we may use outside sources to confirm the information.

Complete **JOINT APPLICANT'S** section only if application is for joint credit.

(A) Applicant — Applicant Information	(B) Joint Applicant or Other Party
FULL NAME (First, Middle, Last): DALIA PAT DAVILA	FULL NAME (First, Middle, Last):
GOV'T ID TYPE: Driver license	GOV'T ID TYPE:
GOV'T ID ISSUED BY:	GOV'T ID ISSUED BY:
GOV'T ID NO.:	GOV'T ID NO.:
GOV'T ID ISSUE DATE: N/A	GOV'T ID ISSUE DATE:
GOV'T ID EXP. DATE: N/A	GOV'T ID EXP. DATE:
DATE OF BIRTH: 05/21/1975	DATE OF BIRTH:
SOC. SEC. NO.: 632-16-2718	SOC. SEC. NO.:
HOME TEL. NO.: 956-244-7284 ☐ CELL	PRIMARY PHONE: ☐ CELL
BUSINESS PHONE: 956-244-7284 ✘ CELL	SECOND PHONE: ☐ CELL
EMAIL ADDRESS:	EMAIL ADDRESS:
STREET ADDRESS: 27783 FM 1561	STREET ADDRESS:
APT#	APT#
HOW LONG? 16 YEARS 3 MONTHS	HOW LONG? YEARS MONTHS
CITY: SAN BENITO	CITY:
STATE: TX	STATE:
ZIP: 78586	ZIP:
MAILING ADDRESS (if different from Street Address):	MAILING ADDRESS (if different from Street Address):
MONTHLY RENT OR MORTGAGE PAYMENT	MONTHLY RENT OR MORTGAGE PAYMENT
RESIDENTIAL STATUS: ☐ OWN ☐ RENT ☐ WITH RELATIVES ☐ WITH FRIENDS ☐ OTHER	RESIDENTIAL STATUS: ☐ OWN ☐ RENT ☐ WITH RELATIVES ☐ WITH FRIENDS ☐ OTHER
LANDLORD MORTGAGE	LANDLORD MORTGAGE
LANDLORD PHONE	LANDLORD PHONE
PREVIOUS ADDRESS (if less than 2 yrs at current address)	PREVIOUS ADDRESS (if less than 2 yrs at current address)
CURRENT EMPLOYER: SELF EMPLOYED	CURRENT EMPLOYER:
GROSS MONTHLY SALARY: 3000.00	GROSS MONTHLY SALARY:
WORK PHONE: 000-000-0000	WORK PHONE:
HOW LONG? 2 YRS 3 MOS	HOW LONG?
OCCUPATION/JOB TITLE	OCCUPATION/JOB TITLE

CALIFORNIA DRIVER LICENSE
DL I1234568
EXP 08/31/2018
CLASS C
END NONE
LN CARDHOLDER
FN IMA
2570 24TH STREET
SACRAMENTO, CA 95816
DOB 08/31/1977
RSTR NONE
0831977
SEX M
HAIR BRN
EYES BRN
HGT 5'-05"
WGT 125 lb
ISS 09/30/2010
DD 09/30/2010622121FD/18
Usa Template

LOAN APPROVAL LETTER

December 23, 2004

RE: Loan Approval for: **<Salutation 2>**
Delivered by fax to:**<B-Agent> <B-Fax>/ <B-Company>**
<L-Agent> <L-Fax>/ <L-Company>

Dear <B-Salutation> and <L-Salutation>

Based on an underwriter's review I am pleased to issue loan approval on behalf of <Salutation 2> for the residential real estate purchase of <Property Address> <City>, <State>, <Zip> based on a qualifying interest rate of <App Rate 1>.

The interest rate used for qualifying is higher than the available market interest rate at the time of approval. Doing so provides for a more conservative loan approval and provides flexibility for the prospective buyer(s) and seller(s) in the event that interest rates increase prior to contract acceptance. Upon receipt of a fully executed contract our client will have the option of protecting (locking-in) an interest rate at current market rates.

Final loan approval is contingent upon a satisfactory appraisal, survey, title commitment, and homeowner's insurance. If loan closing takes place after <Document AU Expiration> are-verification of credit, assets, liabilities, employment, and income may be required.

Please contact my office at <MY:b-Office Ph> if you have any questions or need assistance.

Best Regards,

<MY:Contact>
<MY:b-Title>

End-to-End Neural Network-Based Automation

- **Security and Privacy Implications:**
  - Document Intelligence tasks
  - + Generative tasks
  - + (Symbolic / rules-based) constraints
- **Domain Adaptation**
  - Fine-Tuning with private data

Domain Adaptation:
Neural Network Fine-Tuning
with
Private Data

Your Own Data: Fine-Tuning (Open Source) Models

Yann LeCun • Following

VP & Chief AI Scientist at Meta

3d •

Llama-3 fine-tuning FTW!

Fine-tuned open source models outperform generalist proprietary models on any specific task.

LinkedIn, 5/29/2024

Example Domain Adaptation: Classification

7 Document classes

Document class	Label
Arizona Contract	AZCONTRACT
California Contract	CACONTRACT
Nevada Contract	NVCONTRACT
Buyers Guide	BUYERSGUIDE
Credit Application	CREDITAPP
Contract Disclosure Form	PRECONTRACT
Title Application Form	TITLE

Fine-Tuning Infrastructure

Privacy: Encryption at Rest

Privacy: Encryption End-to-End

Encryption

sanook

ค้นหาใน ข่าว

หน้าแรก เรื่องฮอต ข่าว ข่าวบันเทิง กีฬา รถยนต์ ไอที ดูดวง วัยรุ่น ไลฟ์สไตล์ เศรษฐกิจ คลิป

ข่าว อาชญากรรม การเมือง ต่างประเทศ สังคม บันเทิง ภูมิภาค กีฬา เศรษฐกิจ ข่าวทั้งหมด

หลุดมาจากไหน? สาวซื้อขนมโตเกียว เจอถุงทำจากเอกสาร รพ. ข้อมูลเจ้าหน้าที่ครบ

07 พ.ค. 67 (09:31 น.) | พิมพ์

Encryption

Data (model inputs):

- Document images
- Labels
  - Some sensitive, e.g., for NER
- Layout data (e.g., bounding boxes)

ML Pipeline Steps	Directly on Encrypted Data	Decryption for Processing
Training and test data quality and preparation	• N/A: Human must look at examples • Synthetic training data	• Training set preparation / labeling tool that supports just-in-time decryption Example: Docugym
Model training	• N/A • Homomorphic encryption	• Datasets and data loaders APIs
Model evaluation		
Using the model to make predictions		

Multiparty Document AI

Multiparty Document AI

Multiparty Document AI

Secure Multiparty Document AI

Secure Multiparty Document AI

Homomorphic Encryption

- Computations carried out on ciphertext
- When the result is decrypted, result is the same (very close to) as if the computation was carried out on plaintext
- Algebraic homomorphism
- Supports both symmetric and asymmetric schemes

diagram idea credit to Pascal Paillier, *http://zama.ai*

Homomorphic Encryption

- Public key to encrypt data
- Private decryption key
- Similar to public key encryption
- Allows multiple sources of encrypted data
- Large-scale multi-party data
- Data encrypted at rest, in transit, during processing

Prediction-as-a-Service

Homomorphic Encryption

- **Rivest, Adleman, Dertouzos, 1978**
  - Privacy homomorphism
  - The encryption functions preserves homomorphism over some operations
  - Perform operation on ciphertext, same operation is performed on the cleartext
  - RSA is partially homomorphic for multiplication:
    - Multiply two ciphertexts, you obtain the encryption of their product
- **Partially Homomorphic Encryption (PHE)**
  - Supports only 1 operation an arbitrary number of times
  - RSA: multiplicative
  - Paillier: Additive
- **Somewhat Homomorphic Encryption (SHE)**
  - Supports both addition and multiplication, but for a limited number of operations
- **Fully Homomorphic Encryption (FHE)**
  - Supports both operations an arbitrary number of times
  - Allows general computations
- **Craig Gentry, 2009**
  - First FHE system (ACM Doctoral Dissertation Award)

Homomorphic Encryption

- **Computation as combination of addition and multiplication**
- **Represent programs as arithmetic circuits**
  - Directed acyclic graph
    - Nodes: Operations (addition and multiplication)
    - Edges: Dataflow
  - Inputs: variables
  - Outputs: results

Boolean circuit

- **Since FHE supports both addition and multiplication, any function can be computed under FHE**

- **Adds noise to the encryption:**

- Addition: Noise additive
- Multiplication: Noise multiplicative

Homomorphic Encryption

- Noise can overwhelm the encrypted data, making decryption impossible
- **Bootstrapping (Gentry, 2009):**
  - Remove the noise
  - Very slow
- Noise budget: Allocate to circuit

Credit: Pascal Paillier
http://zama.ai
and
https://fhe.org/

Homomorphic Encryption: TFHE

TFHE: Fast Fully Homomorphic Encryption over the Torus

- Chillotti, et al, 2016, 2018
- Leading scheme in terms of efficiency
  - **Learning With Errors (LWE)**
  - Hardness, based on lattice structures
  - Linear equations perturbed by some small error
  - Computationally hard
- Polynomial rings for structuring data
- Efficient bootstrapping
  - Bootstrapping at every operation
- Encrypts each bit of data
  - Implement any Boolean circuit
  - Not for large amounts of data encryption
- Adds noise to the databit in a way that the data can still be operated on homomorphically

TFHE vs AES

TFHE	AES
Fully homomorphic, Asymmetric	Symmetric key encryption
Encrypts data bit-by-bit	Block cipher (fixed-size blocks, 128 bits)
Slower, more resource-intensive	GB/s processing
Small amounts of data	GB or TB or greater
Data at rest, transit, and processing	Data in rest and transit
Can compute on encrypted data	Cannot compute on encrypted data

Homomorphic Encryption Libraries

- TFHE
- CONCRETE (TFHE)
- SEAL
- HElib
- Palisade
- Lattigo
- nuFHE
- HEAAN
- FV-NFLib

Standardization efforts:
http://homomorphicencryption.org

CONCRETE-ML: FHE for Machine Learning

- **Compiler-based approach:**
  - Compile a ML model into a FHE circuit
  - Supports variety of ML models, including deep networks.
- **For inferencing on encrypted data**
  - Build model on clear data, compile model so it works on encrypted data

Demo

Encrypted Anonymization with CONCRETE-ML:
https://huggingface.co/spaces/zama-fhe/encrypted-anonymization

Just-in-Time Decryption: Datasets and Data Loaders

Pytorch, Tensorflow, HuggingFace

Dataset

- Isolates loading and processing data (document files) examples
- Load individual data samples
- Load from variety of sources (even synthetic data)
- Perform transformations on data items
- `__len__, __getitem__`

Dataloader

- Load batches
- Shuffling, parallel data loading, etc

Encryption: PyTorch Dataset

```python
class EncryptedPDFDataset(Dataset):
  def __init__(self, file_paths, decryption_key):
      self.file_paths = file_paths
      self.decryption_key = decryption_key
      self.fernet = Fernet(decryption_key)

  def __len__(self):
      return len(self.file_paths)

  def decrypt_pdf(self, file_path):
      # Read the encrypted PDF
      with open(file_path, 'rb') as file:
          encrypted_data = file.read()

      # Decrypt the data
      decrypted_data =
          self.fernet.decrypt(encrypted_data)
      return decrypted_data
```

```python
def __getitem__(self, idx):
    file_path = self.file_paths[idx]
    decrypted_pdf_data = self.decrypt_pdf(file_path)
    # Read the decrypted PDF
    pdf_reader =
        PdfFileReader(io.BytesIO(decrypted_pdf_data))

    # Text of the first page as an example
    text = pdf_reader.getPage(0).extractText()

    # Convert text to tensor
    # Same with image, etc.
    sample = torch.tensor(....) # encode
    return sample
```

Encryption: PyTorch Dataset

```
dataset =
    EncryptedPDFDataset(file_paths,
                        decryption_key)
dataloader =
    DataLoader(dataset,
               batch_size=2,
               shuffle=True)

input_size = 100
output_size = 10
class SuperSimpleModel(nn.Module):
  def __init__(self, input_size, output_size):
    super(SuperSimpleModel, self).__init__()
    self.fc = nn.Linear(input_size, output_size)

  def forward(self, x):
    return self.fc(x)
```

```
# Using the model with encrypted files:
model = SuperSimpleModel(input_size, output_size)
criterion = nn.CrossEntropyLoss()
optimizer = optim.Adam(model.parameters(),
                       lr=0.001)
# Training loop
num_epochs = 5
for epoch in range(num_epochs):
   for batch in dataloader:
       outputs = model(batch)
       targets = # Target labels for the item
       loss = criterion(outputs, targets)
       optimizer.zero_grad()
       loss.backward()
       optimizer.step()
```

Just-in-Time Decryption: Data Labeling

Cannot be outsourced to labeling companies

Domain experts:

- Already have access to sensitive documents
- They know the data and documents the best
- Are busy, can only label a handful of documents
- Not labeling experts
- Can only label a few documents, may make mistakes

Private Document Data Labeling: Docugym

- **Documents stored in secure document repositories**
  - Existing enterprise document silos
  - Existing access control and security
- **Define a corpus**
  - One or more repositories
- **Invite domain experts to label a small number of documents**
- **Classification and entity recognition tasks**
  - Coming: Question answering
  - Classification: E.g., document type, fraud detection
- **Redundant labeling**
  - Specify desired labeling redundancy
  - Identifies disagreement between annotators
  - Reconcile discrepancies
- **Generate model input from labeled corpus**
- Secure training corpus curation
- Open-source

Demo

Private Dataset Curation with Docugym

Privacy: For a Single Example

Privacy: Memorization

Language modeling training objectives

- Maximize the probability of the data in training set X
- Minimize the loss function $\mathcal{L}(\theta) = -\log \prod_{i=1}^{n} f_\theta(x_i | x_1, \ldots, x_{i-1})$
- Memorization is the optimal solution!
  - "The address of Washington, DC., resident Joe Biden is…"
    - *xi:* "1600 Pennsylvania Avenue"
  - "The address of Washington, DC., resident Bob Smith is … "
    - *xi:* "12345 K Street, Apt 888"
  - "Bob Smith's bank account number is… "
- First example is much more likely to be in many training examples

Training data extraction attack

- Reconstruct training data points verbatim (not just in a "fuzzy" way)
- Even if data points appear very few times in the training data
- Important to consider for generative models especially

Privacy: The Not-So-Secret Sharer

Data secrecy:

- Gmail SmartCompose
  - Chen, et al, 2019
  - auto-complete
- Learning occurs on secret data

Privacy as Contextual Integrity:

- Helen Nissenbaum (2004)
- Data is not secret
- Data used outside of its intended context

Privacy: *k*-Eidetic Memorization

Eidetic memory: Photographic memory

Carlini, et al., 2021

Eidetic memorization

- Data that the model memorized that has appeared in only a few training examples
- The fewer training examples, the stronger eidetic memorization

Model Knowledge Extraction

- Model $f(\boldsymbol{\Theta})$ knows string s, if s can be extracted by interacting with the model (black-box interaction)

***k*-Eidetic Memorization**

- Model $f(\boldsymbol{\Theta})$ k-eidetic memorized string s, if s is extractable from $f(\boldsymbol{\Theta})$, and s appears at most in k training examples

A measure of how "strong" memorization is

k-eidetic for *small* (Bob Smith) ks vs *large* ks (Joe Biden)

Privacy: *k*-Eidetic Memorization

- Memorization does not require overfitting
- Larger models memorize more data
  - Of 600,000 generated samples, about 0.1% contained memorized text (Carlini, et al.)
- Fine-tuning may cause the pretrained LM to "lose" some of it's eidetic memory
- Extent to which fine-tuning causes a pretrained model to "forget", active research
- Fine-tuning can introduce it's own privacy leakages

Privacy: Anonymization

Insufficient anonymization

- Direct identifiers: name, social security number
- Quasi-identifiers: date of birth, zip code, gender, even car owned
- Removing direct identifiers is insufficient if quasi-identifiers remain intact
- Re-identification:
  - > 63% of US population can be re-Id'd from 3 pieces of data:
    zip code, date of birth, gender
    (Philippe Golle, 2006)

Linkage attacks

- Anonymized datasets can be combined with other datasets to re-identify individuals
- 2006 Netflix movie ratings, combined with IMDb comments dataset
- 2017, Strava exercise data (27 million fitness tracking devices):
  - Revealed location of military bases, secret facilities
    (https://www.zdnet.com/article/strava-anonymized-fitness-tracking-data-government-opsec/)

Privacy: Anonymization

Reduced utility

- Example: Legal court decision publishing
- < 10% of Bavarian court decision documents are public (private communication)
    - Axel Adrian, et al., Friedrich-Alexander Universitat (FAU),
    - Bavarian State Ministry of Justice
- Legally meaningful anonymized legal cases for ML model training
- Develop models to automatically anonymize large volumes of legal cases (*https://www.str2.rw.fau.de/en/honorarprofessor/forschungsprojekte/#collapse_0*)

Privacy: Anonymization

- The police can give a fine or a warning.
- Bob is the only African American young man living in a small Alpine village.
- Bob is also the only person owning a red sports car in the village.
- Bob claims that he always receives a fine, but others driving at the same speed often get a warning.

Privacy: Membership Inference

Data (D) (medical dataset)

```
(Bob)
DOB: xxxxxx
Name: xxxxx
Depression: True
```

```
(Alice)
DOB: xxxxxx
Name: xxxxx
Depression: False
```

Mechanism (M)

$$M(D) => output$$

To what extent does Bob's example being in D influences the Output?

M: Entire ML pipeline:

- Collection of training examples
- Example labeling
- Model building
- Using the predictions

Privacy: Membership Inference

M**:* mechanism, ***D: dataset

Use M to predict if Bob has depression:

A $P(M(D) = Bob\ has\ depression) = 0.55$

Augment the dataset with Bob's record, and predict again:

B $P(M(D + Bob) = Bob\ has\ depression) = 0.57$

B' $P(M(D + Bob) = Bob\ has\ depression) = 0.80$

Adding Bob to the dataset leads to significantly different outcome in **B'**

It's likely that Bob has depression

M leaked information about Bob

Could also be the model was re-trained in a completely different way

Privacy Loss

No example influences the output at all

- Total privacy for examples in *D*
- We learn nothing from *D*

Some loss of privacy

- When we learn something from dataset *D*, we leak information about *D*

  "The mean age of the participants in the study is 42"

  How much information did we reveal about the participants?

  How likely is Bob in *D*, given this information?

- The more a single item in *D* influences the Output, the greater the privacy loss

Goal

- The model should make a (nearly) similar inference about Bob whether or not Bob opts into the dataset
- Limit the extent to which Bob's record can affect the model's predictions
- Quantify the privacy loss

Quantifying the Privacy Loss

- Cannot compare the respective model weights directly
- **Compare the probability of observing the weights**
- Information theory (Shannon):
  - How much do we learn from the occurrence of an event?
  - Low probability event, high information
  - High probability event, low information

$$A = M(D) = Bob\ has\ depression \qquad I(A) = -\log P(A)$$

$$B = M(D + Bob) = Bob\ has\ depression \qquad I(B) = -\log P(B)$$

Privacy loss:

- Information we gained by adding Bob's record to D
- Difference between A and B

Quantifying the Privacy Loss

$$A = M(D) = Bob\ has\ depression \qquad I(A) = -\log P(A)$$

$$B = M(D + Bob) = Bob\ has\ depression \qquad I(B) = -\log P(B)$$

Privacy loss:

- Difference between A and B

$$I(A) - I(B) = (-\log P(A)) - (-\log P(B))$$

$$(-\log P(A)) - (-\log P(B)) = -\log P(A) + \log P(B)$$

$$-\log P(A) + \log P(B) = \log P(B) - \log P(A)$$

$$\log P(B) - \log P(A) = \log \frac{P(B)}{P(A)} \quad = log \frac{(P(M(D + Bob)) = Bob\ has\ depression)}{P(M(D) = Bob\ has\ depression)}$$

B: $\log \frac{0.57}{0.55} = 0.0357$ **B':** $\log \frac{0.8}{0.55} = 0.375$ 10X the privacy loss!

Goal: Limit the privacy loss to an upper bound

ε-Differential Privacy

Let ***D*** and ***D`*** differ in one data record

Let ***S*** be a set of outcomes from mechanism ***M:***

$$privacy\ loss = \log \frac{P(M(D) \in S)}{P(M(D') \in S)}$$

If $privacy\ loss$ is 0, $P(M(D) \in S)$ and $P(M(D') \in S)$ are the same.
We learned nothing!
Instead, specify an upper bound for the privacy loss, $\boldsymbol{\varepsilon}$:

$$\log \frac{P(M(D) \in S)}{P(M(D') \in S)} \leq \epsilon$$

Privacy budget:

- Trade-off how much we learn vs the privacy of the individual (training example)

Dwork, et al. (2006)

ε-δ Differential Privacy

Tune ε to bound the privacy of ***D*** with regard to ***M***:

$$\log \frac{P(M(D) \in S)}{P(M(D') \in S)} \leq \epsilon$$

$$P(M(D) \in S) \leq e^{\epsilon} P(M(D') \in S)$$

Bound the privacy loss for any dataset ***D*** and ***D'*** for any outcomes

M is ε-differentially private

- For every possible model output, the probability of observing the model weights lies within the bounds of e^{ϵ} for both ***D*** and ***D'***
- Condition holds no matter how data is processed

Allow for some small failure probability, δ:

M is ε-δ differentially private

$$P(M(D) \in S) \leq e^{\epsilon} P(M(D') \in S) + \delta$$

Private Training vs Private Prediction

Black-box access to LLM:

- API-based access
- DP Next-Token Prediction (Flemings, et al., 2024)
  - Sampling from the output distribution of the LLM
  - ε-Mollifiers: To smooth the loss function

White-box access to LLM:

- Weights are available for inspection
- Open-weights models
- Mobile inferencing

Differentially Private Training

- **Compute exact answer**
- **Add noise randomly sampled from the Laplace or the Gaussian distributions**
  - The smaller the privacy budget, add more noise (more privacy)
  - The higher the privacy budget, add less noise (less privacy)
  - Noise spreads the range of possible values
  - Smoothes probabilities, prevents one outcome from being much more likely
- Example:
  - Bob's age is 35
  - Add some noise randomly sampled from the Gaussian distribution with a standard deviation of 5
  - About 95% of "noisy" age false between 25 and 45

Bob's age (mean of distribution): $\mu = 35$

Standard deviation of noise: $\delta = 5$

2 standard deviations: $2\delta = 2 * 5 = 10$

Lower bound: $\mu - 2\delta = 35 - 10 = 25$

Upper bound: $\mu + 2\delta = 35 + 10 = 45$

Empirical rule of normal distribution:
68% of data within 1 standard deviation
95% of data within 2 standard deviations
99.7% of data within 3 standard deviations

DP-SGD

Abadi, et al., 2016

- **Gradient clipping**
- Opacus: PyTorch implementation
- TensorFlow Privacy
- DiffPrivLib: IBM (can be used with scikit-learn, etc)
- JAX-Privacy (Google DeepMind)
- PySyft: Federated Learning and distributed data science framework
- CrypTen (not recently updated)
- autodp: General DP library

Opacus: DP-SGD in PyTorch

- Miranov, et al. 2020
- **Modifies the SGD optimization to make it differentially private**
- Access to parameter gradients (gradients of the loss wrt to each parameter)
- Per sample gradients
- How much noise to add?
  - Looks at each batch, picks the largest norm of the gradient in a minibatch
  - Sample most at risk of exposure
  - Noise multiplier, and a bound on the gradient norm
  - Clips gradients that are larger than a given max norm
  - Clipping threshold
  - Microbatch method: Not optimal for performance
- Opacus uses a more efficient method than microbatch
- Uses Renyi DP (Miranov, 2017)
- Simple API, `make_private()`
- Adds randomness in the data loader

DP Training with Opacus

```
model = Net()
optimizer = SGD(model.parameters(), lr=0.05)
data_loader = torch.utils.data.DataLoader(dataset, batch_size=1024)
privacy_engine = PrivacyEngine()

model, optimizer, data_loader = privacy_engine.make_private(
    module=model,
    optimizer=optimizer,
    data_loader=data_loader,
    noise_multiplier=1.1,
    max_grad_norm=1.0,
)

model, optimizer, data_loader = privacy_engine.make_private_with_epsilon(
    module=model,
    optimizer=optimizer,
    data_loader=data_loader,
    target_epsilon=50.0
    target_delta=0.00001
)
```

Synthetic Training Data

Mode Size vs Training Examples

- *"A Few More Examples May be Worth a Billions of Parameters"*
- Kirstain, et al 2021
- Scaling parameters consistently improves performance
- Scaling the amount of training examples depends on task:
  - Significant benefit: Classification, extractive question answering, multiple choice
  - Less significant benefit: Open question answering

Differentially Private synthetic data

- Rosenblatt, et al 2020
- Efficiently capture unique aspects of the "real" dataset

Synthetic document generation

- SynthDog
- Part of Donut project
- Kim, et al, 2022

Privacy: For a Person

Split Silos

- Hospitals, wanting to create unified models
- Financial institutions, fraud detection
- Cross-national models

Privacy: Federated Learning

- No single training dataset
  - Distributed training set database
  - Distributed organizational training dataset:
    - Customers residing across jurisdictions, unified dataset impossible
  - Mobile device dataset
- Training participant membership attack (Identify that Alice was a participant)
- Vs individual training examples
- Trusted weight aggregation

Federated Learning

Federated training round:

1. Initialize model. Random or pretrained weights

2. Send pre-initialized based model, or part of the model, to clients

3. Clients further train the model on their own data

4. Clients send updated model weights to server. Server aggregates the clients' weights

Federated Learning

- Move the learning (computation) to the data
- vs the data to the learning
- Weight federating strategy:
  - Federated Average (FedAvg), (McMahan, et al., 2016)
  - Weighted average by number of training examples
- Clients may train for one epoch or even just few steps (mini-batch)
- Clients can send full model weights or just accumulated gradients
- Clients have intrinsic ID (esp on mobile, etc)
  - User privacy vs per sample privacy
  - DP as an optimization strategy to ensure the client examples' privacy
- FL + DP:
  - Model performance drops due to DP
  - Federated model performance can still exceed performance of single model

Federated Learning + Differential Privacy

2. Global model with DP

3. Model updates with DP

1. Local model with DP

Federated Learning + Differential Privacy

Federated Learning Frameworks

Project	Maintainer	Contributors	License
PySyft	OpenMined	424	Apache 2
FLOWER	Flower	121	Apache 2
TensorFlow Federated	Google	107	Apache 2
FATE	WeBank	86	Apache 2
OpenFL	Linux Foundation	78	Apache 2
Substra	Owkin	34	Apache 2
NVIDIA FLARE	NVIDIA	34	Apache 2

Document AI Privacy and Security Framework

	Private data	Problems	Techniques

1	**Regulatory compliance**	• Comply with privacy regulations:
		  • (a) Training on private data
		  • (b) Predicting with private data
			Localized training
			Localized predictions

2	**Encryption**	• Training on encrypted examples • Inference on encrypted documents	Data Loaders Homomorphic encryption (FHE)

3	**Privacy:**
	Sensitive, private information in training documents
		• **Verbatim regurgitation:**
		  • Model can leak training data
		• **Privacy for a single example:**
		  • Ensure that a specific training example does not reveal a person
		• **Privacy for a user:**
		  • Ensure that a user's aggregated data does not reveal a person
			Differential Privacy (DP)

4	**Distributed training corpus**	• It may not be feasible to assemble a single training example corpus	Federated Learning (FL)

Fine Tuning

Pre-trained Transformer Models

- **Model architecture**
- **Pretrained weights**
  - **BERT:**
    - BookCorpus
    - English Wikipedia
  - **LayoutLMv3:**
    - Text modality initialized from RoBERTa weights
    - IIT-CDIP, 11 million documents
  - **Donut:**
    - IIT-CDIP, 11 million
    - Synthetic docs generated via SynthDog (English, Chinese, Japanese, Korean), 0.5m each
  - **idefics2**:
    - Cauldron, dataset specifically curated for idefics
- Architecture and pretrained weight licenses may be different

Pre-trained Transformer Models

DistilBERT Backbone

```
model = AutoModel.from_pretrained("distilbert-base-uncased")

DistilBertModel(
  (embeddings): Embeddings(
    (word_embeddings): Embedding(30522, 768, padding_idx=0)
    (position_embeddings): Embedding(512, 768)
    (LayerNorm): LayerNorm((768,), eps=1e-12,
elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False)
  )
  (transformer): Transformer(
    (layer): ModuleList(
      (0-5): 6 x TransformerBlock(
          ….
    )
  )
)
```

DistilBERT for with Sequence Classification Head

```
model = AutoModelForSequenceClassification.from_pretrained('distilbert-base-uncased', num_labels=3)
```

```
(model): DistilBertForSequenceClassification(
    (distilbert): DistilBertModel(
      (embeddings): Embeddings(
        (word_embeddings): Embedding(30522, 768, padding_idx=0)
        (position_embeddings): Embedding(512, 768)
        (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
        (dropout): Dropout(p=0.1, inplace=False)
      )
      (transformer): Transformer(
        (layer): ModuleList(
          (0-5): 6 x TransformerBlock(
              …..
          )
        )
      )
    )
    (pre_classifier): Linear(in_features=768, out_features=768, bias=True)
    (classifier): Linear(in_features=768, out_features=3, bias=True)
    (dropout): Dropout(p=0.2, inplace=False)
```

Use Model with Pre-Trained Weights

- **Via API from Model-as-a-Service**
  - OpenAI (Chat-GPT, DALL-E)
  - Google (Genimi)
  - Specialist services:
    - Azure document analytics
    - AWS Reconition (OCR)
    - Google Cloud document AP
- **Download model with pre-trained weights**
  - HuggingFace hub
  - Be mindful of licensing
- Both can be fine-tuned

Model with Pre-Trained Weights from HuggingFace

Hugging Face | Search models, datasets, us | Models | Datasets | Spaces | Posts | Docs | Pricing

Transformers

Search documentation ⌘K

V4.41.3 | EN | 127,119

- Grounding DINO
- GroupViT
- IDEFICS
- Idefics2
- InstructBLIP
- KOSMOS-2
- LayoutLM
- LayoutLMV2
- LayoutLMV3
- LayoutXLM
- LiLT
- Llava
- LLaVA-NeXT
- LXMERT

LayoutLM

Overview

The LayoutLM model was proposed in the paper LayoutLM: Pre-training of Text and Layout for Document Image Understanding by Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. It's a simple but effective pretraining method of text and layout for document image understanding and information extraction tasks, such as form understanding and receipt understanding. It obtains state-of-the-art results on several downstream tasks:

- form understanding: the FUNSD dataset (a

LayoutLM
- Overview
- Usage tips
- Resources
- LayoutLMConfig
- LayoutLMTokenizer
- LayoutLMTokenizerFast
- LayoutLMModel
- LayoutLMForMaskedLM
- LayoutLMForSequence Classification
- LayoutLMForToken Classification
- LayoutLMForQuestion Answering

Zero-Shot vs Fine-Tuning

- **Zero-Shot**
  - May provide acceptable performance without any special tuning
- **Fine-Tuning**
  - Improve performance:
    - For your own data (e.g., your own classes, your own entities, custom-tailor)
    - For your own task
    - Inference speed and efficiency
- **Remote or local fine-tuning**
- **Overall fine-tuning goal**
  - Minimize effort (examples, setup, validation, etc)
  - Minimize cost (training infrastructure, etc)
  - Measure->Iterate, Measure->Iterate, Measure->Iterate, etc

Privacy-Aware Fine-Tuning

1. **Work with the fewest number of training examples possible**
   - While still producing an effective model
   - Few 100s of examples should suffice in many document AI scenarios
   - Curate example corpus for iterative fine-tuning
     - Incrementally tune for different tasks
     - Incrementally add examples, as needed
     - Carefully amend with synthetic training data, if necessary
2. **Modify the fewest amount of model parameters**
   - Resource efficiency, modest hardware resources
   - Iteration velocity
3. **Iterate**
   - Start with few examples and aim to modify no or few parameters
   - Evaluate the results
   - Repeat as needed

Iterative Parameter-Efficient Fine-Tuning

	Iteration	Tuning Concept
1	Start with a simple prompt and observe the response	**In-context learning** Model weights do not change
2	Write more detailed prompt ("mega prompt")	
3	Add examples to the prompt. Many-shot learning	
4	Fine-tune the model	**Parametric learning** Partial or full adaptation of model weights to domain
5	Use several LLMs: Break down a task into subtasks, agentic workflow	**Agents**

At each step, evaluate the model on the private document test or evaluation examples

Parameter-Efficient Fine-Tuning (PEFT)

- **No parameters modified**
  - **In-context learning:**
    - "Hard" prompt tuning:
      - Try different prompts, few-shot learning from handful of examples
    - "Soft" prompt tuning
      - Train prompt tensors prefix, prefix tuning
  - **Adaptors**: Insert adapter layers between transformer blocks
- **Feature-based tuning**
  - Keep pre-trained model frozen
  - Use pretrained model as feature extractor for downstream task
- **Fine-tune only the output layers**
  - No need to backpropagate to entire network
  - Keep backbone frozen
- **Update all layer weights (full fine-tuning)**
- **Reparameterize model weights**
  - Low rank adaptation (LoRA)

More examples
More resources

Iterative Tuning Strategy Trade-Offs

Tuning Concept	Examples	Privacy Risk	Cost	Effort
In-context learning Model weights do not change	None or few Few-shot: Dozens at most	**Via API endpoint:** ● Regulatory risk ● LLM vendor policies **Locally running LLM**: Low (sweetspot)	Per token API fees Volume-based Low	Low / medium
Parametric learning Partial or full adaptation of model weights to domain	100s to 1000s in each class	● Encrypted dataset? ● Distributed dataset? ● DP maybe required	Medium to high Training set QA Infrastructure cost	Medium / high
Agents	Hybrid	Hybrid	Hybrid	Medium / high

In-Context Learning

- **Model's parameters are not changed**
  - Only input prompt changed
  - Does not require dataset, or requires only a small dataset for in-context examples
  - May benefit from small set of examples for few-shot learning
  - Fast feedback loop
- **"Hard" prompt tuning**
  - Prompt not differentiable
  - Prompt engineering ("mega-prompts")
  - Prompt chaining
  - Function calling
- **Soft prompt tuning**
  - Prompt differentiable
  - Can use linear regression to find optimal prompt

In-Context Learning Enablers

- **Increasing context windows**
  - GPT-4o: 128,000 input tokens
  - Claude 3 Opus: 200,000 tokens
  - Gemini 1.5 Pro: 2 million token context window
- **Increasing ability to follow complex instructions**
- Prompts can be several pages long
- Prompts can be highly structured
- Prompts can include dozens (or more) examples
  - "Few-shot" learning
- Mega-prompt

"Hard" Prompt Tuning

- Specify some combination of prompt language that results in the desired output
- Just text
- Non-differentiable, trial-and-error
- Prompt the LLM to output structured data
- Normalize the LLM output with an appropriate tool

Prompting for Structured Output

```
"""You are an expert at information extraction from images of automobile loan
contracts.

  Given this page of an automobile loan contract, extract the following
information:
    - The name of the customer
    - The address of the customer
    - The Vehicle Identification Number (VIN)
    - The car make and model

    Do not guess. If some information is missing just return "N/A" in the
relevant field.
    If you determine that the image is not of an automobile loan contract,
just set all the fields in the formatting instructions to "N/A".

    You must obey the output format under all circumstances. Please follow the
formatting instructions exactly.
    Do not return any additional comments or explanation."""
```

LangChain

- **Prompt templates**
  - Reusable and dynamic prompts
  - Prompts can be adjusted to different tasks and different models
- **Chains**
  - Combine multiple calls to LLMs in support of workflows
  - Can use branching
- **Agents**
  - Supports complex decision making in the context of invoking LLMs
  - Make decisions based on inputs and actions
- **Context**
  - Store context of LLM interactions

Demo

Structured Data Extraction with Prompt Tuning GPT-4 Turbo via LangChain

Mega-Prompting API Costs

- API charges are per token
- Longer prompts = more tokens, potentially higher costs
- Images and text are tokenized in the input
  - Large and numerous images = lots of input tokens

Soft Prompt Tuning

- **Prepend a trainable parameter tensor to embedded tokens**
- Differentiable
- Tune the prepended tensor to improve model performance using gradient descent
- Soft tensor has same output dimensions as the embedded inputs

```
x = EmbeddingLayer(input_ids)
x = concatenate([soft_prompt_tensor, x], dim=seq_len)
output = model(x)
```

Prefix Tuning

- **Prepend a trainable parameter tensor to each transformer block**
- Implement a soft prompt tensor
  - Process soft prompt through fully connected layers
  - Concatenate with the input prompt
- Rest of the transformer block same as usual

```
transformer_block(x):
  soft_prompt_tensor = FullyConnectedLayer(soft_prompt)
  x = concatenate([soft_prompt_tensor, x], dim=seq_len)
  res = x
  x = SelfAttention(x)
  x = LayerNorm(x + res)
  res = x
  x = FullyConnectedLayer(x)
  x = LayerNorm(x + res)
  return x
```

Feature Extractor-Based Tuning

- **Use pre-trained model as a feature extractor**
  - Pretrained model already captures high-dimensional, informative features from input
  - Extracts complex patterns and relationships
- Keep model frozen
- Last hidden state on forward pass
- Train a downstream task on these features embeddings
- Precomputed for a given dataset when training for multiple epochs

Demo

Pre-Trained Model as a Feature-Extractor

Demo

Retrieval-Augmented Generation

Fine-Tune Outer Layers

- **Keep backbone frozen**
- **Add one more layer to the LLM**
- Update parameters only in new layers
- No need to backpropage through entire network

Fine-Tuning the Outer Layer

Demo

Fine-Tuning all the Layers

Hands-On Fine Tuning

LayoutLMv3

- Huang, et al, 2022
- BASE (133M params) and LARGE (368M params)
- Text:
  - Masked Language Modeling (MLM)
  - Pre-initialized with RoBERTa weights
- Linear image embedding
  - Vs CNN or Faster R-CNN
  - **Encodes image patches**
    - Objective: Reconstruct discrete image tokens of masked image patches vs raw pixels or image regions
  - Captures high-level layout structures
- **Word Patch Alignment (WPA)**
- Far fewer parameters needed

LayoutLMv3

"LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking"

Huang, Lv, Cui, Lu, Wei, 2022

LayoutLMv3$_{BASE}$

Model	Params	Modality	Image Embedding	FUNSD F1	CORD F1	RVL-CDIP Accuracy	DocVQA ANLS
BERT$_{BASE}$	110M	T		60.26	89.68	89.81	63.72
RoBERTa$_{BASE}$	125M	T		66.48	93.54	90.06	66.42
BROS$_{BASE}$	110M	T+L		83.05	95.73		
LiLT$_{BASE}$		T+L+I (region)		88.41	96.07	95.68	
LayoutLM$_{BASE}$	160M	T+L+I (region)	ResNet-101	79.27		94.42	
SelfDoc		T+L+I (region)	ResNeXt-101	83.36		92.81	
TILT$_{BASE}$	230M	T+L+I (region)	U-Net		95.11	95.25	
XYLayoutLM$_{MBASE}$		T+L+I (grid)	ResNeXt-101	83.35			
LayoutLMv2$_{BASE}$	200M	T+L+I (grid)	ResNeXt-101-FPN	82.76	94.95	95.25	78.08
LayoutLMv3$_{BASE}$	133M	T+L+I (patch)	Linear	90.29	96.56	95.44	78.76

Data Preparation for Fine-Tuning LayoutLMv3

Demo
ICMR24 Notebook 5

Fine-Tuning LayoutLMv3 with TorchLight

Demo
ICMR24 Notebook 6

Evaluating Fine-Tuned LayoutLMv3 with Custom Document Dataset