Web3-based Privacy-Enhanced AI Chatbot Framework with Clio-X Integration

A comprehensive study on designing and implementing a privacy-preserving AI chatbot framework integrated with Clio-X in a Web3 environment.

2025.07.19

TEAM: PowerBlock

Background & Motivation

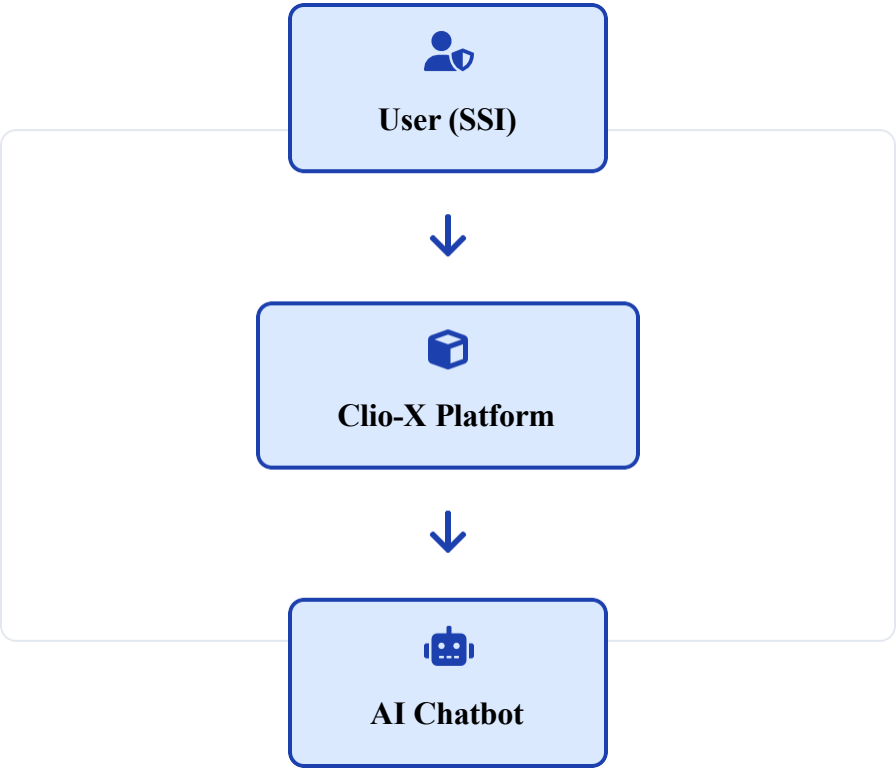
- The acceleration of digital transformation post COVID-19 has highlighted the need for privacy-preserving AI solutions.
- Web3 environments emphasize decentralized control and self-sovereign identity (SSI).
- Increasing demand for AI chatbots that guarantee data privacy in Web3-native architectures.

Privacy-preserving solutions are becoming critical in the expanding Web3 ecosystem

Clio-X Platform & Self-Sovereign Identity (SSI)

- **Clio-X:** A core Web3 platform enabling decentralized data management, user authentication, and ownership validation on the blockchain.
- **Users gain full control over data** through Self-Sovereign Identity (SSI).
- **Critical role** in supporting privacy and user-driven data in Web3 services.

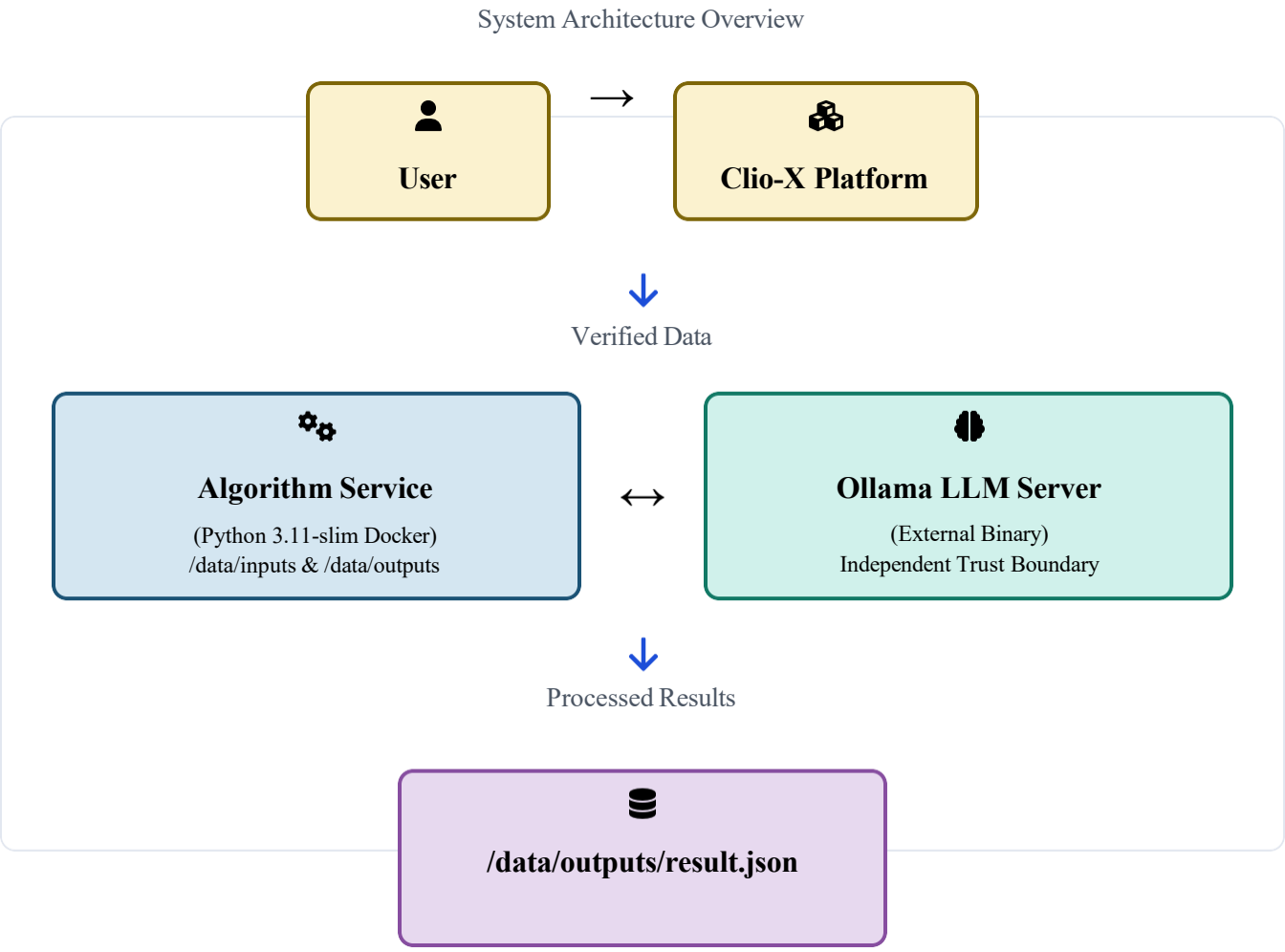
> *"Clio-X operates as the trust layer between blockchain data verification and AI processing, ensuring only user-approved data enters the AI pipeline."*



System Architecture Overview

- **Algorithm Service:** Python 3.11-slim Docker container with secure data handling via isolated volumes (/data/inputs and /data/outputs).
- **Ollama LLM Server:** External independent service operating as a binary on the host machine, ensuring clear trust boundary separation.
- **Data Flow Control:** Only verified data from Clio-X is processed, maintaining Web3-native security principles.

> *"The architecture creates clear trust boundaries between the AI processing components and Web3 authentication, ensuring privacy by design."*



Privacy Protection Mechanisms

- **Hybrid masking approach:** Combination of spaCy Named Entity Recognition (NER) and regex pattern matching for comprehensive sensitive entity detection
- **Entity types masked:** PERSON, EMAIL, and GPE (Geo-Political Entity) data are automatically replaced with **[MASKED]** before any LLM processing
- **Privacy-by-design principle:** All sensitive information masking happens automatically before any data leaves the Algorithm service

> *"All data masking operations are performed on the Clio-X verified data as a mandatory step before any external LLM calls, ensuring no sensitive information ever leaves the trusted environment."*

Masking Process Flow

Input Text:

"Summarize Cameroon decrees mentioning Mr. John Doe from Paris and email john.doe@example.com"

↓

spaCy NER Processing:

```
doc = nlp("Input text") for ent in
doc.ents:
        if ent.label_ in ["PERSON", "GPE"]: # Replace with
                [MASKED]
```

↓

Regex Enhancement for EMAIL:

```
email_pattern = r'[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}' text = re.sub
(email_pattern, '[MASKED]', text)
```

↓

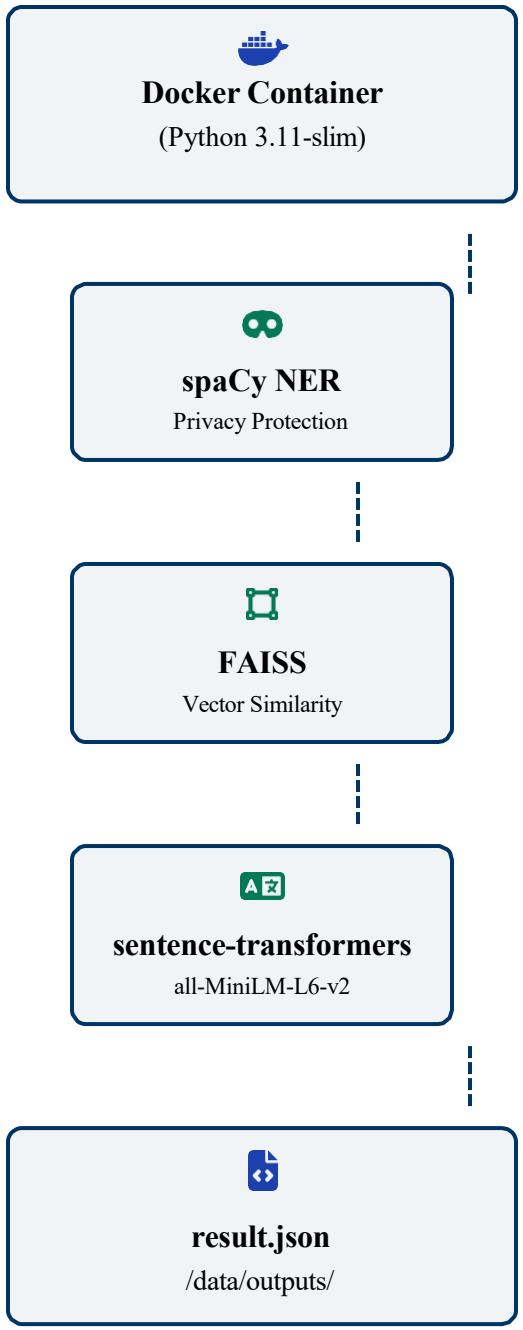
Processed Output:

"Summarize Cameroon decrees mentioning **[MASKED]** from **[MASKED]** and email **[MASKED]**"

Technical Implementation Details

- **Algorithm service** runs in Python 3.11-slim Docker container; utilizes **/data/inputs** and **/data/outputs** volumes linked to Clio-X.
- **FAISS-based similarity search** with sentence-transformers all-MiniLM-L6-v2 for context retrieval and question processing.
- Results are exported to **/data/outputs/result.json** for downstream use by Clio-X platform.

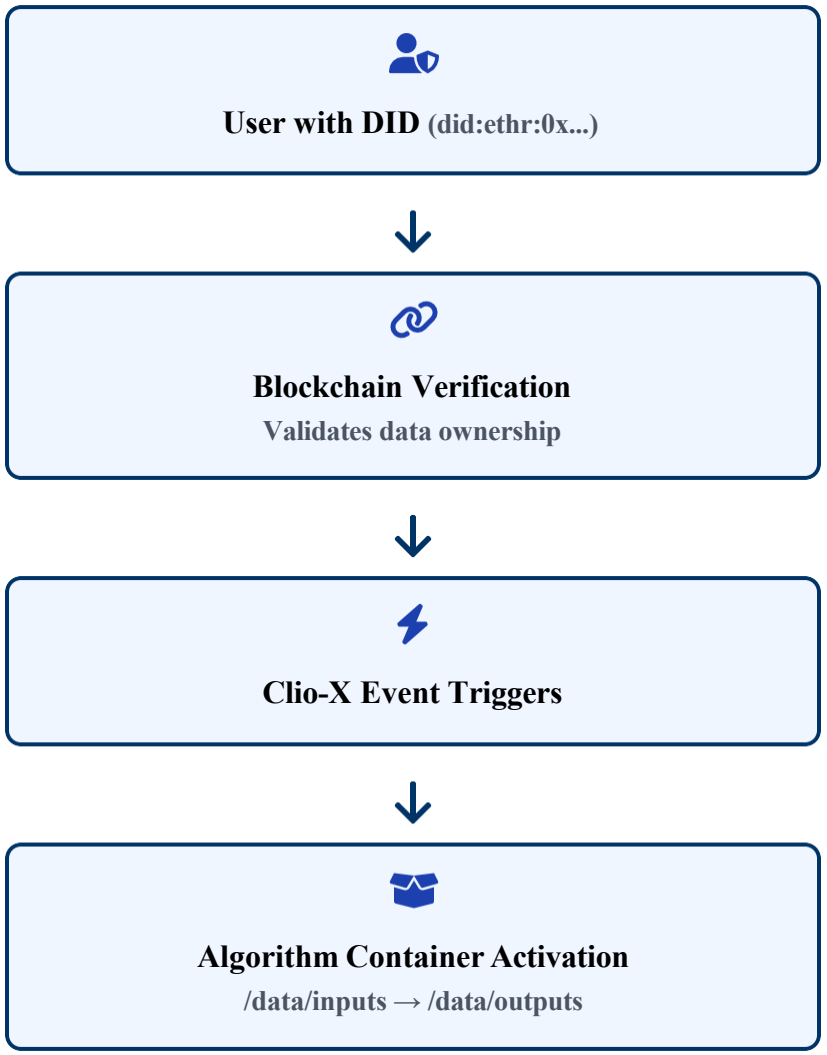
> *"The containerized architecture ensures isolation and security, while the use of volumes enables seamless data exchange with the Clio-X platform."*



Web3-Native Integration & Data Provenance

- **DID-based authentication** and on-chain data verification integrated via Clio-X APIs.
- **Docker Compose** enables on-demand algorithm service execution triggered by Clio-X events.
- **Data provenance** and complete auditability guaranteed throughout the processing pipeline.

> *"Web3-native integration ensures that every AI interaction is verifiable, traceable, and user-controlled, maintaining the core principles of decentralized applications."*



Experimental Results & Testing Scenarios

- **Multiple privacy scenarios tested** using inputs containing sensitive information across different entity categories.
- **PERSON and GPE masking** were successfully processed by spaCy NER component with high accuracy.
- **EMAIL masking** required regex enhancement to complement spaCy NER for complete coverage.

> Results are stored in /data/outputs/result.jsonfor secure further processing by Clio-X, maintaining the Web3-native privacy preservation chain.

Our Algorithm

```
Successfully built 69034c74f407
Successfully tagged template_algorithm:latest
Creating template_algorithm_1 ... done
Attaching to template_algorithm_1
algorithm_1   | Starting Algorithm container...
algorithm_1   | Running in DEV mode
algorithm_1   | INFO:sentence_transformers.SentenceTransformer:Use pytorch device_name: cpu
algorithm_1   | INFO:sentence_transformers.SentenceTransformer:Load pretrained SentenceTransformer: all-MiniLM-L6-v2
algorithm_1   | INFO:implementation.algorithm:Reading input from: /data/inputs/sample_input/0
Batches: 100%|[redacted]| 1/1 [00:00<00:00, 30.17it/s]
algorithm_1   | INFO:implementation.algorithm:Using Ollama URL: http://192.168.50.102:11434/api/generate
algorithm_1   | INFO:implementation.algorithm:Saved result to /data/outputs/result.json
template_algorithm_1 exited with code 0
(venv) root@soonduck2:~/blockaton/Cliox-Blockathon25/template# cd _data/outputs/
(venv) root@soonduck2:~/blockaton/Cliox-Blockathon25/template/_data/outputs# cat result.json

  "question": "Summarize Cameroon decrees mentioning Mr. John Doe from Paris and email john.doe@example.com",
  "sanitized_question": "Summarize Cameroon decrees mentioning Mr. [MASKED] from [MASKED] and email [MASKED]",
  "dataset": "cameroon",
  "context": "Appointment of Attach\u00e9 at the Civil Cabinet of the President of the Republic Mr. Aboubakary Abdoulaye, Administrative Officer, is, with effect from the date of signat
ure of this decree, appointed Attach\u00e9 at the Civil Cabinet of the President of the Republic. This decree shall be registered and published in the Official Gazette in English and Fr
ench.\nAppointment of Charg\u00e9 de Mission at the Civil Cabinet of the President of the Republic Mrs. Onamb\u00e91\u00e9, n\u00e9e Rose Amvuma, High Schools Teacher, is, with effect f
rom the date of signature of this decree, appointed Charg\u00e9 de Mission at the Civil Cabinet of the President of the Republic. This decree shall be registered and published in the Of
ficial Gazette in English and French.\n",
  "response": " The decrees appoint Mr. Aboubakary Abdoulaye as Attach\u00e9 at the Civil Cabinet of the President of the Republic, effective from the date of signature. No mention was
made about the city or email. For Mrs. Onamb\u00e91\u00e9, no specific name or email was mentioned in the provided decrees."
```

Privacy Masking Example

Input Query:

"Summarize Cameroon decrees mentioning Mr. John Doe from Paris and email john.doe@example.com"

↓

"Summarize Cameroon decrees mentioning Mr. [MASKED] from [MASKED] and email [MASKED]"

After Masking:

Entity Detection Results

PERSON: "John Doe" — Successful

GPE: "Paris" — Successful

EMAIL: "john.doe@example.com" — Regex Enhanced

Ollama (External)

```
load_tensors: offloading 32 repeating layers to GPU
load_tensors: offloading output layer to GPU
load_tensors: offloaded 33/33 layers to GPU
load_tensors:        CUDA0 model buffer size =  4097.52 MiB
load_tensors:   CPU_Mapped model buffer size =    72.00 MiB
llama_context: constructing llama_context
llama_context: n_seq_max     = 2
llama_context: n_ctx         = 8192
llama_context: n_ctx_per_seq = 4096
llama_context: n_batch       = 1024
llama_context: n_ubatch      = 512
llama_context: causal_attn   = 1
llama_context: flash_attn    = 0
llama_context: freq_base     = 1000000.0
llama_context: freq_scale    = 1
llama_context: n_ctx_per_seq (4096) < n_ctx_train (32768) -- the full capacity of the model will not be utilized
llama_context:  CUDA_Host  output buffer size =     0.28 MiB
llama_kv_cache_unified: kv_size = 8192, type_k = 'f16', type_v = 'f16', n_layer = 32, can_shift = 1, padding = 32
llama_kv_cache_unified:      CUDA0 KV buffer size =  1024.00 MiB
llama_kv_cache_unified: KV self size  = 1024.00 MiB, K (f16):  512.00 MiB, V (f16):  512.00 MiB
llama_context:      CUDA0 compute buffer size =   560.00 MiB
llama_context:  CUDA_Host compute buffer size =    24.01 MiB
llama_context: graph nodes  = 1094
llama_context: graph splits = 2
time=2025-07-19T13:48:13.128Z level=INFO source=server.go:637 msg="llama runner started in 1.51 seconds"
[GIN] 2025/07/19 - 13:48:13 | 200 |  2.779631257s |       172.18.0.2 | POST     "/api/generate"
```

Technical Contributions

1. **Web3 Integration with Privacy-by-Design**

   Implementation of DID-based authentication and data ownership verification through Clio-X platform, ensuring complete user control over personal data.

2. **Trust Boundary Separation**

   External Ollama LLM server maintained as independent service, creating a clear security boundary between Web3 infrastructure and AI processing components.

3. **Automated Sensitive Entity Masking**

   Fully automatic detection and masking of all sensitive information (PERSON, EMAIL, GPE) through combined NER and regex approaches before external processing.

Future Work & Extensions

- **Multilingual expansion** to support diverse global Web3 communities beyond English-centric interfaces.
- **On-chain AI service integration** for fully decentralized, transparent AI execution with smart contract governance.
- **Enhanced privacy protection** through differential privacy techniques to further prevent information leakage.
- **Scalability improvements** for Web3-Enterprise and public Web3 service applications.

> *"The future roadmap aims to establish this framework as the standard for privacy-preserving AI systems in decentralized environments."*

Multilingual Support

Cross-language privacy preservation with culture-specific entity recognition

On-chain AI Services

Blockchain-verified AI execution with decentralized governance

Differential Privacy

Advanced statistical techniques to prevent inference attacks

Enterprise Scalability

High-throughput architecture for organizational Web3 adoption

Conclusion

- Demonstrated a robust framework for privacy-preserving AI chatbots in Web3, addressing critical needs in post-pandemic digital transformation.
- Clio-X integration ensures user control, auditability, and native Web3 compatibility through decentralized identity verification and blockchain validation.
- Sets the foundation for secure, privacy-aware conversational AI in decentralized ecosystems with potential applications in Web3- Enterprise and public services.

Thank You

Questions?