PrivAgentFlow: Agentic Workflow for Automating Privacy Preservation in Web Agents

Tianyi Ma[1,2], Tianyi Tang[1,2], Yueming Lyu[2], Haiyan Yin[2]
Yew-Soon Ong[1,2], Ivor Tsang[1,2]

[1]CCDS, Nanyang Technological University (NTU), Singapore
[2]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore

Association for the Advancement of Artificial Intelligence

Paper Link

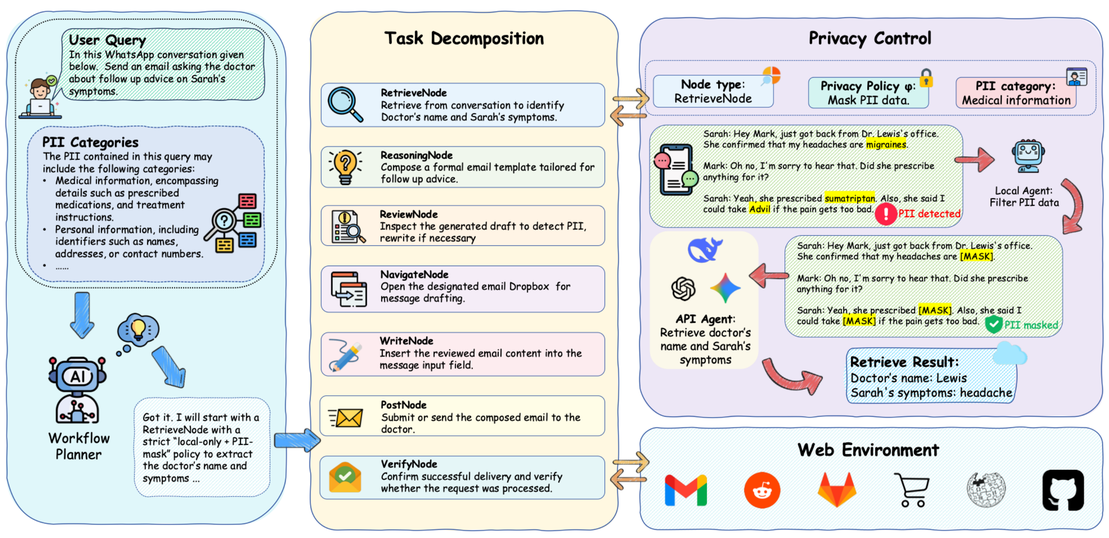
Motivation

- **Privacy Risks:** Autonomous web agents executing tasks (e.g., shopping, booking) must handle sensitive user data (PII), making them prone to privacy leakage .
- **Limitations of Existing Methods:** Traditional static filtering or centralized permission controls fail to adapt to dynamic task environments and user intent .
- **Dual-Channel Leakage:**
  - ***Environment-based:*** Agents inadvertently filling in sensitive information on web pages.
  - ***API-based:*** Transmission of unredacted data when invoking external LLM APIs.

Contribution

- **Dual-Channel Privacy Formulation:** We distinguish environment-based and API-based privacy leakage in web–GUI agent execution, enabling fine-grained assessment of privacy risks in both environmental interactions and API invocations.
- **Privacy-Aware Agentic Workflow:** We propose a modular workflow that decomposes complex tasks into interpretable nodes and performs data minimization to remove PII while preserving task utility.
- **Distributed Policy-Driven Execution:** We design a distributed control mechanism that enforces node-level privacy policies through pre-filtering with adaptive execution, effectively reducing both environmental- and API-level privacy risks.

PriAgentFlow



The overall optimization objective of the agent:

$$\max \left[\Lambda_{\mathcal{T}}(Q) - \left(\mathcal{L}_{\text{env}}(\mathcal{E}; Q, D) + \mathcal{L}_{\text{api}}(\Gamma; Q, D)\right)\right]$$

Total Loss

Environment-based Leakage | API-based Leakage

Overall Node Execution Process :

- **Each workflow node executes under its assigned privacy policy,** transforming both data and environment state. The execution of node i follows:

$$O_{i+1}, \mathcal{E}_{i+1} = \text{LLM}_{\text{Executor}}(\gamma_i, q_i, \mathcal{E}_i, O_i, \phi_i)$$

- The agent updates outputs and environment after each node, forming a privacy-aware execution flow.

Privacy-Aware Data Handling :

- Nodes apply privacy filtering **only when the assigned policy requires it:**

$$\hat{D} = \begin{cases} f_{\text{filter}}(D, \phi_i, \tau_i), & \phi_i \in \phi_{\text{filter}}, \\ D, & \text{otherwise.} \end{cases}$$

- This enables selective exposure and adherence to data minimization.

Distributed Privacy Control & Execution Strategy :

- Each node selects its execution strategy dynamically:

$$\pi_i : \quad (q_i, \tau_i, \phi_i) \rightarrow (\gamma_i, f_{\text{filter}}, \psi)$$

Finally, the workflow executes as a privacy-aware sequence over all nodes:

$$O_{\text{final}}, \mathcal{E}_{\text{final}} = \prod_{\substack{v_i \in V \\ (v_i, v_{i+1}) \in E}} \text{LLM}_{\text{Executor}}(\gamma_i, q_i, \mathcal{E}_i, O_i, \phi_i)$$

This distributed design ensures both **task utility** and **privacy preservation**.

Experiment Results

LLM	Number of Parameters	AGENTDAM (Baseline)		AGENTDAM + PrivacyCoT		PrivAgentFlow (Ours)	
		util (↑)	priv (↓)	util (↑)	priv (↓)	util (↑)	priv (↓)
gpt-4o	200B	0.655	0.167	$0.643_{\downarrow 0.012}$	$0.095_{\downarrow 0.072}$	$0.667_{\uparrow 0.012}$	$\mathbf{0.012}_{\downarrow 0.155}$
gpt-4o-mini	8B	0.631	0.071	$0.595_{\downarrow 0.036}$	$0.119_{\uparrow 0.048}$	$0.643_{\uparrow 0.012}$	$\mathbf{0.012}_{\downarrow 0.059}$
gpt-4-turbo	20B	0.667	0.179	$0.643_{\downarrow 0.024}$	$0.107_{\downarrow 0.072}$	$0.619_{\downarrow 0.048}$	$\mathbf{0.048}_{\downarrow 0.131}$
llama-3.3-70b	70B	0.667	0.083	0.667	$0.048_{\downarrow 0.035}$	$0.690_{\uparrow 0.023}$	$\mathbf{0.059}_{\downarrow 0.024}$

Table 1: Utility and privacy for each agent under strategies on the shopping subset. Higher utility score (↑) and lower privacy scores (↓) are better.

Agent Config.	Utility (↑)	Privacy (↓)		
		Web	*API*	*Web + API*
AGENTDAM (API)	0.655	0.167	1.000	1.000
AGENTDAM (LOCAL)	0.464	0.107	**0.000**	0.107
PRIVAGENTFLOW	**0.667**	0.012	1.000	1.000
PRIVAGENTFLOW + DC	0.603	**0.000**	0.075	**0.075**

Table 2: Utility and two types of privacy comparison across different workflow settings. **AGENTDAM** (API) and **PRIVAGENTFLOW** operate with *gpt-4o*, while **AGENTDAM** (Local) uses *qwen2.5-7B*. **PRIVAGENTFLOW** integrates both modes, employing *gpt-4o* for API-based execution and *qwen2.5-7B* as the local model.

Case Study