EXTENDED ABSTRACT

CensorAlert - Leveraging LLM Agents for Automated Censorship Report Aggregation and Analysis

Ali Zohaib, Jade Sheffey, Mingshi Wu, Amir Houmansadr

FOCI 2026

University of Massachusetts Amherst

GFW Report

MOTIVATION

Censorship Reporting is Fragmented

Measurement Platforms

- OONI, Censored Planet, Cloudflare Radar, NetBlocks
- Continuous, systematic monitoring
- Limited by geography & protocol coverage

Crowdsourced Reports

- Forums, chat groups, social media, issue trackers
- Multilingual, informal, scattered
- Often the first signal of new censorship

No single platform to track censorship incidents across all these channels

MOTIVATION

Why This is a Problem

- Measurement platforms miss localized or novel censorship methods
- Crowdsourced reports are scattered across **dozens of platforms and languages**
- Critical findings depend on whether a **researcher sees the right post**

Examples:

- GFW blocking of fully encrypted traffic → discovered via user forum post
- SNI-based QUIC Blocking in China → identified from a Telegram post

Current pipeline relies on a small group of overburdened volunteers

OUR APPROACH

Introducing CensorAlert

A platform that **automates** aggregation, assessment, and delivery of censorship reports

Aggregate

Collect from measurement APIs, forums, social media, news websites etc

Analyze

LLMs classify, translate, summarize, and score each report

Alert

Deliver a ranked feed and real-time notifications to users

CensorAlert

https://censoralert.org

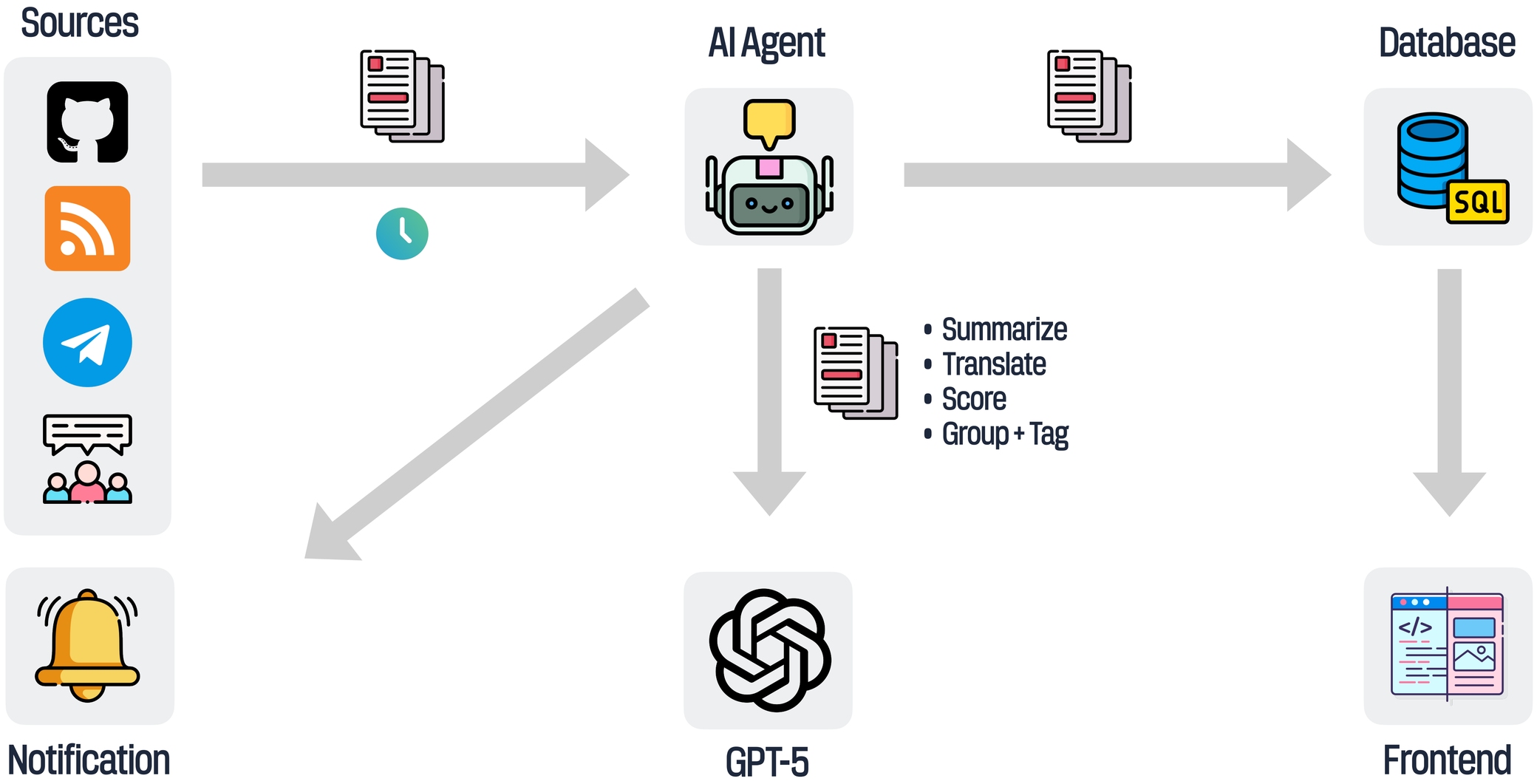
SYSTEM DESIGN

Architecture Overview

- Built on **N8N**: open-source workflow automation platform
- Each workflow: fetches data → calls AI agent → parses results → writes to DB → sends alerts
- Triggered on a **timed interval** (every 2 hours)
- Modular and easily extendable

SYSTEM DESIGN

How it Works



SYSTEM DESIGN

Data Sources

Measurement Platforms	OONI, Cloudflare Radar, NetBlocks, IODA
Social Media & Forums	NTC Party, Mastodon, Net4People BBS, Telegram Channels, Github Issues e.g. XTLS, Hysteria
Research Papers	arXiv

SYSTEM DESIGN

LLM Processing

- For each ingested item, the LLM agent performs:
  - **Classification:** Is this relevant to censorship?
  - **Summarization:** Rewrite title, generate English summary, add tags
  - **Significance Scoring:** Credibility, Novelty, Impact, Timeliness
- LLM behavior constrained via system prompt
- **Output:** Each item gets a normalized score between 0-10

SYSTEM DESIGN

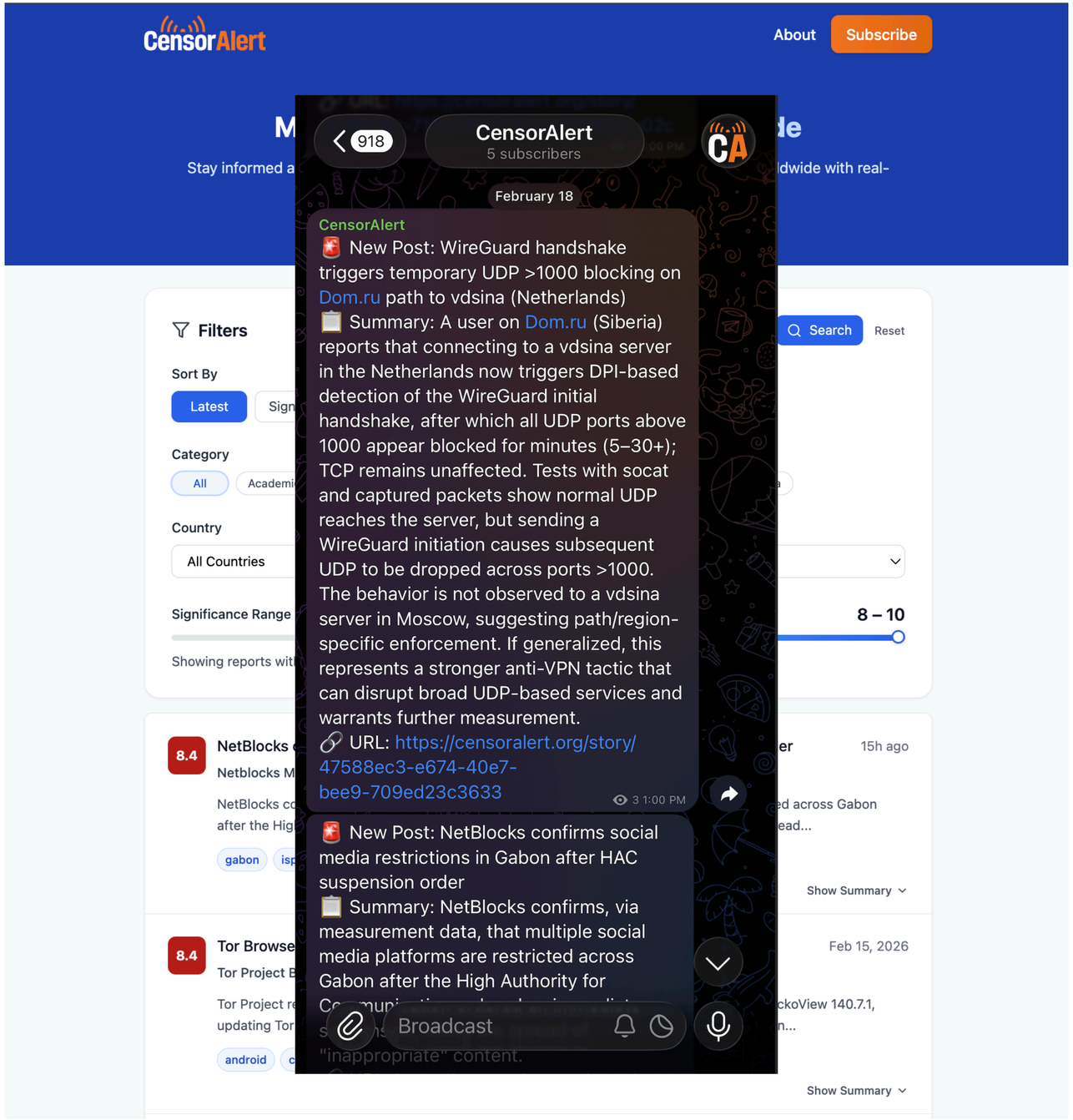
Clustering & Deduplication

- An event might get reported on multiple sources: deduplication and clustering prevents feed/notification overload.
  - We cluster semantically similar items.
  - Collapse near-duplicates (reposts, translations) into a single entry.
  - All original source URLs and metadata are **preserved.**

SYSTEM DESIGN

Frontend

https://censoralert.org



SYSTEM DESIGN

Future Work

- Expand source coverage.
- Add more notification channels: Signal, RSS etc.
- **Long-term:** AI-driven measurement and circumvention.

We value your feedback. Please try censoralert.org and tell us what's working, what's missing, and what would make this useful for you.

Thank you!