


SecurePrompt: a production-grade prompt and file scrubber with controlled de-scrubbing and audit intelligence for banking LLMs

Objective

Develop a secure, auditable, and reversible prompt / file scrubber tailored for use with LLMs in banking. The system must:

- Scrub sensitive data from prompts and files
- Log all transformations into an append-only audit trail
- Allow controlled de-scrubbing of specific entities
- Provide explainability, adaptive sensitivity (C2-C4), and metrics visibility

Data Classification

ING 		Data Classification
C4 Secret / C3 Sensitive	Customer sensitive data (Ethnic origin, Trade-union membership, Criminal convictions/ offences/related information, Political opinions, Health (incl. disabilities), Religious or philosophical beliefs, Sexual orientation/ Sex life, Genetic Data, Biometric Data (fingerprint/face biometrics, etc.)	
	Authentication Data (PIN, Biometrics, Passwords)	
	Credit card data necessary for transactions, Fraud investigations, Restrictions, Flags, Risk ratings, Warning lists, Sanctions lists, Interdictions	
	Employee screening results	
	Whistleblower details	
C3 Confidential	Customer/Employee data (First Name, Last Name, CNP, CIF, ID Series and Number, Country, Postal Address, Postal Address Identification, Street Name, Phone number(s), Date of birth, Age, Education, Financial data, Transaction data, Family data, Marital Status, Gender, Citizenship, E-mail address, Employment data, Household data, Video recordings, Voice recordings, IP Address, Location Data, Hobbies, Interest, Carrier Track, Expenses	
	Link between ING and customers	
	Link between customers and ING products, Products and Services Used/Held; Credit History, Customer segment, Bank guarantee, Company mandate, Title	
	Reports over customers and employees	
	Agreements and contracts with customers / suppliers	
C2 Restricted	Payment orders or report with payment details, Transaction data, Expenses, Credit history, BC/ANAF/CRC reports	
	ING software source code and configuration, Security/System log files, Infrastructure assets configuration, change management, problem management, incident management, access lists, firewall rules, proxy settings, etc	
	NFRD Reports, Sox Reports, Risk events, KYC/CDD reports	
	Employee Data used for contacting purposes First Name, Last Name, Phone number(s), e-mail address	
	Product and Service specification (Product development, System development, Calculations, Legal and Tax options, Reports over Products, ING internal references, status, dates)	
C1* Public	Payment (transfer order initiated by ING)	
	Physical/IaaS (Hypervisor/ESX; Storage: physical & management; Network: physical & management; Configurations for Virtual Systems; VM's; Storage; Network interfaces; System names; Designs documents)	
	Definitions, policies, guidelines, Processes, Configuration data (CMDB)	
	Product and service offering (not linked to customers) - Published description	
	Public information - Press release, Annual report, Public website (without personal data)	
	*only if it was already published by ING (site, press release, etc.) and it is used for the purpose it was published	

Features

1. Prompt scrubbing

- Detect and redact sensitive information (C2-C4)
 - o First phase: use static anonymization
 - regular expressions, references, etc.
 - o Second phase: build anonymization intelligence
 - Model that predicts data as being confidential, assigns a confidentiality class (C2 – C4) and confidence score.
- Provide explanation for each redaction (“detected as IBAN”)
- Include confidence scores and allow override with justification
- Replace scrubbed data with an identifier
 - o Ensure the LLM does not lose information / context due to the identifier replacing the actual information.
- **Bonus:** Include pre-processing that avoids prompts bypassing the scrubbing procedure

2. File scrubbing

- Accept and process files in multiple formats: pdf, docx, txt, html, csv, etc.
 - o Screenshots (.png) will be the most prominent format – opportunity to leverage OCR
- Extract and scrub text using same principle as prompt scrubbing
- Return a redacted version of the file with confidential data being replaced or removed

3. Audit logging

- Log
 - o Timestamp
 - o corporate key
 - o logon & logoff actions
 - o opening & closing session
 - o identification of client
 - device identification
 - browser identification
 - location identification (MAC)
 - o original & scrubbed content
 - highlight when a prompt attempts to:
 - Search on customer name
 - Inject data for a customer
 - Update the customer’s information
 - Delete customer information
 - Share / export customer information
 - o scrubbing actions & metadata

- closure without continuation or result
 - confidence level
- Use append-only, tamper-proof logs (integrity must be ensured)

4. De-Scrubbing module

- Allow authorized users to request reintroduction of scrubbed data
- Support
 - full de-scrubbing
 - selective entity restoration
- Require
 - Justification for de-scrubbing
 - Logging of de-scrubbing actions

5. Adaptive scrubbing sensitivity

- Dynamically adjust scrubbing strictness based on prompt context (or parameters)
- Example parameter: Risk level of the LLM task (C2 – C4)
 - if the target LLM risk level is C2, then all C2 data should be scrubbed.
 - Per label: name, IBAN, address, ...

6. Optional: Scrubbing metrics dashboard

- Track & visualize
 - Number of prompts scrubbed
 - Types of entities detected
 - De-scrubbing requests
 - Override frequency and reasons
- Provide insights into system usage and risk trends
 - Predict future request volume

Performance measure

As a financial institution, ING is risk averse: leakage of sensitive data to the external world might result in financial or reputational damage. As such, anonymization can only be considered successful if no sensitive information was overlooked.

Deliverables

- We expect the project to be implemented in Python 3 and all required modules of your choice. These modules are expected to be free and open source. The project will be delivered as one or multiple standalone modules, depending on your implementation choices.
- Each significant method should be covered by unit tests to guarantee the success rate of scrubbing and showcase the precision of the devised module(s).
- The delivered modules should strive for maximal efficiency – speed of execution is a critical factor of success.
- Extensive documentation is expected, with but not limited to:
 - Regular single line comments to clarify operations
 - A detailed docstring for each function or method, providing the necessary information on the input, operations and return.