



# SECURE PROMPT

## LLM DATA PROTECTION

EMMANUEL GOLDBERG  
HAJER SMIAI  
MARC VAN GOOLEN  
SOFIA MURILLO

# SCENARIOS & FOCUS OF TOOL



SECURE PROMPT  
LLM DATA PROTECTION

1

**Risk 1:**  
**Public info into  
LLM**

## Example:

- Draft a LinkedIn post promoting the 2024 Annual Report with this link: <link to Annual-report-2024.pdf> Keep tone professional and include a call to read the full report.

2

**Risk 2:**  
**Sensitive info into  
LLM & connect to  
ING systems**

## Example:

- Please initiate retirement for App\_31. It's Installed / Operational in Production, IT Custodian Joseph Johnson, Config group T34407, Cost Center 68805132. Reference IT Governance Policy

3

**Risk 3:**  
**Sensitive info into  
LLM**

## Example:

- Give me the avg monthly account balance for 2025 for Linda Williams, Account Number and National ID 90.06.25-173.69
- Review this data (text, screenshot, file) and give me the top 10 customers with largest overdraft in September 2025

**Audit Log**

# ARCHITECTURE



SECURE PROMPT  
LLM DATA PROTECTION

## DATA FLOW



### ACTIONS

- |  |   |  |  |   |  |  |
|--|---|--|--|---|--|--|
| <ul style="list-style-type: none"><li>• Get prompt</li><li>• Auto-detect sensitivity level</li></ul> | <ul style="list-style-type: none"><li>• Detection</li><li>• Encryption</li><li>• Sensitivity classification</li><li>• Scrubbing &amp; De-scrubbing</li><li>• Performance monitoring</li></ul> | <ul style="list-style-type: none"><li>• Analyze / generate</li></ul> | <ul style="list-style-type: none"><li>• Sanitized response</li></ul> | <ul style="list-style-type: none"><li>• Check permissions (admin login)</li></ul> | <ul style="list-style-type: none"><li>• De-scrub</li></ul> | <ul style="list-style-type: none"><li>• Response</li></ul> |
|--|---|--|--|---|--|--|

### TOOLS

- |  |  |   |  |  |  |  |
|--|--|---|--|--|--|--|
| <ul style="list-style-type: none"><li>• User Interface</li></ul> | <ul style="list-style-type: none"><li>• Fast API</li><li>• Presidio Analyzer</li><li>• SpaCy NER</li><li>• Custom regex</li><li>• YAML rules</li></ul> | <ul style="list-style-type: none"><li>• Gemini / Co-pilot</li></ul> | <ul style="list-style-type: none"><li>• User Interface</li></ul> | <ul style="list-style-type: none"><li>• Based on user ID</li></ul> |  | <ul style="list-style-type: none"><li>• User Interface</li></ul> |
|--|--|---|--|--|--|--|

# API END POINTS - SCRUB

127.0.0.1:8000/docs#/

SecurePrompt

0.1.0

OAS 3.1

/openapi.json

default

GET

/

Root

POST

/scrub/prompt

Scrub Prompt

POST

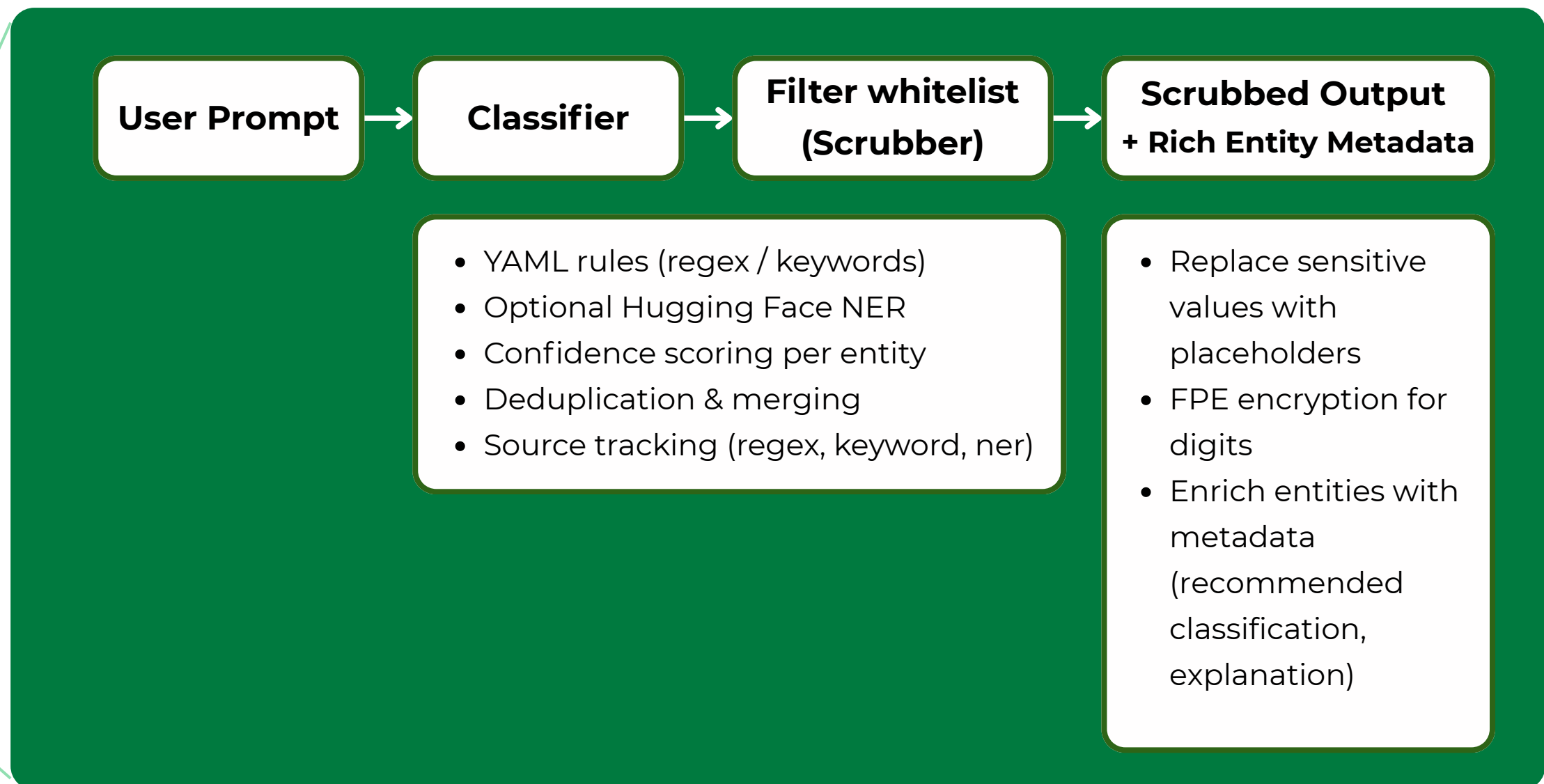
/descrub/prompt

Descrub Prompt

POST

/ask-llm

Ask Llm



# API END POINTS - DESCRUB



127.0.0.1:8000/docs#/

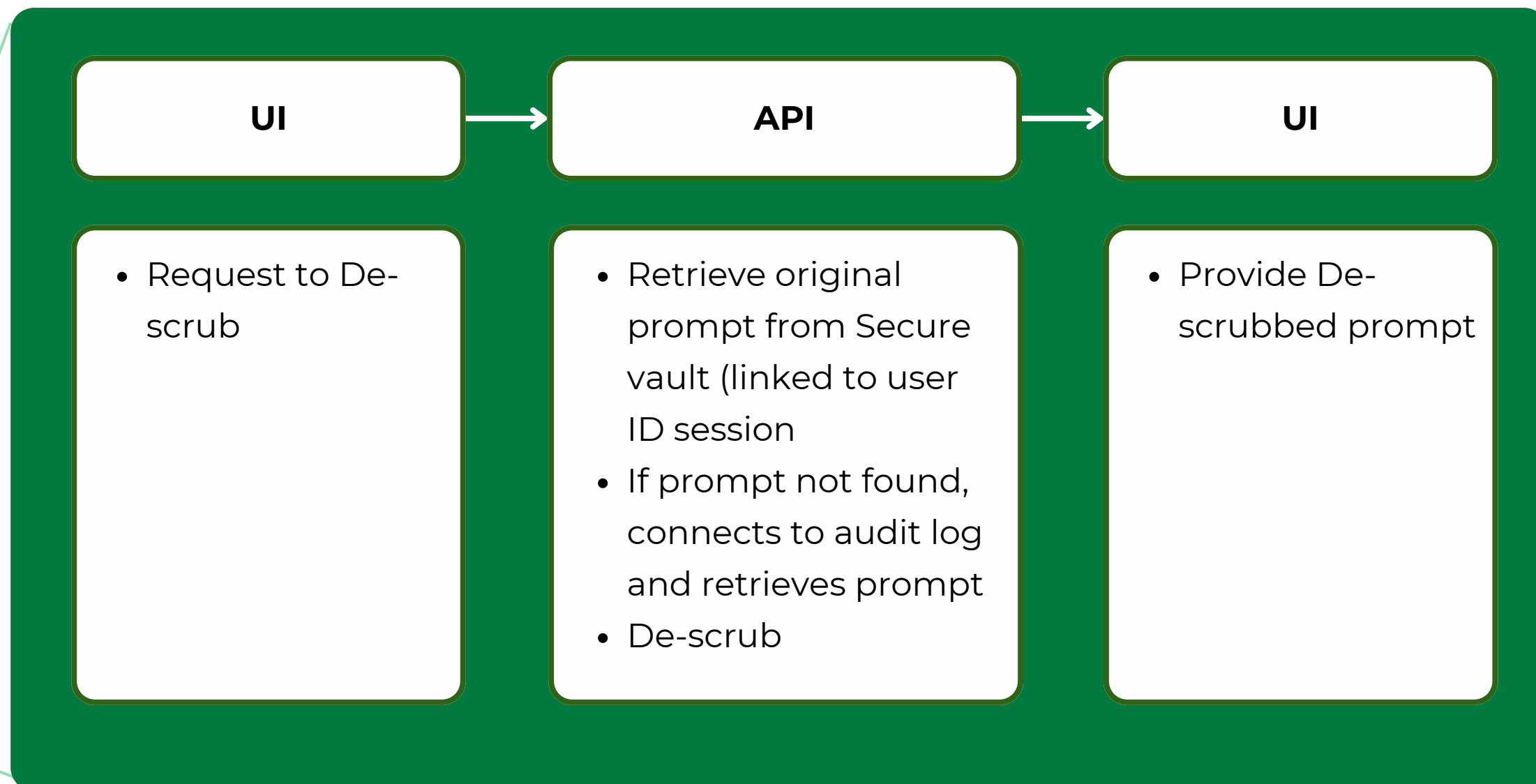
## SecurePrompt

0.1.0 OAS 3.1

/openapi.json

### default

- GET** / Root
- POST** /scrub/prompt Scrub Prompt
- POST** /describ/prompt Describ Prompt
- POST** /ask-llm Ask Llm



# API END POINTS - DESCRUB



SECURE PROMPT  
LLM DATA PROTECTION

## Curl

```
curl -X 'POST' \
  'http://127.0.0.1:8000/describ/prompt' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "scrubbed_prompt": "{{NAME_2}} {{COST_CENTER_5}} BE{{PHONE}} for review after {{NAME}} transfer{{PHONE_2}}b-4b8f-8b92-ab{{AGE_2}}cdab7faaf ({{COST_CENTER_8}} {{ACCOU
    "placeholders": [
      "{{AGE_2}}", "{{PHONE}}", "{{PHONE_2}}", "{{DATE}}", "{{COST_CENTER}}", "{{ACCOUNT_NUMBER}}", "{{ACCOUNT_BALANCE}}", "{{NAME}}", "{{FILENAME}}", "{{YEAR}}", "{{NAM
    ],
    "user_id": "Laura",
    "user_role": "Auditor",
    "justification": "Audit purposes"
  }'
```

## Request URL

http://127.0.0.1:8000/describ/prompt

## Server response

Code	Details
------	---------

200

## Response body

```
{
  "describbed_prompt": "Flag IBAN BE62 5474 2787 2767 for review after Failed transfer 93151103-357b-4b8f-8b92-ab7cdab7faaf (EUR 7,813.25) on 2025-08-16.",
  "restored_entities": [
    {
      "placeholder_id": "{{AGE_2}}",
```

# API END POINTS - LLM

→ ↺ ⓘ 127.0.0.1:8000/docs#/

SecurePrompt

0.1.0 OAS 3.1

/openapi.json

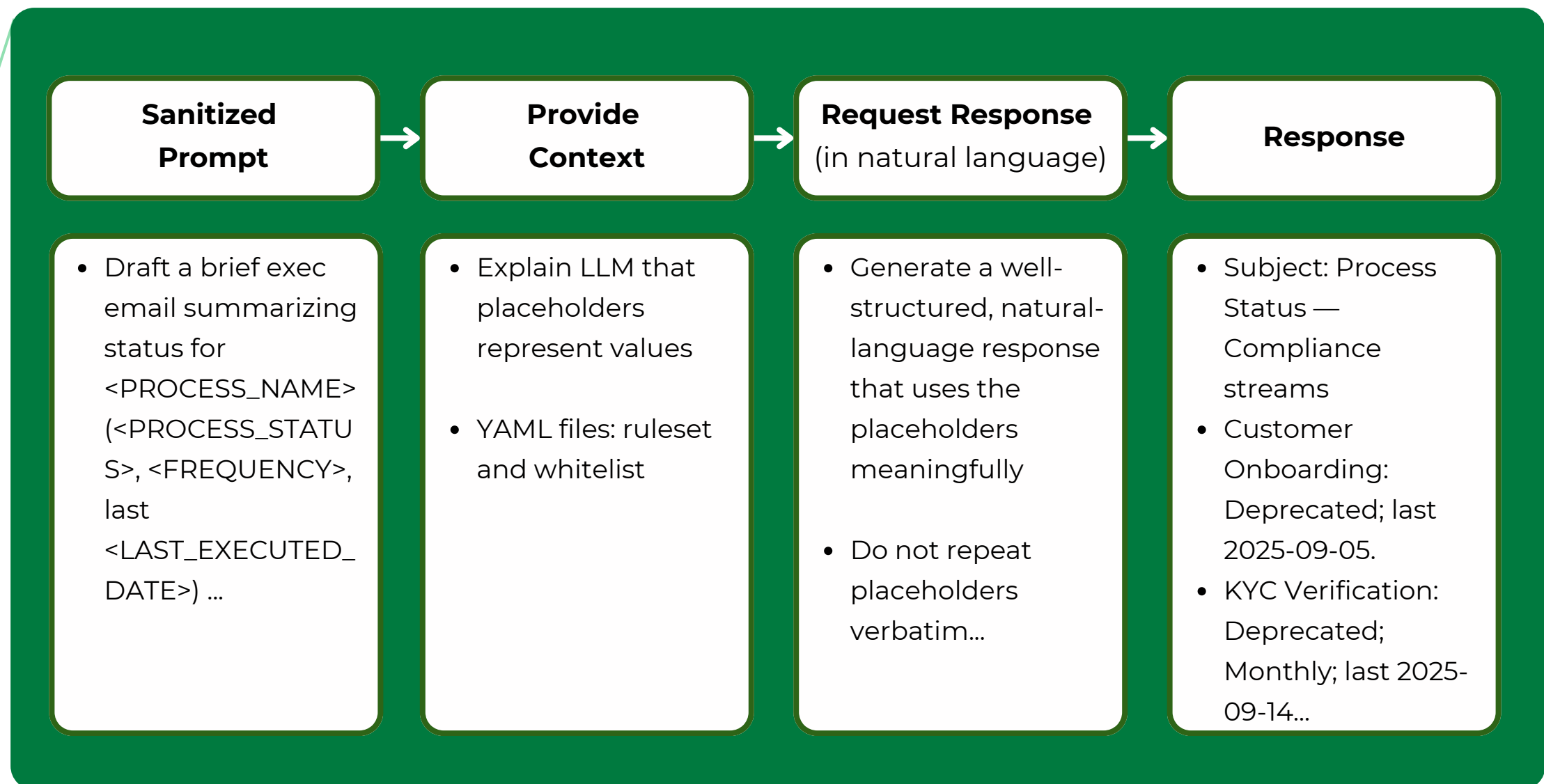
default

GET / Root

POST /scrub/prompt Scrub Prompt

POST /descrub/prompt Descrub Prompt

POST /ask-llm Ask Llm





# API END POINTS - LLM



SECURE PROMPT  
LLM DATA PROTECTION

→ 127.0.0.1:8000/docs#/default/ask\_llm\_ask\_llm\_post



```
curl -X 'POST' \
  'http://127.0.0.1:8000/ask-llm' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "scrubbed_prompt": "Prepare a 5-point agenda: align <POLICY_NAME> (<POLICY_STATUS>, last reviewed <LAST_REVIEWED_DATE>, approval <APPROVAL_LEVEL>, responsible <RESPONSIBLE_DEPT>.",
    "user_id": "Jane"
  }'
```

Request URL

http://127.0.0.1:8000/ask-llm

Server response

Code

Details

200

Response body

```
{
  "llm_response": "{NAME} has put forward a comprehensive agenda, comprising {ACCOUNT_BALANCE} key points. The primary objective is to ensure robust alignment between the '<POLICY_NAME>' policy and the '<PROCESS_NAME>' process.\n\nThe '<POLICY_NAME>' policy is currently in a <POLICY_STATUS> status. It was last reviewed on <LAST_REVIEWED_DATE> and holds an <APPROVAL_LEVEL> approval level, with the <RESPONSIBLE_DEPT> department responsible for its management.\n\nConcurrently, the '<PROCESS_NAME>' process maintains a <PROCESS_STATUS> operational status and features an <AUTOMATION_STATUS> level of automation. This process is associated with the {COST_CENTER} cost center, and the <RESP_TEAM> is designated as its owner. Its last recorded execution, a procedure that often incorporates a masked PIN for secure access or verification, took place on <LAST_EXECUTED_DATE>.",
  "audit_id": "110d8be1e1ee6a94668f90b001cd7c640376b2ec318a3918517cf178568badfe"
}
```



Download



# DEMO



## Prompt Scrubbing

Secure your prompts before sending them to Large Language Models.

Compliance: GDPR & PCI-DSS

### Original Prompt

Type your prompt or upload a file (PDF, image, PPTX, TXT...)

Prompt (original)

Enter your prompt here...

Example:

Transfer €5,000 to IBAN BE68539007547034

Contact: john.doe@ing.be

Phone: +32 471 23 45 67

Card: 4532 1234 5678 9010


All sensitive data will be automatically detected and scrubbed.

### Secured Prompt



Secured content will appear here  
Click "Secure & Scrub" to begin

### LLM Response

 Ask LLM

# CONFIDENCE LEVEL - APPROACH



## DEFINITION

- A number **between 0.0 and 1.0**, shown as 0-100% in the User Interface
- It expresses **how confident the scrubbing pipelines is in entities it detected** (e.g. IBAN, credit card, email, etc.)
- It is not a model safety or sensitivity level, it is **a measure tied to detection quality**

## CALCULATION

- The scrubber runs **multiple detectors** and normalizes the outputs into a unified entity schema.
- Detectors include:
  - **Rule / regex detectors** for banking PII: IBAN, SWIFT/BIC, Credit card (with Luhn check), Amounts, Email, etc.
  - **Context guards** to reduce false positives
  - **SpaCy / Presidio NER / Recognizers** for generic PERSON/ORG/GPE
  - **Whitelist filtering** to ignore approved terms (e.g. "ING")
- Post-processing & merging to deduplicate and resolve overlaps
- Each detector assigns a **per-entity confidence** and then calculates an average for the full prompt

# CONFIDENCE LEVEL - DETAIL



Entity type	Typical source	Confidence logic (typical)
-----	-----	-----
<b>**IBAN**</b>	Regex + <b>**MOD97**</b> checksum	Base high ( $\approx 0.90-0.95$ ) + <b>**+0.03-0.05**</b> if checksum passes
<b>**Credit card**</b>	Regex + <b>**Luhn**</b> checksum	Base ( $\approx 0.88-0.92$ ) + <b>**+0.05-0.10**</b> if Luhn passes; minus if spacing/length is suspicious
<b>**Email**</b>	RFC-esque regex	Base ( $\approx 0.93-0.97$ ); may reduce slightly for unusual TLDs
<b>**SWIFT/BIC**</b>	Regex (8/11 chars)	Base ( $\approx 0.90-0.95$ ); slight boost for valid length/pattern
<b>**Money**</b>	Regex (currency + amount)	Base ( $\approx 0.85-0.92$ ); higher with explicit symbol/ISO code
<b>**Phone**</b>	Regex + <b>**guards**</b>	<b>**0.80**</b> in the provided code when not excluded by guards
<b>**Generic NER**</b>	spaCy/Presidio	Usually mapped to a conservative fixed range ( $\approx 0.60-0.75$ ) unless the library provides a score

## Guards & Validators:

- **Luhn**, detects mistyped or invalid credit card numbers
- **MOD97** checks IBAN control digits
- Using **real industry standards** and contextual filters

## Overlap Resolution:

- Resolves conflicts using a **priority map** (IBAN > Credit Card > SWIFT > Email > Phone > person > org > location)
- Entity with higher priority or higher confidence prevails

## Additional Rules:

- If a match fails a validator it is discarded or assigned lower confidence
- If an entity looks like another entity (e.g. 16 digit phone) it is skipped or downgraded
- Whitelist terms are removed



# AUDIT LOG

```
{
  "event": {
    "action": "scrub",
    "timestamp": 1759306823.4307806,
    "corporate_key": "ING_CORP_KEY_001",
    "user_id": "Jean",
    "device": "unknown_device",
    "browser": "unknown_browser",
    "location": "unknown_location",
    "original": "Place a quarter-end freeze on App_53 (Installed / Operational / Production; Config T14899; Cost B8089966) and keep Transaction Monitoring (Deprecated) jobs disabled during the freeze.",
    "scrubbed": "[REDACTED]",
    "entities": [
      {
        "id": "f8b4db-1",
        "entity": "Name",
        "value": "Operational",
        "classification": null,
        "confidence": 1.0,
        "explanation": ""
      },
      {
        "id": "d8b5b0-2",
        "entity": "Name",
        "value": "Deprecated",
        "classification": null,
        "confidence": 1.0,
        "explanation": ""
      },
      {
        "id": "035685-3",
        "entity": "Name",
        "value": "Place",
        "classification": null,
        "confidence": 0.9999999999999998,
        "explanation": ""
      },
      {
        "id": "d63b9f-4",
        "entity": "Name",
        "value": "Installed",
        "classification": null,
        "confidence": 0.9999999999999998,
        "explanation": ""
      },
      {
        "id": "883f44-5",
        "entity": "Name",
        "value": "Production",
        "classification": null,
        "confidence": 0.9999999999999998,
        "explanation": ""
      },
      {
        "id": "c2cea1-6",
        "entity": "Name",
        "value": "Config",
        "classification": null,
        "confidence": 0.9999999999999998,
        "explanation": ""
      },
      {
        "id": "a3bd48-7",
        "entity": "Name",
        "value": "Cost",
        "classification": null,
        "confidence": 0.9999999999999998,
        "explanation": ""
      },
      {
        "id": "d15bfb-1",
        "entity": "Cost Center",
        "value": "B8089966",
        "classification": null,
        "confidence": 0.9999999999999998,
        "explanation": ""
      },
      {
        "id": "98709c-2",
        "entity": "Cost Center",
        "value": "T14899",
        "classification": null,
        "confidence": 0.9999999999999998,
        "explanation": ""
      },
      {
        "id": "0dd30e-8",
        "entity": "Name",
        "value": "Transaction Monitoring",
        "classification": null,
        "confidence": 0.9999999999999998,
        "explanation": ""
      },
      {
        "id": "c14fc7-1",
        "entity": "VM Name",
        "value": "freeze",
        "classification": null,
        "confidence": 0.99999984375,
        "explanation": ""
      },
      {
        "id": "83e70c-1",
        "entity": "First Name",
        "value": "Transaction",
        "classification": null,
        "confidence": 0.999999375,
        "explanation": ""
      },
      {
        "id": "665c6e-2",
        "entity": "First Name",
        "value": "Monitoring",
        "classification": null,
        "confidence": 0.9999996875,
        "explanation": ""
      },
      {
        "id": "0f5de4-1",
        "entity": "Policy Name",
        "value": "Place a quarter-end freeze on App_53 (Installed / Operational / Production; Config T14899; Cost B8089966) and keep Trans",
        "classification": null,
        "confidence": 0.9999996875,
        "explanation": ""
      },
      {
        "id": "d82765-2",
        "entity": "Policy Name",
        "value": "action Monitoring (Deprecated) jobs disabled during the freeze.",
        "classification": null,
        "confidence": 0.9999996875,
        "explanation": ""
      },
      {
        "id": "9cd0e4-1",
        "entity": "Application Name",
        "value": "Installed",
        "classification": null,
        "confidence": 0.99999375,
        "explanation": ""
      },
      {
        "id": "118b5f-2",
        "entity": "Application Name",
        "value": "Operational",
        "classification": null,
        "confidence": 0.99999375,
        "explanation": ""
      },
      {
        "id": "8f9102-3",
        "entity": "Application Name",
        "value": "Production",
        "classification": null,
        "confidence": 0.99999375,
        "explanation": ""
      },
      {
        "id": "3a721d-4",
        "entity": "Application Name",
        "value": "Config T14899",
        "classification": null,
        "confidence": 0.99999375,
        "explanation": ""
      },
      {
        "id": "97e664-5",
        "entity": "Application Name",
        "value": "Cost B8089966",
        "classification": null,
        "confidence": 0.99999375,
        "explanation": ""
      },
      {
        "id": "3099da-6",
        "entity": "Application Name",
        "value": "and keep Transaction Monitoring",
        "classification": null,
        "confidence": 0.99999375,
        "explanation": ""
      },
      {
        "id": "9b39bb-1",
        "entity": "Phone Number(s)",
        "value": "8089966",
        "classification": null,
        "confidence": 0.999875,
        "explanation": ""
      },
      {
        "id": "b87691-7",
        "entity": "Application Name",
        "value": "jobs disabled during the freeze",
        "classification": null,
        "confidence": 0.999875,
        "explanation": ""
      },
      {
        "id": "a0333a-2",
        "entity": "VM Name",
        "value": "quarter-end",
        "classification": null,
        "confidence": 0.999875,
        "explanation": ""
      },
      {
        "id": "924409-3",
        "entity": "VM Name",
        "value": "App_53",
        "classification": null,
        "confidence": 0.999875,
        "explanation": ""
      }
    ]
  }
}
```

## Audit Log

- Audit ID
- Event (scrub, ask LLM, response from LLM, de-scrub)
- Time stamp
- User ID
- Original prompt
- Entities & details (definition, classification level, value)
- Scrubbed prompt
- Justification for de-scrub
- LLM response
- Further improvements: Session ID, Dashboard with metrics and warnings

# FURTHER DEVELOPMENTS



SECURE PROMPT  
LLM DATA PROTECTION

## Prompt/ File Scrub

- Improved multi-pass detection with **Custom ML**
- **Build fusion hierarchy** picks canonical label + confidence
- Explainability: method + validator + confidence
- ML ADD-ON (Recall Booster)

## Scrubbing Sensitivity

- **Consistent Policy** (C2-C4) **implementation** UI + FastAPI
- YAML policies: thresholds, labels, strictness
- Runtime policy banner + policy-hash in audit

## De-scrubbing

- Token vault (KMS), RBAC & approvals (dual for C4?)
- Selective/full restore; immutable audit of who/when/why

## Audit Logging & Search

- Append-only receipts incl. user/session/device
- Search by label/text; confidence & rationale logged

## Feature Dashboard (Bonus)

- Volumes, entity mix, FP/FN, de-scrub, latency
- Trends & forecasts; per-policy observability



Q & A