WOPA: Intelligent Chat Safeguarder

Philosophy is All Agent Needs for Collaboration

Students: Shucheng Fang, Yongcheng Liu
Advisor: Larz White

Introduction

Motivation

- In *real-time chat* environments (SMS, WhatsApp, Discord, etc.), users often encounter unfamiliar people, files, messages, and links, but they must access due to fast-pace nature of real-time chats.
- Even security-savvy individuals struggle to *quickly judge credibility and safety* in these situations
- Users are vulnerable to attacks like phishing, malware, and/or privacy issues.

WOPA

- WOPA is a PoC (proof of concept) project exploring and showcasing *using LLMs and sandboxing* to provide proactive, user-friendly *protection in real time messaging* scenarios.
- WOPA integrates AI-empowered sandboxing, log analysis, and visual-based behavioral simulation tools to conduct checks against messages, links, files, and app users encounter during chat.

Background

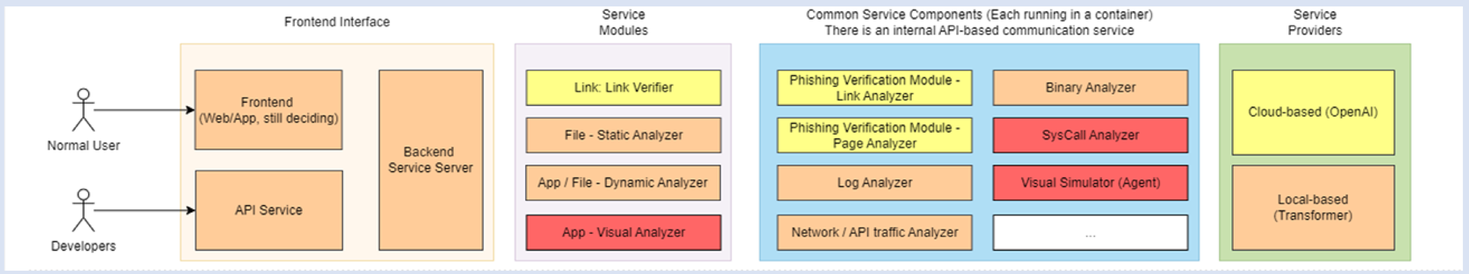
Current Approaches

- Traditional solutions rely heavily on *pattern-matching*, static ML models, and *large labeled datasets*.
- Traditional solutions are *limited to known threats*, they *struggle to adapt* to new, unseen attacks..

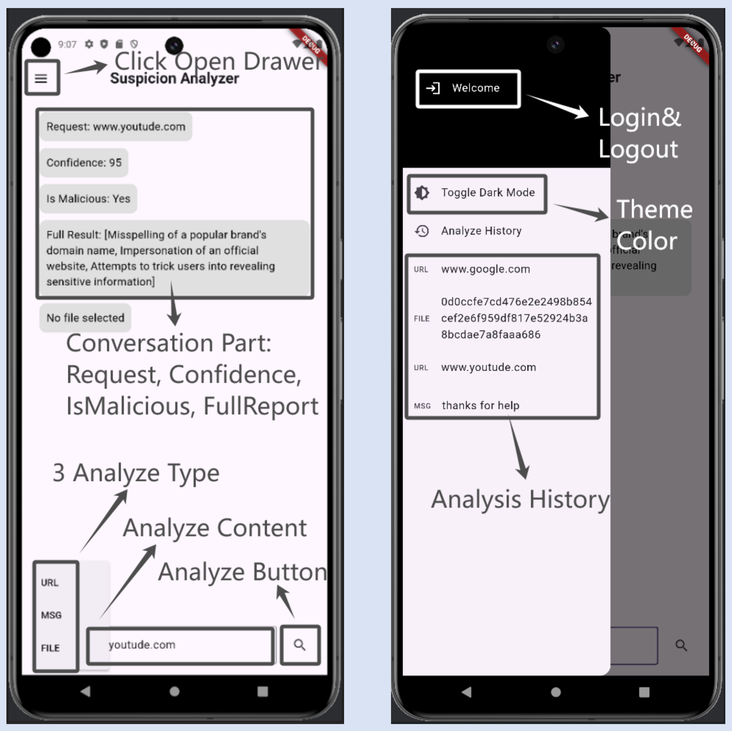
Emerging LLM Techs

- Emerging Large Language Models (LLMs) offer *potential to deliver more intelligent, context-aware analysis* beyond static rule-based checks.
- LLMs can interpret context, language nuances, and subtle signals in messages and websites, making them well-suited for dynamic threat detection.

System Architecture



Demo



Sample Workflow

(Link Analysis)

STEP	*EXAMPLE DATA*
[0] Got Link Analysis Task	Input URL Example: https://example.com
[1] Test Reachability	Check get https://example.com return code == 200
[2] Analyze page content	2.1 Fetch ALL page content (HTML + JavaScripts) 2.2 LLM analyze each content
[3] LLM Analysis & Aggregation	3.1 Ask LLM: Does "https://example.com" seem valid? 3.2 Aggregate (weighted) all results from above

Core Functionalities

Chat

Worried about a scam your grandmother received? Let her ask WOPA! Fast and clear advice, no technical background needed.

Message Analysis

Not sure if a strange message is legit? WOPA reads the text, checks context, and help you identify all potential issues.

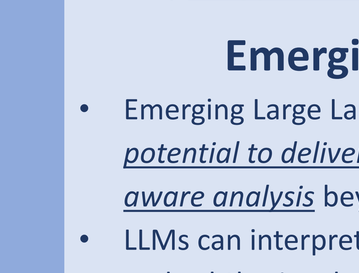
Link Analysis

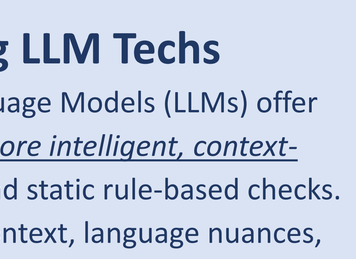
Got a suspicious URL from your "friend"? WOPA scans and analyzes the link, taking the risk for you to ensure you are all time safe.

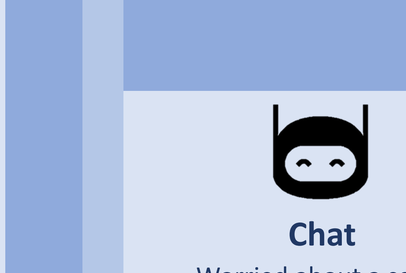
File Analysis

Friend send you "game to earn $100 when play"? WOPA runs it in a safe environment, visually play it as you, and identify threats before you enjoy.

Evaluations & Findings







Insights (What's Working Well):

- **Adaptability and Deep Reasoning:** The system runs without any training (zero-shot), provides "smart" responses – it can even analyze the smallest piece of context (like JavaScript) on the target and use it to yield meaningful results.
- **Speed:** Both local models and online models (gpt-4o/gpt-4o-mini) can perform comprehensive message analysis within seconds and conduct link analysis (requiring scraping & analyzing the source of the page contents, which is very extensive) within minutes (with high variation between models → this needs model-specific adjustment).

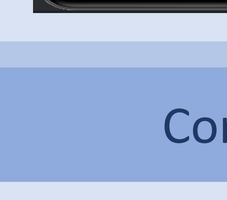
Challenges:

- **High Error Rate & False Positives (Reliability):** The evaluations reveal that LLM suffer from high error rates (especially local), and none of the models have 90%+ accuracy. We investigated this and found this is due to:
  - **JSON Validations:** All local model errors are due to inability to respond in JSON format after 3 trials.
  - **High False Positives:** All models suffered from high false positive rates (which is the direct and only contributor to accuracy issues), and we speculate it relating to LLM hallucinations issues..

Findings:

- **LLM security workers are working logically, though fine tuning is needed to reduce false positive issues**

The Future – Philosophy is All Agent Needs

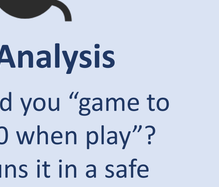


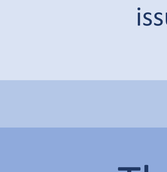
From Separation to Collaboration: Philosophy as the Last Missing Piece

- **Existing Practices – Agent Chaining:** Current LLM-based workflows often group or chain multiple "agents" to tackle separate steps. While this can solve isolated, mutually independent tasks, when it comes to dependent projects, it frequently lacks a cohesive vision, sometimes leading to fragmented or suboptimal results.
- **The Missing Piece:** From this project's explorations & experiments, we speculate that the lacking piece is a core set of guiding philosophies—high-level design principles that teach how all agents should think, plan, and act, alike human worker onboarding.

A 7-Day Miracle: What We Get after Giving Agents Philosophies?

- **Our New Approach:** Inspired by the methodologies used in this project, we automated them for our agents. Surprisingly, this gave every worker a clearer "big picture" perspective, significantly reducing their errors and improving overall consistency.
- **Consistency and Quality:** As a direct outcome, we rebuilt the entire codebase from scratch in just one week with minimal human intervention. The result was a more coherent, modular, adaptable, and scalable system full of comments.
- **Easy Integration with Any Code:** We also tried to integrate in an open-source novel cybersecurity research idea, MobileAgent (using LLM-vision for automating app control), and within hours, our system is able to use the code designs from it to automate UI tests.





Attention is All You Need, Philosophy is All Agent Needs

- **Philosophy is a special type of Memory:** Philosophies empower agents to adapt, learn, and retain context from a record of key principles captured at critical moments.
- **Backbone Work Guidelines:** Agents can use philosophies as high-level guidelines to maintain consistency and coherence, navigate complexity, and evolve over time without losing sight of their overarching goals.