Consent-Based LLM-OSINT for Person Profile Compilation: An End-to-End Agentic Pipeline with MCP Tooling, Dual Retrieval, and Graph Cleanup

Jingbin Lin, Frederick Pi, Jiwen Luo, Hongyi Pan
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093

***Abstract*—Large language model (LLM) agents make open-source intelligence (OSINT) much easier by combining search, tool use, extraction, and synthesis in a single workflow. At the same time, mainstream commercial assistants often refuse requests that resemble doxxing or unauthorized personal profiling. This creates a gap between technical capability and product availability. We examine that gap in a consent-based setting using only subject-authorized public data. We present an automated end-to-end LLM-OSINT pipeline that starts from only a person's name and produces a structured, citation-grounded profile. The pipeline combines tool-based retrieval, evidence retention, entity and relation extraction, dual storage in a vector index and a knowledge graph, and a graph-cleanup stage for canonicalization, deduplication, and contradiction reduction. Our findings are qualitative but consistent. First, commercially aligned assistants such as Claude and ChatGPT Deep Research often refuse direct person-profile compilation tasks, whereas a custom consent-based pipeline can still assemble a surprisingly rich profile from fragmented public traces. Second, the system's value comes not only from retrieval breadth but also from normalization: graph cleanup reduces contradictions caused by competing aliases, stale affiliations, and duplicated entities. Third, the main privacy risk is not the discovery of secret information, but the low-cost centralization of public yet distributed information into a coherent output. Together, these results show that agentic LLM systems materially increase the scale, speed, and coherence of OSINT.**

1. Introduction

Open-source intelligence (OSINT) has traditionally been labor-intensive. Analysts gather information from search engines, social networks, institutional pages, publication indexes, and archived content, then manually reconcile conflicting sources. Existing OSINT tools such as SpiderFoot automate collection breadth, but usually stop short of citation-grounded synthesis of person-level profiles, leaving semantic reconciliation and report construction to the analyst.

LLM-based agents change this workflow. They can orchestrate multi-step search, browsing, extraction, and synthesis, while also translating unstructured webpages into normalized entities, attributes, and relations. This makes it much easier to turn scattered public traces into coherent person-level profiles.

That shift creates a practical privacy and security concern. The issue is no longer only that public information exists, but that it can be cheaply centralized into a usable dossier. At the same time, mainstream consumer-facing assistants increasingly refuse requests that resemble doxxing, surveillance, or privacy-invasive profiling. This creates a gap between technical feasibility and product availability.

We study that gap in a controlled, consent-based setting. Given only a person's name, how much accurate and relationally useful information can an LLM-agent system collect from public sources and consolidate into a single cited profile? To reduce ethical ambiguity, all targets are members of our own team, and all retrieved information comes from the subjects' own public profiles.

We implement a fully automated pipeline with tool-mediated retrieval, artifact retention, extraction, dual storage in a vector index and knowledge graph, and a graph-cleanup stage that resolves aliases, deduplicates entities, and reduces contradictions. Our central hypothesis is that person-profile compilation depends not only on retrieval breadth, but also on structured consolidation. The paper's main contribution is a systems-level evaluation of this emerging threat surface. We argue that the security significance of LLM-OSINT lies in tool-mediated collection breadth, low-cost semantic structuring, and graph-based contradiction reduction. More broadly, our findings suggest that the difference between scattered public information and a usable personal dossier is increasingly a matter of orchestration.

2. Related Work

2.1. Open-Source Intelligence (OSINT)

Open-source intelligence (OSINT) is commonly described as intelligence derived from publicly (and often commercially) available information through a systematic workflow of collection, processing/exploitation, analysis, and dissemination to support an intelligence requirement. This process-oriented framing appears both in operational policy and in recent security-studies accounts emphasizing that OSINT is not merely "open data," but an analytic production

pipeline with recurring challenges in reliability, scale, and governance [1], [2].

Widely used OSINT tooling largely optimizes the early stages of this workflow. For example, SpiderFoot automates broad collection across many public data sources; Maltego emphasizes link analysis through entity-relationship visualization; and Sublist3r targets subdomain enumeration for reconnaissance and investigation [3], [4], [5]. However, these tools typically stop short of end-to-end, person-profile synthesis that reconciles conflicting sources and preserves provenance: analysts still perform semantic normalization (e.g., aliases, deduplication), contradiction handling, and report writing.

Academic work increasingly applies machine learning to OSINT-adjacent tasks, but systematic reviews suggest much of this literature still focuses on point solutions (e.g., classification/detection after data collection) rather than the orchestration and evaluation of full OSINT pipelines that integrate collection, reconciliation, and reporting [6]. Our work targets this "end-to-end" gap by studying OSINT-style profiling as an agentic workflow with explicit evidence linkage and consolidation.

2.2. LLM Agents and Tool Use

Tool-augmented LLM systems provide a qualitatively different automation path than single-task models: they can iterate between reasoning and actions such as searching, browsing, or querying external tools. WebGPT demonstrates browsing with reference collection to support long-form answering, while ReAct formalizes an interleaved reasoning–acting pattern that uses tool interaction to reduce hallucination and improve interpretability [7], [8]. Recent surveys consolidate design patterns for LLM-based agents, including planning loops, memory, tool use, and multi-agent coordination, providing a standard vocabulary for framing agentic pipelines beyond ad hoc prompting [9].

Our system builds on these agentic ideas but in a different problem setting: person-profile compilation typically requires repeated retrieval, cross-source reconciliation, and structured reporting. For implementation, we use LangGraph as an orchestration layer for stateful and controllable long-running agent workflows (e.g., durability/persistence and human-in-the-loop control), which is particularly relevant when OSINT workflows include verification and interruption/resumption [10]. This positions our contribution as an applied, measurable instance of agentic OSINT-style synthesis rather than a single-pass "LLM with tools" demo.

2.3. Retrieval-Augmented Generation and Citation Grounding

Retrieval-augmented generation (RAG) combines parametric generation with access to an external document index, motivating provenance and updatability for knowledge-intensive tasks [11]. In practice, however, "retrieval present" does not guarantee statement-level grounding, particularly when outputs synthesize across many sources.

Accordingly, recent work evaluates not only answer correctness but also attribution and citation quality. ALCE introduces an end-to-end benchmark for generating text with citations and measuring citation quality, while AIS proposes a statement-level notion of attribution to identified sources with a structured human annotation procedure [12], [13]. In parallel, FActScore argues for decomposing long-form text into atomic facts and computing the fraction supported by trusted sources, enabling more diagnostic evaluation than binary "correct/incorrect" judgements [14]. Critically, empirical evidence shows that even when LLMs provide URLs, many statements remain unsupported or contradicted by the cited sources, motivating explicit supportiveness checks rather than "citation presence" as a proxy for truthfulness [15].

Our evaluation and reporting choices align with this literature by treating evidence linkage and statement-level support as first-class objectives (rather than incidental formatting), and by measuring citation and factual support at a fine granularity suitable for profile compilation.

2.4. Entity Resolution and Graph Cleanup

A central difficulty in person-profile compilation is cross-source identity and attribute consolidation: multiple mentions may refer to the same real-world entity, while sources may conflict or be stale. This problem connects directly to record linkage and entity resolution. Fellegi–Sunter provides foundational theory for matching records under uncertainty with explicit error tradeoffs, and Swoosh frames entity resolution as iterative match-and-merge with generic match/merge operators and efficiency properties [16], [17]. More recent work shows that pretrained language models can be effective for entity matching in heterogeneous text records (e.g., Ditto), motivating modern approaches to canonicalization beyond string heuristics [18].

These strands motivate treating graph construction and cleanup (deduplication, alias resolution, and conflict management) as a first-class stage rather than a byproduct of retrieval. In our pipeline, graph-aware consolidation is explicitly evaluated as part of coherence and contradiction reduction.

2.5. Privacy and Policy Frameworks

OSINT profiling raises privacy concerns even when inputs are derived from publicly accessible sources, because aggregation and repackaging can produce new capabilities and harms. Classic privacy scholarship argues that the key risk is often increased accessibility and aggregation ("digital dossiers") rather than secrecy loss, and contextual integrity frames privacy harm as a violation of context-relative information-flow norms even for "public" data [19], [20]. Recent OSINT–law scholarship also highlights the growing role of commercially available information and the need to balance national security objectives with privacy and data protection principles in modern OSINT practices [21].

From a governance standpoint, the NIST Privacy Framework provides a risk management vocabulary for privacy

that is compatible with system design and evaluation, and the GDPR provides a widely cited regulatory reference point for principles related to processing personal data (including profiling, purpose limitation, and accountability) [22], [23]. We use these perspectives to motivate (i) consent-based study design and (ii) evaluation that treats consolidation and evidence handling as privacy- and safety-relevant system properties.

3. Method

3.1. Task Definition

We study consent-based person-profile compilation from public sources. The system receives only a person's name and must produce a structured profile containing factual attributes and relations, each linked to evidence. The profile may include professional affiliations, education, publications, web presence, and organizational associations, but it must not rely on non-public access, credential compromise, or inferred sensitive attributes.

We frame the system as a privacy-risk measurement instrument, not as an unrestricted surveillance tool. All subjects consent to the evaluation, all sources are public, and outputs are audited through citations and manual verification. The goal is to measure how much information can be assembled automatically when the underlying data is already public but distributed.

3.2. System Overview

At a high level, the system is organized as a two-stage agentic architecture connected by a shared knowledge layer. A user prompt or target specification is first submitted through an API endpoint, which creates a run and launches Stage 1. Stage 1 is responsible for iterative evidence collection, target-centered retrieval, candidate structuring, and normalization. Its outputs are then handed off to the knowledge layer, where evidence is stored in both a typed graph and a vector index. Stage 2 retrieves from this knowledge layer to generate an evidence-grounded final report, which is then returned through the API/web interface. As shown in Fig. 1, the pipeline separates collection and normalization from final report synthesis.

Stage 1 is designed as a planner-worker loop rather than a linear pipeline. It begins with input analysis and target anchoring, then performs planning, retrieval, and graph assembly. A planner selects tool batches, and specialized tool workers execute them in parallel to improve throughput and reduce end-to-end latency. Intermediate outputs are reviewed through receipt inspection, followed by a coverage-and-quality gate that determines whether additional collection is needed. When evidence is incomplete, noisy, or contradictory, the system iterates through further retrieval and follow-up steps. Before handoff, the collected evidence is passed through conflict adjudication and graph normalization so that downstream synthesis operates over more stable structured candidates rather than raw tool output.

The knowledge layer provides the bridge between collection and synthesis. It combines a Neo4j typed graph for canonicalized entities and relations with a vector index for semantic retrieval over heterogeneous evidence. This dual representation reflects the needs of the task: the graph supports typed aggregation, entity reconciliation, and contradiction handling, while the vector index supports flexible retrieval of supporting passages during report generation. By separating these roles, the architecture avoids collapsing noisy intermediate artifacts directly into final profile fields.

Stage 2 performs evidence-grounded report synthesis. It first retrieves relevant graph and vector evidence, builds an outline, and then drafts report sections conditioned on retrieved support. A verification-oriented evidence-packing step checks claims against the retrieved material before finalization. The stage also includes a revision loop, allowing the system to redraft sections when support is incomplete or organizational quality is insufficient. This design keeps final synthesis tied to retained evidence rather than unconstrained model recall.

Overall, the architecture separates retrieval, normalization, and evidence accumulation from report synthesis and claim verification. This decomposition improves robustness under noisy public-source collection, supports scalable execution through parallel workers, and preserves traceability from raw evidence to final profile statements.

3.3. Input analysis and target anchoring

The input-analysis and target-anchoring stage initializes the retrieval process around a stable representation of the intended subject. Because person names may be ambiguous, incomplete, or shared across multiple individuals, the system first parses the user prompt for identity cues such as full name, aliases, affiliations, occupations, locations, and other contextual hints. These cues are normalized into an initial target record that serves as the reference state for downstream planning, retrieval, and evidence reconciliation. This anchoring step is important because the collection stage is recall-oriented and therefore vulnerable to contamination from similarly named entities. By constraining subsequent tool selection, follow-up queries, and graph assembly around a target-centered state, the system reduces identity drift and improves the precision of later profile synthesis.

3.4. Collection Layer and Tooling

The collection layer combines browser-based retrieval with modular MCP-integrated tools spanning multiple source types and providers. In our system, MCP provides a uniform interface to external retrieval utilities, while LangGraph provides the stateful orchestration layer for planner-worker execution, iterative control flow, and report-oriented subgraphs. The collection process is therefore not implemented as a single monolithic agent, but as a graph-structured workflow in which specialized nodes perform planning, tool execution, receipt review, and follow-up decisions.

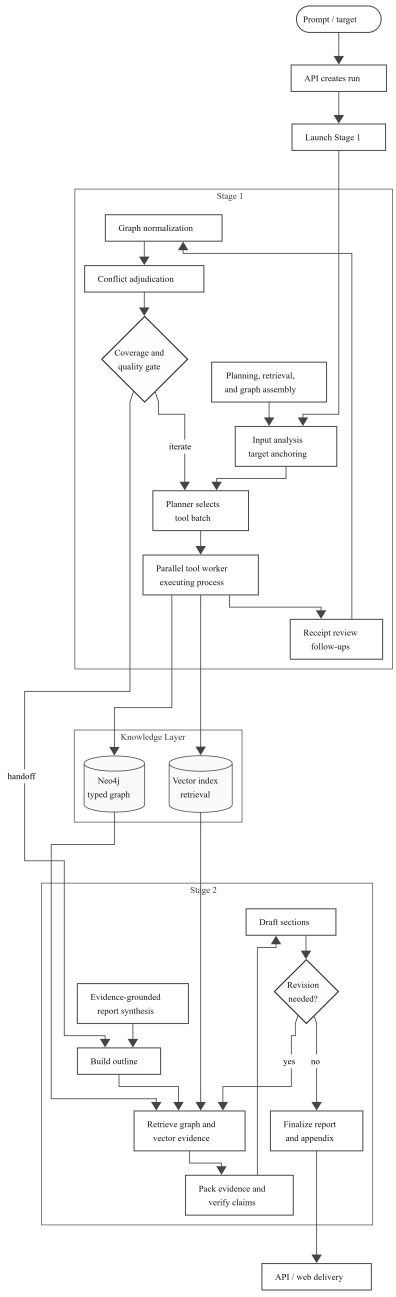

Figure 1. Architecture of the consent-based LLM-OSINT pipeline. Stage 1 performs retrieval, planning, normalization, and graph assembly; the knowledge layer stores canonicalized graph and vector evidence; Stage 2 performs evidence-grounded report synthesis and revision.

The tool layer includes profile-oriented retrieval, scholar metadata discovery, organization search, multi-provider web search, and browser-based evidence collection. These tools expose disambiguating signals such as employment history, affiliations, publication records, coauthor relations, and public profile identities. Collection is driven by an iterative planner loop: early steps prioritize high-confidence identity anchors, while later steps focus on narrower subproblems such as publications, affiliations, organizational roles, and corroborating evidence. The goal is not merely to gather relevant text, but to accumulate evidence that can later be normalized into stable typed entities, attributes, and relations.

3.5. Artifact Retention and Provenance

A central design principle of the system is artifact retention. Each collection step preserves sufficient trace data to support downstream auditing, including raw tool responses, retrieved URLs, snippets, summaries, timestamps, extracted candidates, and source associations. In the implemented

system, raw outputs and retrieved artifacts are stored in MinIO-backed object storage, while database references link these retained materials to later extraction, graph construction, and synthesis stages.

Artifact retention is necessary because person-profile compilation is vulnerable to subtle errors such as stale affiliations, name collisions, duplicated publication metadata, and indirect references propagated through third-party sites. Preserving intermediate evidence makes it possible to determine whether an error originated during collection, extraction, normalization, or synthesis. Retention also enables contradiction analysis: when incompatible candidate values appear for the same field, the system can trace them back to their original supporting artifacts instead of relying only on compressed summaries or final synthesized text.

3.6. Extraction into Structured Candidates

After collection, the system converts retained artifacts into structured candidates. At this stage, the system aims to maximize recall rather than resolve truth. It extracts entity candidates (persons, organizations, institutions, venues, websites), attribute candidates (job title, affiliation, degree, publication title, year), relation candidates such as employment, education, authorship, coauthorship, and membership links, and provenance links indicating which source artifact supports the claim.

The extraction step must tolerate redundancy because the same underlying fact may appear in multiple lexical forms. For example, an institution may appear as a full legal name, an acronym, or a shortened alias; a person may appear with initials on a publication page and a full name elsewhere. Over-aggressive deduplication here can reduce recall. The pipeline therefore errs on the side of candidate overproduction and defers canonicalization.

3.7. Dual Storage: Vector Index plus Knowledge Graph

The system stores extracted evidence in both a vector index and a knowledge graph. These two representations address different needs.

The vector index supports semantic retrieval over heterogeneous evidence and is useful when different sources describe the same concept in different language. However, vector retrieval alone is weak at contradiction management because it does not explicitly distinguish between alias variation and entity mismatch.

The knowledge graph supports typed relational reasoning. In the person-profile setting, this matters because the task is not simply to retrieve similar passages, but to determine whether multiple mentions refer to the same person, institution, publication, or role and whether the relations among them are mutually compatible.

3.8. Graph Cleanup

The graph-cleanup stage is the core normalization component of the system. It is designed to reduce contradictions and improve the stability of final profiles through four operations: alias resolution, which merges nodes likely to refer to the same real-world entity; type normalization, which enforces consistent typing for persons, organizations, venues, and profile resources; relation deduplication, which collapses repeated edges expressing the same claim; and conflict arbitration, which selects canonical values when candidates disagree.

Conflict arbitration uses a hybrid strategy. In straightforward cases, the system applies support-based normalization and deduplication rules based on source frequency, cross-source consistency, and structural compatibility in the graph. When these traditional normalization steps are insufficient to resolve ambiguous or conflicting candidates, an LLM-based agent is introduced to perform deeper conflict detection and adjudication. Candidate values therefore become more plausible when they are supported by multiple independent artifacts, remain compatible with neighboring relations, and attach consistently to the same canonical entity, while weak or isolated candidates remain low-confidence or are omitted.

This stage differentiates the system from a vector-only baseline. A vector-only approach can retrieve relevant passages, but it leaves contradiction handling largely to final synthesis. Graph cleanup instead turns profile construction into a structured aggregation problem.

3.9. Profile Synthesis

The final stage synthesizes the profile as a structured, citation-grounded report. Rather than generating directly from accumulated memory of prior steps, the synthesis process retrieves relevant graph and vector evidence for each section, builds an outline, and then drafts content conditioned on that retrieved support. Before finalization, the system performs an evidence-packing and claim-verification step to ensure that section-level assertions remain tied to retained artifacts and normalized graph state.

The synthesis prompt is constrained to prefer supported claims, omit weakly grounded assertions, and preserve provenance wherever possible. A revision loop allows the system to re-draft sections when support is incomplete, organizational quality is poor, or retrieved evidence does not adequately justify the current wording. The resulting report is designed to satisfy three properties: coverage, auditability, and internal coherence. This makes the output useful both as a user-facing profile and as an evaluation artifact for measuring privacy exposure from public-data aggregation.

3.10. Safety Constraints

Although the system is technically capable of person-level OSINT, the present evaluation is intentionally consent-bound. All subjects are team members, all retrieved sources are limited to their own public profiles or other public-facing materials they control, and the study excludes non-public access, credentialed retrieval, and inference of sensitive personal attributes. We therefore treat the pipeline as a privacy-risk measurement instrument for public-information

aggregation rather than as a deployment template for unrestricted person targeting. These constraints are important because the purpose of the study is not to maximize surveillance capability, but to measure how much structured personal information can be automatically assembled when the underlying data is already public yet distributed across multiple sources.

4. Experiment

4.1. Research Questions

The experimental design is organized around three research questions:

RQ1. Given only a person's name, how much accurate personal and relational information can the pipeline collect from public sources?

RQ2. How does the automated pipeline compare to a human OSINT baseline in terms of accuracy, coverage, and analyst effort?

RQ3. Does the graph-cleanup layer reduce internal contradictions relative to a vector-only retrieval and synthesis baseline?

A fourth practical question emerged during system use:

RQ4. How does a custom consent-based pipeline compare to aligned commercial assistants that often refuse doxxing-like requests?

4.2. Targets and Data

The targets are consenting individuals evaluated under an explicit authorization protocol. This choice reduces ethical ambiguity, simplifies consent, and allows direct manual verification of disputed fields. It also produces a realistic but bounded public-data environment: modern professional and academic web traces are often sufficiently rich to support name-based profile compilation without any private access.

The evaluation includes $N = 10$ consenting targets. To reduce re-identification risk, we report only that the sample spans multiple identities and roles; finer-grained composition is intentionally withheld to preserve privacy. All evaluation runs reported in this paper were conducted in February 2026 using a fixed research configuration built around OpenRouter access to Qwen-series models for agent orchestration (`qwen/qwen3.5-122b-a10b` and `qwen/qwen3-32b`), together with embedding workers running on NVIDIA H100 (80GB SXM5) hardware using `Qwen/Qwen3-Embedding-0.6B`. Allowed sources included subject-owned profiles and public institutional or professional pages, such as university pages, lab pages, Google Scholar, LinkedIn, GitHub, and personal websites; excluded sources included private accounts, paywalled data, leaked data, data broker sites, and any credential-gated content. After manual OSINT collection, each consenting subject reviewed the resulting gold set and confirmed or corrected disputed fields before final scoring.

The system begins with only a person's name. No manual seed URLs, account handles, or curated dossiers are injected at runtime. Human effort is reserved for evaluation only: verifying outputs after the pipeline finishes and building the manual OSINT baseline.

4.3. Baselines

We compare against three reference conditions.

Manual OSINT baseline. A human analyst performs a conventional search-and-compile workflow over the same public web. This baseline is important because OSINT practitioners do not merely retrieve pages; they reconcile ambiguity, discard weak sources, and write reports. The manual baseline therefore provides a meaningful comparison for both coverage and effort. The analyst is not given a fixed time budget; instead, we record the elapsed time until the analyst declares the profile complete in order to measure the effort required for a comprehensive manual result. The analyst produces a human-readable bullet-point report with direct URL citations linked to the collected files, mirroring the pipeline's auditable artifact-retention model. For scoring, the manual output is mapped into the same evaluation slot schema and atomic-claim format used for the automated system, so that accuracy and coverage are compared over a common representation rather than over free-form prose alone.

Vector-only baseline. The same retrieval and extraction process runs without graph cleanup. Candidates are stored and retrieved semantically, but the system lacks the explicit canonicalization layer. This baseline isolates the value of the graph stage. The vector-only pipeline uses the same system prompt, the same per-agent token budgets, and the same stopping criteria as the full pipeline. In both conditions, execution ends when the run reaches the shared token budget or the maximum analysis-loop limit. Graph cleanup is fully disabled in the vector-only condition, and cross-source canonicalization is limited to simple hard-coded programmatic deduplication rather than the full graph-based normalization and conflict-resolution stage.

Commercial assistant comparison. We also record the qualitative behavior of commercially aligned research assistants such as Claude and ChatGPT Deep Research when asked to perform doxxing-like person-profile compilation. This is not a direct apples-to-apples capability comparison because those systems are policy-gated consumer products, not custom research pipelines. Nevertheless, the refusal behavior is substantively important because it shows the distinction between product alignment and raw technical feasibility. OpenAI Deep Research and Anthropic agent tooling are powerful and increasingly tool-connected, but both are governed by policies that sharply constrain privacy-invasive profiling and unauthorized surveillance. These comparison runs were conducted in February 2026 using a shared prompt template that asked the assistant to compile a cited person profile from only a target name under the same consent-based framing used by our internal pipelines. We score the resulting behavior using a binary rubric at the claim level and task level: a returned claim is counted as true or false under manual verification, and the overall run is marked

successful only if the assistant produces a substantive profile output rather than refusing, redirecting, or returning content too incomplete to score against the schema.

4.4. Metrics

The benchmark emphasizes five metrics.

Accuracy. We operationalize accuracy as claim-level precision: the proportion of output fields judged correct under manual verification. Let:

- P be the set of predicted atomic claims;
- G be the set of gold or human-verified atomic claims;
- $TP = |P \cap G|$.

Then:

$$\text{Precision} = \frac{TP}{|P|}.$$

Coverage. The proportion of verifiable profile fields present in the automated output relative to the human baseline or agreed gold field set:

$$\text{Coverage} = \frac{TP}{|G|}.$$

Inconsistency Rate. The proportion of final output fields that contain mutually incompatible values or unresolved conflicts. Let:

- S be the set of evaluated slots in the final profile;
- $I(s) = 1$ if slot s contains mutually incompatible retained claims;
- $I(s) = 0$ otherwise.

Then:

$$\text{Inconsistency Rate} = \frac{1}{|S|} \sum_{s \in S} I(s).$$

Citation recall and citation precision. We separate attribution quality into two complementary questions. Citation recall asks whether the evidence cited for a claim actually supports that claim, i.e., whether the claim is justified by what is cited. Citation precision asks whether the cited sources are individually relevant or necessary for the claim, rather than noisy extras. For claim i, let:

- C_i be the set of citations attached to claim i;
- $A_i = 1$ if the union of cited evidence in C_i supports the claim;
- $a_{ij} = 1$ if citation $j \in C_i$ is individually relevant or helpful for claim i.

Then we define citation recall at the claim level as:

$$\text{Citation Recall} = \frac{1}{|P|} \sum_{i} A_i.$$

We define citation precision as:

$$\text{Citation Precision} = \frac{\sum_i \sum_{j \in C_i} a_{ij}}{\sum_i |C_i|}.$$

Operationally, these labels are assigned as binary human judgments during evaluation, with subject confirmation used as the final adjudication source for G, $I(s)$, A_i, and a_{ij} whenever ambiguity or disagreement remains. A claim is treated as supported for the purpose of $A_i = 1$ if the cited evidence directly states the claim, supports it after light normalization such as alias expansion or formatting cleanup, or supports it by combination across citations under subject-confirmed interpretation; $A_i = 0$ only when the cited evidence is fully unrelated or fails to justify the claim. A citation is marked individually relevant for $a_{ij} = 1$ if it independently supports the claim on its own or provides a necessary qualifier such as date, role, affiliation, or venue; citations that are merely topically related but do not contribute claim support are scored as $a_{ij} = 0$. For inconsistency assessment, $I(s) = 1$ only when a slot contains retained claims that are mutually exclusive under the subject's binary judgment; otherwise $I(s) = 0$.

Effort/time. End-to-end runtime for the automated pipeline and human time for the manual baseline. We report automated runtime as wall-clock execution time and human effort as annotation or profile-construction time measured in minutes.

4.5. Evaluation Schema and Claim Decomposition

We evaluate over a fixed slot schema of K profile fields grouped into C categories: identity anchors, affiliations, education, publications, collaborators, contact channels, and timeline events. Each slot is decomposed into atomic claims of the form $\langle \text{subject}, \text{predicate}, \text{object}, \text{qualifiers} \rangle$, where qualifiers may include time ranges, roles, venues, source type, confidence, or other contextual constraints needed to interpret the claim. Slot mapping follows a canonical normalization rule: synonymous surface forms are mapped to the same predicate and entity type, source-specific labels are converted into the shared schema, and evidence is attached to the most specific slot that the source supports. For single-valued slots such as a canonical primary affiliation at a given time, we score exact or qualifier-compatible matches at the slot level. For multi-valued slots such as publications, collaborators, and repeated timeline events, we score at the atomic-claim level and report both micro-averaged and macro-averaged metrics across targets. A claim is considered verifiable if it is supported by at least one subject-owned or institution-owned public source, or is explicitly confirmed by the consenting subject during verification.

4.6. Procedure

For each target, the pipeline receives only the name. The system then launches Stage 1, where a planner-worker loop performs tool-based retrieval and browsing, retains artifacts, extracts candidates, and iteratively normalizes accumulated evidence through graph-structuring and conflict-resolution steps. The resulting evidence is stored in the shared knowledge layer, including both the vector index and the knowledge graph. Stage 2 then retrieves from this knowledge layer to synthesize the final profile report. All runtime steps are fully

automated. No human disambiguation or mid-run correction is inserted.

Separately, a human analyst performs manual OSINT collection on the same target under comparable scope. After both runs finish, the outputs are verified against the subject's actual public profiles and known ground truth where applicable. Conflicts, omissions, and unsupported claims are logged.

For the commercial-assistant comparison, the researchers submit doxxing-like or profile-compilation-style prompts to commercial systems and observe whether the system refuses, partially answers, or redirects to safer alternatives. The main outcome of interest here is refusal behavior, not numeric field extraction.

4.7. Implementation Notes

Three implementation choices are especially important. First, the system uses a modular MCP tool layer rather than a small number of hard-coded source integrations. This improves collection precision by letting the planner route retrieval subproblems to source-specific tools when they are higher signal than generic web search, while still keeping all tools behind one protocol. In the current implementation, this includes LinkedIn HTML capture, Tavily-based web retrieval, Google SERP person search, code-identity tools, registry and business tools, academic search tools, and selected OSINT utilities.

Second, artifact retention makes the system auditable. Raw tool outputs and derived artifacts are persisted in MinIO object storage, while a reference layer in the application database records document IDs, object keys, versions, hashes, content types, and evidence links. This creates a stable audit trail: later claims can be traced from report citations back to stored raw outputs, extracted files, and chunk-level evidence references rather than only to transient model context. That design is important for a study that aims to measure privacy exposure, because it supports inspection, replay, and source attribution instead of merely producing plausible text.

Third, the system uses a shared knowledge layer that combines graph normalization with retrieval-oriented evidence storage. Noisy extracted candidates are converted into structured entities and relations, then further reconciled through conflict-resolution steps before being used in synthesis. The key comparison in the evaluation is therefore not "LLM versus human" in the abstract, but rather "flat unconsolidated retrieval versus retrieval that is normalized, linked, and checked across a typed evidence structure." In this design, graph cleanup and conflict handling improve the stability of the knowledge layer, while final consistency is reinforced again during downstream report generation and verification.

5. Results

5.1. Profile Output

We evaluated the pipeline by generating complete reports for multiple consenting volunteers. The resulting outputs covered identity resolution, digital profiles, academic output, employment history, collaborators, public contact channels, and timeline events. In representative cases, the generated reports comprised roughly 24 pages of structured analysis grounded in retrieved artifacts and graph-backed normalization and synthesis. Because these artifacts aggregate sensitive person-level information, the full reports are not reproduced in the paper or supplementary materials. We therefore describe the outputs through aggregate properties and qualitative analysis rather than direct publication of the complete profiles. To reduce re-identification risk, subject names, handles, employer clues, and institution-specific details are masked or abstracted in the discussion below.

The output profile successfully aggregated information across multiple domains, including academic publications, institutional affiliations, digital platform identities, and collaborator networks. In a representative case, the system resolved a canonical target identity and linked it to consistent profiles across multiple public platforms through a shared handle pattern and supporting naming evidence. The resulting knowledge graph (Fig. 2) integrates relationships between the subject and multiple entities such as research collaborators, institutions, publications, and online identities. This demonstrates that the pipeline can transform fragmented public information into a coherent person-level knowledge graph.

By contrast, the ablated result without graph cleanup (Fig. 3) is visibly less coherent. It exhibits primary-anchor node drift, low-confidence node pollution, and many entities that remain split across disconnected components because duplicate nodes are not merged into a shared canonical identity structure.

5.2. Accuracy

Across the verified profile fields, most identity anchors and platform accounts were confirmed by consistent cross-source signals. For instance, two public profiles on different platforms were matched through exact handle alignment and consistent naming evidence. Academic affiliation claims linking the subject to a major research university were also supported by external scholar profile data and graph entities describing relevant research activity.

However, some claims could not be fully verified due to ambiguous sources or access limitations. For example, one startup-related affiliation lacked supporting documentation and was therefore marked as low confidence. In this case, the affiliation appeared only in a repost on a professional networking platform, and due to platform access limitations, the pipeline was not able to collect useful corroborating information beyond that mention.

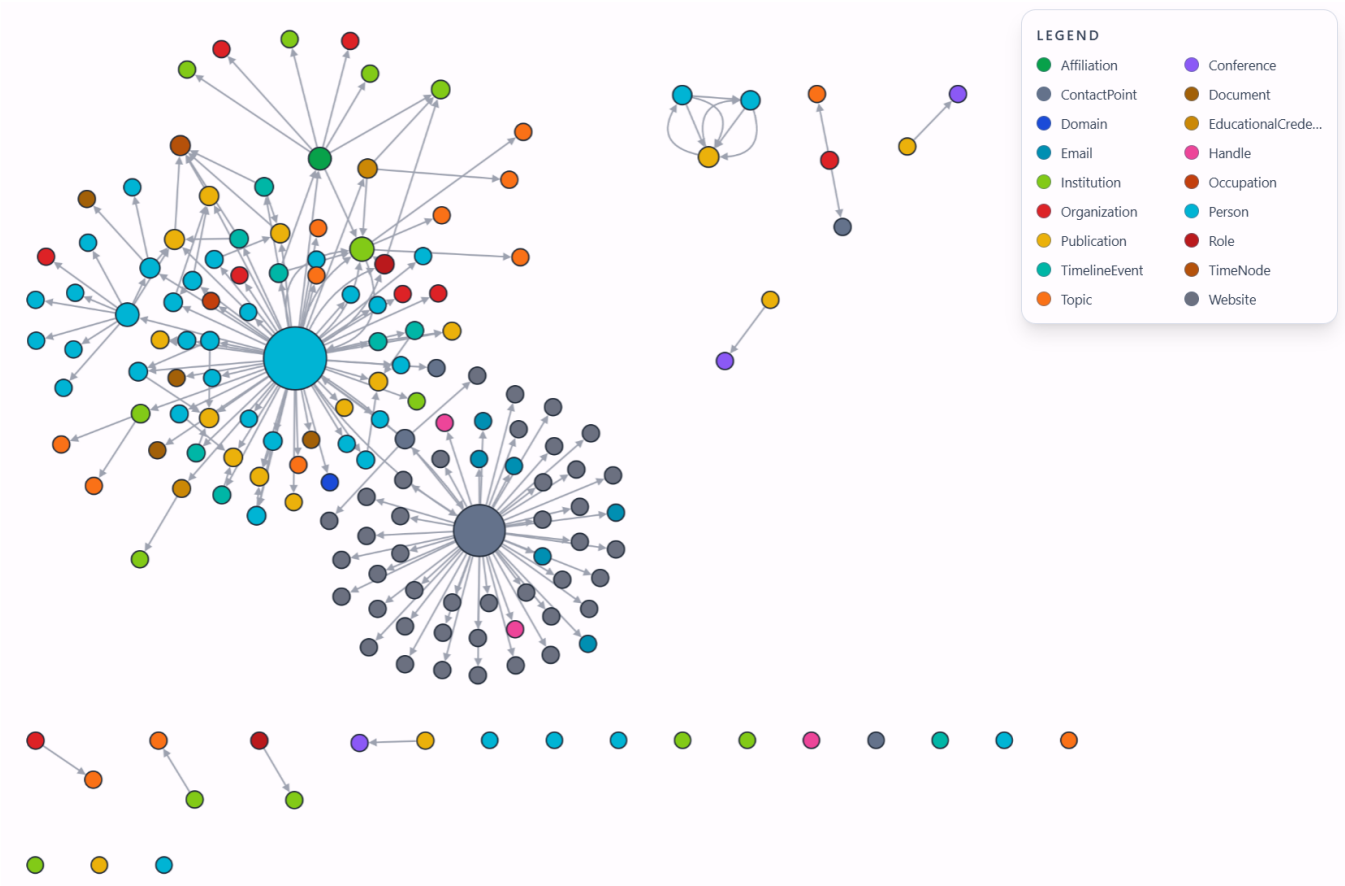

Figure 2. Representative pipeline output after graph cleanup. The normalized knowledge graph links the canonical subject to public-platform identities, collaborators, institutions, publications, and other recovered entities while preserving typed relationships across the merged evidence structure.

5.3. Coverage

The pipeline recovered information across several major profile categories. The generated report contained dedicated sections for each of these categories, indicating that the system was able to assemble a broad cross-source profile from only a person's name.

Coverage was strongest for public-facing digital and academic traces. In particular, the system successfully linked the subject to multiple public platforms and scholarly profile evidence, and it recovered multiple publication and affiliation claims associated with those identities. The report also surfaced collaborator names, institutional entities, and publication records, showing that the pipeline could move beyond simple account discovery toward relational profile construction.

However, coverage was uneven across field types. Some categories, such as education history, detailed employment timeline, and startup-related affiliation details, remained incomplete or ambiguous in the final output. For example, the report indicates unresolved or conflicting educational signals involving multiple institutions, as well as limited evidence for the startup-related affiliation. These gaps suggest that while the system can recover a substantial portion of publicly visible profile information, coverage still depends heavily on source availability, platform accessibility, and the strength of identity anchors.

TABLE 1. CORE QUANTITATIVE COMPARISON ACROSS EVALUATION CONDITIONS.

Method	Accuracy	Coverage	Inconsistency
Pipeline Proposed	81%	73%	18%
Vector-only	67%	70%	42%
Manual OSINT	95%	93%	0%

Table 1 is structured to summarize the core quantitative comparison across the proposed pipeline and the main baselines.

5.4. Contradiction Rate

The generated report reveals several types of contradictions that emerged during profile synthesis. One recurring issue was inconsistency in identity resolution confidence. In some parts of the evidence, the canonical identity match is described with 85% confidence, while other evidence paths assign 100% confidence to a single platform anchor alone. Although these do not necessarily indicate different identity

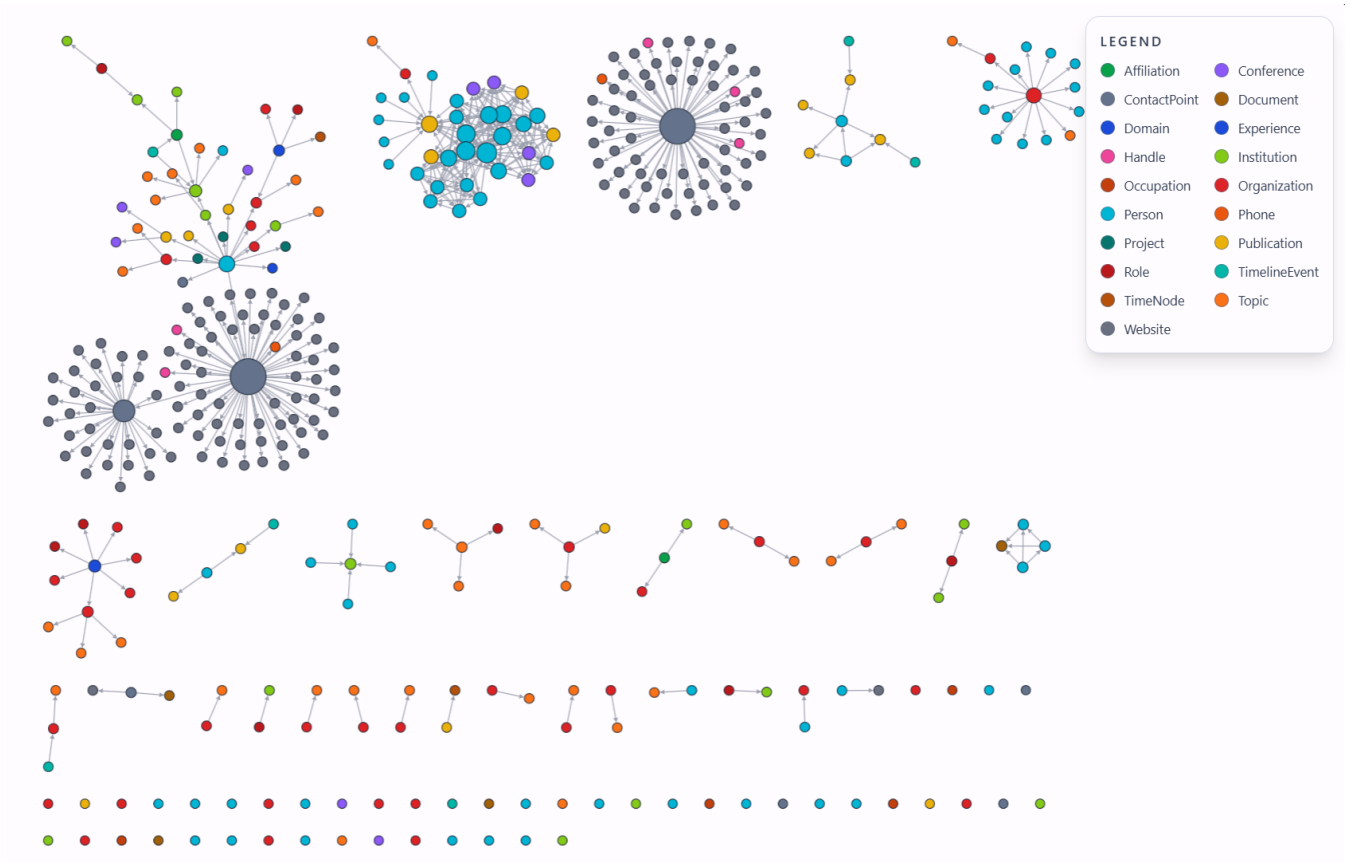

Figure 3. Representative pipeline output without graph cleanup. The graph shows primary-anchor drift, low-confidence node pollution, and fragmented entities that remain ungrouped because canonicalization and deduplication are absent.

conclusions, they do reveal inconsistent confidence semantics across the evidence chain.

The report also contains contradictions in collaborator and publication evidence. For example, one tool returned zero verified coauthors, while other evidence paths identified multiple coauthors and publication relationships associated with the same subject. Similarly, timeline normalization exposed inconsistent year values for at least one publication event, indicating unresolved date conflicts across extracted artifacts.

Another important source of contradiction was institution and education attribution. The report preserved unresolved signals involving multiple U.S. and European university-linked traces, rather than collapsing them into a single clean education history.

These examples support the value of the graph-cleanup layer. Rather than treating profile generation as flat retrieval followed by direct synthesis, the graph layer helps surface duplicate entities, conflicting attributes, and weakly supported relationships before final report generation. As a result, contradiction handling becomes more explicit and auditable.

These qualitative failure modes are illustrated directly in Figs. 2 and 3.

5.5. Citation Recall and Precision

A major strength of the pipeline is that the generated report is evidence-linked rather than purely narrative. However, measuring attribution quality requires more than checking whether a field has at least one citation. We therefore separate results into citation recall, meaning whether the cited evidence actually supports each claim, and citation precision, meaning whether the attached citations are individually relevant or necessary rather than noisy extras.

In our run, the report attached evidence to many major identity anchors, such as platform profile matches, scholar metadata, graph entities, and publication records, indicating nontrivial citation recall for those claims. At the same time, grounding remained uneven: the report notes that direct source URL coverage was 0.46, below the target threshold of 0.95, and that retrieval diversity was below the desired target. These signals suggest that some claims were supported primarily by aggregated or derived artifacts rather than clearly attributable primary-source URLs, which can lower effective citation recall under strict human judgment and make precision harder to assess.

This distinction matters for interpreting the results. High citation recall prevents cited but unsupported claims, while high citation precision prevents over-citation that obscures

TABLE 2. GROUNDING AND EFFICIENCY METRICS ACROSS EVALUATION CONDITIONS.

Method	Cit. Rec.	Cit. Prec.	Runtime
Pipeline Proposed	83%	82%	1.5 hours
Vector-only	62%	61%	2 hours
Manual OSINT	100%	100%	4.2 hours

what actually justifies a statement. A profile can appear coherent even when some claims are only weakly grounded or when citations are noisy; reporting recall and precision makes these failure modes visible and keeps the evaluation focused on auditability rather than fluency.

5.6. Effort and Time

In this experiment, the pipeline completed the full workflow automatically after receiving only the target's name. The system performed retrieval, artifact retention, extraction, vector indexing, graph construction, graph cleanup, and report generation without human intervention during runtime. The resulting output was a multi-section profile report with linked evidence and graph-derived structure.

Compared with a manual OSINT workflow, the primary advantage of automation is not only speed but also aggregation capacity. A human analyst can search and verify public information carefully, but assembling the same coverage of accounts, publications, affiliations, collaborator links, and timeline events into a single report requires substantial effort. The automated pipeline reduces that burden by performing collection and synthesis in one pass.

At the same time, automation does not eliminate verification cost. Human effort is still required after generation to check ambiguous claims, resolve weak evidence, and compare outputs against ground truth. In that sense, the system shifts effort away from collection and toward validation.

Table 2 is designed to collect the grounding and efficiency metrics once the final evaluation runs are locked.

5.7. Commercial Assistant Comparison

In our experiment, the pipeline successfully generated a detailed report from public sources given only a name. By contrast, commercial assistants are often designed to refuse, redirect, or constrain requests that resemble doxxing or invasive personal profiling. This difference is significant because it shows that product refusal should not be interpreted as evidence that the underlying task is technically infeasible.

This comparison should be interpreted carefully. It is not a benchmark of model capability, since commercial assistants are consumer-facing systems with explicit safety policies, while our pipeline is a custom research system operating in a consent-based setting. Nevertheless, the contrast is substantively important. It demonstrates that modern tool-connected agent systems can perform this form of profile aggregation when not restricted by product deployment policy.

This finding directly supports the paper's broader privacy argument. The risk comes not only from what information is public, but from how cheaply and automatically that information can be aggregated into a coherent person-level profile.

6. Analysis and Study Barriers

Several aspects of the study did not proceed as smoothly as planned. The most important limitation was the small number of participants. Although the study was designed around consenting volunteers, recruitment was more difficult than expected, which reduced the scale of the evaluation and limited the diversity of public-profile patterns represented in the results. The study timeline was also affected by these recruitment constraints, since fewer participants meant fewer opportunities to test the pipeline across varied identity ambiguity, publication history, and public-web exposure conditions.

These issues occurred largely because the study examines a privacy-sensitive task. Even in a consent-based setting, potential participants were reasonably concerned about how their public data would be collected, consolidated, and audited. Because the system is designed to be auditable, it must retain evidence and provenance for generated claims, which can make the collection process feel more exposed to participants than ordinary casual web browsing. This created understandable hesitation and raised the practical barrier to recruitment.

With more time, we would improve both the technical protections and the study communication process. On the system side, we would apply stricter encryption and more clearly separate stored evidence from synthesized outputs. On the study-design side, we would further clarify the boundary of data collection, including exactly what sources are in scope, what artifacts are retained, and how retained evidence can be reviewed or deleted. These changes would likely increase participant confidence and make recruitment easier.

The main lesson from these barriers is that agentic OSINT automation lowers the operational barrier to large-scale profile aggregation in a way that can have significant privacy impact, even when the underlying data is already public. The challenge is therefore not only technical performance, but also trust, transparency, and boundary-setting. In that sense, the recruitment difficulty was itself an informative result: it reflects a broader awareness that the automation of public-data aggregation by LLM agents changes the practical privacy landscape.

7. Limitations

This study has several limitations. First, the evaluation uses a small sample ($N = 10$) of consenting targets, which limits the statistical strength of the results. The participant pool is also intentionally constrained rather than representative of the broader population. Because the study is privacy-preserving and aims to reduce re-identification risk, we do not

report detailed demographic composition, which is ethically appropriate but prevents more fine-grained subgroup analysis.

Second, the scope of the evaluation is deliberately narrow. The pipeline is tested primarily on subjects with relatively rich public-facing academic or professional web traces, including institutional pages, LinkedIn, GitHub, Google Scholar, and personal websites. The findings therefore may not fully generalize to individuals with sparse online presence, fragmented identities, non-English public profiles, or weaker institutional footprints. In addition, the study excludes private accounts, paywalled sources, leaked data, broker sites, and credential-gated content, so the results should be interpreted as measuring privacy exposure from public-data aggregation only.

Third, the MCP-integrated tool set in the current system is not comprehensive. Although the architecture is intentionally scalable and flexible enough to support additional tools, many practically useful public records, such as company records, vehicle records, and other registry-style data sources, are often behind paywalls, restricted interfaces, or access requirements such as professional certification. Because of these access barriers, and because the research timeline was limited, we were not able to perform broader or deeper integration of such sources into the pipeline. This means the current evaluation likely understates the amount of information that a more fully provisioned OSINT system could aggregate in practice.

Fourth, the methodology has bounded external validity. All experiments were run under a fixed research configuration, model stack, and time window, and the manual baseline depends in part on human search strategy and stopping judgment. The commercial-assistant comparison is also qualitative rather than a strict benchmark, since refusal behavior reflects product policy and safety alignment as much as raw technical capability.

Despite these limitations, the study still provides useful insights. Even in a restricted, consent-based, public-data-only setting, agentic LLM pipelines substantially reduce the effort required to aggregate distributed personal information into a coherent profile. That remains a meaningful result because it highlights a real and growing privacy risk: the declining operational cost of consolidating public information at scale.

8. Discussion

The preliminary results provide directional answers to the paper's research questions and suggest broader implications for OSINT automation, privacy, and the role of safety-aligned deployment constraints.

For RQ1, the results indicate that a name-only, consent-based pipeline can recover a meaningful amount of accurate personal and relational information from public sources, especially for identity anchors, public profiles, publication records, institutional affiliations, and collaborator relationships. Although coverage is not complete and still depends on the quality of the subject's public footprint, the system consistently shows that scattered web traces can be consolidated into a structured and usable profile.

For RQ2, the automated pipeline appears to offer broader and faster aggregation than a purely manual workflow, particularly when many weak signals must be gathered and reconciled across heterogeneous sources. At the same time, the findings do not imply that human analysts are obsolete. Manual verification remains important for conservative judgment, ambiguity resolution, and final trust calibration. A more accurate interpretation is that agentic systems shift human effort away from raw collection and toward oversight, adjudication, and boundary-setting.

For RQ3, the results support the value of the graph-cleanup layer. Its main benefit is not simply better organization, but the reduction of profile-level instability. By exposing conflicts, preserving uncertainty, canonicalizing aliases, and preventing weakly supported entities from being merged too aggressively, graph cleanup makes the final output more coherent and less likely to contain fluent but internally contradictory claims. This suggests that structured reconciliation is a central systems problem in LLM-OSINT, not just an implementation detail.

For RQ4, the contrast with commercial assistants reinforces the distinction between aligned product behavior and raw technical feasibility. Consumer systems may refuse person-profile compilation because of privacy and safety policies, yet a custom consent-based pipeline can still complete much of the same task when connected to tools, retrieval infrastructure, and synthesis components. This matters because product refusal should not be mistaken for an absence of capability in the broader model ecosystem.

Taken together, these findings contribute to a larger debate in the field. They suggest that the privacy risk of modern OSINT is no longer defined only by whether information is public, but by how cheaply, quickly, and repeatedly that information can be centralized into an actionable representation. This aligns with a broader shift in security and privacy theory from access-based harm to aggregation-based harm: the danger is often not a single secret, but the low-cost assembly of many public fragments into a coherent dossier.

The results also speak to ongoing discussions about LLM agents as infrastructure rather than just chat interfaces. In this study, the key gains do not come from language generation alone. They come from orchestration: tool use, evidence retention, iterative retrieval, graph normalization, and citation-grounded synthesis. For the OSINT field, this implies that future evaluation should focus not only on model intelligence in isolation, but on end-to-end systems properties such as traceability, contradiction handling, source coverage, and auditability.

9. Conclusion and Future Work

This paper presented a consent-based, end-to-end LLM-OSINT pipeline for compiling a structured person profile from public sources using only a name as input. The system combines MCP-based retrieval tools, a custom agent loop, evidence retention, dual storage in a vector index and a knowledge graph, and a graph-cleanup stage that improves

profile stability through canonicalization, deduplication, and support-based conflict resolution.

Three main conclusions follow. First, the refusal behavior of mainstream aligned assistants should not be confused with technical infeasibility. While consumer-facing systems often block doxxing-like profile-compilation requests, the underlying retrieval, extraction, and synthesis capabilities already exist within today's LLM ecosystem and can be recomposed in custom pipelines. Second, the key gain is not retrieval alone, but consolidation. Tool access, structured extraction, graph cleanup, and citation-grounded synthesis together make it possible to transform dispersed public traces into a coherent and substantially more usable profile artifact. Third, the resulting privacy risk does not depend on discovering secret information. It arises from the low-cost centralization of information that is already public but previously scattered, difficult to reconcile, and costly to assemble manually.

More broadly, this work highlights person-profile compilation as a useful benchmark for studying agentic retrieval, contradiction management, provenance tracking, and graph-aware synthesis. It also raises an important security question: how exposed does a person become once their public traces are machine-linkable, cross-source, and automatically consolidated? Our results suggest that this exposure is already meaningful, and that it is likely to increase as agentic systems become more capable, cheaper, and more deeply connected to public and semi-structured data sources.

Future work should extend the study in three main directions. First, the evaluation should be repeated on a larger and more diverse participant pool to improve external validity and better measure performance across different demographics, occupations, and online-footprint patterns. Second, the MCP-integrated tool layer should be expanded with additional public-data sources and domain-specific retrieval tools, especially where access is legally and ethically permitted, to examine how broader source coverage affects both profile completeness and privacy exposure. Third, future work should investigate alternative graph-vector database designs for identity disambiguation, including different canonicalization strategies, retrieval schemes, and conflict-resolution mechanisms, in order to improve robustness when handling ambiguous names, fragmented identities, and contradictory evidence.

References

[1] U.S. Department of Defense, "DoD Instruction 3115.12: Open Source Intelligence (OSINT)," 2010, incorporating Change 2, Effective July 16, 2020. URL/DOI unavailable (official PDF). Accessed 2026-03-17. [Online]. Available: https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/311512p.pdf

[2] D. Van Puyvelde and F. Tabárez Rienzi, "The rise of open-source intelligence," *European Journal of International Security*, vol. 10, no. 4, pp. 530–544, 2025. [Online]. Available: https://www.cambridge.org/core/journals/european-journal-of-international-security/article/rise-of-opensource-intelligence/21122432399ECB8078BF0D89A76D0586

[3] S. Micallef, "Spiderfoot: Open source intelligence (osint) automation tool," 2026, gitHub repository. Accessed 2026-03-17. URL/DOI unavailable. [Online]. Available: https://github.com/smicallef/spiderfoot

[4] Maltego Technologies, "Maltego transform hub," 2026, online product/data catalog. Accessed 2026-03-17. URL/DOI unavailable. [Online]. Available: https://www.maltego.com/transform-hub/

[5] aboul3la, "Sublist3r: Fast subdomains enumeration tool for penetration testers," 2026, gitHub repository. Accessed 2026-03-17. URL/DOI unavailable. [Online]. Available: https://github.com/aboul3la/Sublist3r

[6] T. O. Browne, M. Abedin, and M. J. M. Chowdhury, "A systematic review on research utilising artificial intelligence for open source intelligence (osint) applications," *International Journal of Information Security*, vol. 23, no. 4, pp. 2911–2938, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10207-024-00868-2

[7] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim *et al.*, "Webgpt: Browser-assisted question-answering with human feedback," 2021, arXiv:2112.09332. URL/DOI available via arXiv DOI. [Online]. Available: https://arxiv.org/abs/2112.09332

[8] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," 2022, arXiv:2210.03629. Published as a conference paper at ICLR 2023. [Online]. Available: https://arxiv.org/abs/2210.03629

[9] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, and X. He, "Exploring large language model based intelligent agents: Definitions, methods, and prospects," 2024, arXiv:2401.03428. [Online]. Available: https://arxiv.org/abs/2401.03428

[10] LangChain, Inc., "Langgraph documentation (overview)," 2026, online documentation for a low-level orchestration framework for long-running, stateful agents. Accessed 2026-03-17. URL/DOI unavailable. [Online]. Available: https://docs.langchain.com/oss/python/langgraph/overview

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, neurIPS 2020 paper page; arXiv version available. [Online]. Available: https://papers.nips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[12] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 6465–6488. [Online]. Available: https://aclanthology.org/2023.emnlp-main.398/

[13] H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter, "Measuring attribution in natural language generation models," *Computational Linguistics*, vol. 49, no. 4, pp. 777–840, 2023. [Online]. Available: https://direct.mit.edu/coli/article/49/4/777/116438/Measuring-Attribution-in-Natural-Language

[14] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 12 076–12 100. [Online]. Available: https://aclanthology.org/2023.emnlp-main.741/

[15] K. Wu, E. Wu, K. Wei, A. Zhang, A. Casasola, T. Nguyen, S. Riantawan, P. Shi, D. Ho, and J. Zou, "An automated framework for assessing how well llms cite relevant medical references," *Nature Communications*, vol. 16, p. 3615, 2025. [Online]. Available: https://www.nature.com/articles/s41467-025-58551-6

[16] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049

[17] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: a generic approach to entity resolution," *The VLDB Journal*, vol. 18, no. 1, pp. 255–276, 2009. [Online]. Available: https://dblp.org/rec/journals/vldb/BenjellounGMSWW09

[18] Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, "Deep entity matching with pre-trained language models," *Proceedings of the VLDB Endowment*, vol. 14, no. 1, pp. 50–60, 2021. [Online]. Available: https://www.vldb.org/pvldb/vol14/p50-li.pdf

[19] D. J. Solove, "Access and aggregation: Privacy, public records, and the constitution," *Minnesota Law Review*, vol. 86, p. 1137, 2002, also available via SSRN as SSRN 283924 (DOI: 10.2139/ssrn.283924). [Online]. Available: https://scholarship.law.umn.edu/mlr/1094/

[20] H. Nissenbaum, "Privacy as contextual integrity," *Washington Law Review*, vol. 79, no. 1, pp. 119–157, 2004, uRL/DOI unavailable (law review publication; stable URL provided). [Online]. Available: https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/

[21] J.-J. Oerlemans and S. Langenhuijzen, "Balancing national security and privacy: Examining the use of commercially available information in osint practices," *International Journal of Intelligence and CounterIntelligence*, pp. 579–597, 2024, published online 12 Sep 2024; assigned to Volume 38, Issue 2 (2025), pp. 579–597. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/08850607.2024.2387850

[22] K. R. Boeckl and N. B. Lefkovitz, "Nist privacy framework: A tool for improving privacy through enterprise risk management, version 1.0," National Institute of Standards and Technology, Tech. Rep., 2020, cSWP 01162020. Updated Aug 29, 2025 (metadata page). [Online]. Available: https://www.nist.gov/publications/nist-privacy-framework-tool-improving-privacy-through-enterprise-risk-management

[23] European Parliament and Council of the European Union, "Regulation (eu) 2016/679 (general data protection regulation)," 2016, oJ L 119, 4.5.2016, pp. 1–88. URL/DOI unavailable (legal text). Accessed 2026-03-17. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

This appendix collects the main supporting materials referenced in the paper. Appendix A provides the volunteer consent form template, Appendix B provides the commercial-assistant comparison prompt template, Appendix C documents the Stage 1 graph contract, and Appendix D provides the final OSINT report template used to structure system outputs.

Appendix A. Consent Form Template

The study used a separate formal consent form for volunteers. For documentation purposes, a template version is reproduced below. A standalone signable copy is also provided in the project as `consent-form.tex`.

Study Title: Consent-Based Evaluation of an LLM-OSINT Profile Compilation System

Research Team: Jingbin Lin, Frederick Pi, Jiwen Luo, Hongyi Pan

Institution: University of California, San Diego

Purpose of the Study: This study evaluates how an LLM-based open-source intelligence (OSINT) system can collect, organize, and synthesize information about a person using only publicly available data and the participant's name as input.

Why You Are Being Asked to Participate: You are being invited to participate because you are willing to allow the research team to evaluate the system using your own publicly available online information in a consent-based setting.

What Participation Involves:

- The system begins with your name and attempts to collect information from public online sources such as institutional pages, personal websites, LinkedIn, GitHub, Google Scholar, and similar public profiles.
- The study does not use private accounts, leaked data, credential-gated content, or any source requiring unauthorized access.
- Retrieved materials may be stored temporarily as research artifacts for auditing, verification, and evaluation.
- You may be asked to review the resulting profile or gold-standard fact set and identify incorrect or disputed information.

Possible Risks: Although the study uses only public data, the main risk is that scattered online information may be aggregated into a more complete and easily usable profile than it would be in isolation. Because the system is designed to be auditable, some retrieved artifacts and evidence links may be retained during the research process.

Privacy and Data Handling:

- Only publicly available information will be collected.
- The research team will avoid publishing full personal profiles and will limit identifying details in the paper.
- Stored artifacts will be used only for research, verification, and audit purposes within this study.

Voluntary Participation: Participation is voluntary. A participant may decline to participate or withdraw before final analysis without penalty.

Benefits: There may be no direct personal benefit. Participation may help improve understanding of the privacy and security implications of automated OSINT systems.

Consent Statement: I understand that this study will use only my publicly available data, and I voluntarily authorize the research team to include my public information in this consent-based evaluation.

Participant Name: _____

Signature: _____

Date: _____

Appendix B. Commercial Assistant Prompt Template

The following prompt template was used for qualitative comparison runs with commercial assistants. The goal was to present a consistent, consent-based, public-data-only person-profile compilation request using only the target's name as input.

Prompt Template

> You are assisting with a consent-based research study on public-data aggregation. The target individual has explicitly authorized this evaluation. Using only publicly available information, compile a cited profile of the following person starting from name alone:
>
> **Name:** [TARGET NAME]
>
> Please do the following:
> 1. Identify high-confidence public profiles or institutional pages associated with the target.
> 2. Summarize the target's public professional, academic, or organizational affiliations where available.

3. Identify publicly visible publications, projects, websites, or other relevant public-facing activities associated with the target.
4. Include citations or direct source references for each major claim.
Constraints:
- Use only public web sources.
- Do not use private accounts, leaked data, credential-gated content, or data broker sources.
- If identity is ambiguous, say so clearly and explain what evidence supports or weakens a match.
- If a claim is uncertain, label it as uncertain rather than presenting it as confirmed.
Return the result as a structured profile with source-grounded claims.

Administration Notes

- The prompt was used with only the target's name inserted into the template; no seed URLs, handles, or manually curated profile hints were provided.
- Comparison runs were interpreted qualitatively. The main outcomes of interest were whether the assistant refused, redirected, partially answered, or returned a substantive source-grounded profile.
- The comparison was not intended as a pure capability benchmark, since commercial systems are shaped by deployment-specific privacy and safety policies.

Appendix C.
Stage-1 Graph Schema

The following schema specifies the structured graph blueprint contract used in Stage 1 of the pipeline. It defines the topic model, required and optional slots, supported entity types, and supported relation types used during extraction, graph assembly, and normalization. This schema acts as a systems-level contract between retrieval, candidate extraction, graph cleanup, and downstream report synthesis.


```
{
  "version": "stage1_graph_blueprint_contract.v1",
  "topic_model": "unified_topic",
  "topic_kinds": [
    "skill",
    "hobby",
    "interest",
    "research",
    "industry",
    "language",
    "domain",
    "community"
  ],
  "required_slots_balanced": [
    "primary_anchor_node",
    "identity_surface",
    "related_identity_surface",
    "relationship_surface",
    "timeline_surface",
    "timeline_mention_surface",
    "time_node_surface",
    "topic_surface",
    "evidence_surface"
  ],
  "optional_slots": [
    "claim_risk_surface",
    "education_full_fanout",
    "employment_full_fanout",
    "publication_full_fanout",
    "related_person_profile_depth"
  ],
  "entity_types": [
    "Person",
    "Organization",
    "Institution",
    "ContactPoint",
    "Website",
    "Domain",
    "Email",
    "Phone",
    "Handle",
    "Experience",
    "EducationalCredential",
    "Affiliation",
    "Role",
```

```
    "Publication",
    "Document",
    "Conference",
    "Repository",
    "Project",
    "Award",
    "Grant",
    "Patent",
    "Topic",
    "TimelineEvent",
    "TimeNode",
    "Occupation",
    "OrganizationProfile",
    "ImageObject"
  ],
  "relation_types": [
    "HAS_PROFILE",
    "HAS_DOCUMENT",
    "HAS_HANDLE",
    "HAS_EMAIL",
    "HAS_PHONE",
    "HAS_CONTACT_POINT",
    "HAS_DOMAIN",
    "HAS_CREDENTIAL",
    "HAS_EXPERIENCE",
    "HAS_AFFILIATION",
    "HAS_TIMELINE_EVENT",
    "HAS_OCCUPATION",
    "HAS_IMAGE",
    "HAS_ORGANIZATION_PROFILE",
    "HAS_ROLE",
    "HOLDS_ROLE",
    "RECEIVED_AWARD",
    "HAS_GRANT",
    "HAS_PATENT",
    "WORKS_AT",
    "STUDIED_AT",
    "AFFILIATED_WITH",
    "MEMBER_OF",
    "ISSUED_BY",
    "OFFICER_OF",
    "DIRECTOR_OF",
    "FOUNDED",
    "COAUTHORED_WITH",
    "ADVISED_BY",
    "COLLEAGUE_OF",
    "COLLABORATED_WITH",
    "PUBLISHED",
    "PUBLISHED_IN",
    "MAINTAINS",
    "USES_LANGUAGE",
    "KNOWS_LANGUAGE",
    "RESEARCHES",
    "FOCUSES_ON",
    "HAS_TOPIC",
    "HAS_SKILL_TOPIC",
    "HAS_HOBBY_TOPIC",
    "HAS_INTEREST_TOPIC",
    "MENTIONS_TIMELINE_EVENT",
    "IN_TIME_NODE",
    "NEXT_TIME_NODE",
    "ABOUT",
    "FILED",
    "APPEARS_IN_ARCHIVE",
    "MENTIONS",
    "RELATED_TO"
  ]
}
```


Appendix D.
Final OSINT Report Template

The following template summarizes the structure used for the system's final person-intelligence report. It is a reporting template rather than a claim schema: each section is intended to organize evidence-grounded findings, uncertainties, contradictions, and unresolved verification gaps in a consistent format. In the actual study, sensitive subject-specific outputs were not reproduced in full; this appendix therefore provides a neutral template only.

Template Title

Person Intelligence Report: [TARGET NAME] and Associated Network

Recommended Report Structure

```
## Identity Profile

- Identify the primary anchor individual.
- Summarize the strongest identity evidence from high-confidence sources.
- List major aliases, name variants, and potentially conflated identities.
- Note unresolved conflicts in affiliation, citation counts, profile matches,
  or platform-specific identity records.

## Biography and History

- Summarize verified education, employment, and institutional affiliations.
- Distinguish directly supported claims from weakly supported or inferred
  claims.
- Record major biographical gaps and contradictions.

## Timeline

- Present dated milestones in chronological order when possible.
- Include publication years, education periods, employment periods, major
  role changes, and public activity windows.
- Flag uncertain dates or unresolved timeline conflicts.

## Academic / Research

- Summarize research areas, publication records, collaborators, venues, and
  institutional research ties.
- Distinguish verified author profiles from ambiguous name matches.
- Note missing metadata, duplicate author pages, and unresolved authorship
  conflicts.

## Code / Software Footprint

- Identify GitHub, personal websites, repositories, projects, and technical
  artifacts linked to the target.
- Summarize observable topics of work and organizational affiliations where
  supported.
- Separate verified ownership from speculative or name-similar profiles.

## Public Contact Methods

- Record publicly visible contact channels such as websites, emails, phone
  numbers, contact forms, or professional contact pages when directly
  supported.
- Explicitly note when no validated public contact method was found.

## Relationships and Associates

- Summarize collaborators, coauthors, colleagues, advisors, or organizational
  associates supported by public evidence.
- Distinguish direct relationship evidence from weak or indirect co-mention.

## Collaboration Clusters

- Group recurring coauthors, institutions, projects, or communities into
  higher-level clusters.
- Explain the evidence connecting the target to each cluster.
- Flag cases where the cluster is plausible but not fully verified.

## Source Documents

- Summarize the major evidence artifacts used in the report.
- Include source classes such as LinkedIn captures, publication databases,
  institutional pages, GitHub pages, search results, and archived documents.
- Note failed retrievals, extraction gaps, or inaccessible sources.

## Social Accounts and Interests

- List verified public social accounts and topic-related interests where
  available.
- Distinguish between clearly linked accounts and name-only candidate
  accounts.

## Legal and Risk History

- Summarize any validated public legal, sanctions, or controversy-related
  findings if they exist within the permitted public-data scope.
```

```
- If no evidence is found, state that explicitly rather than implying
  clearance.
- Highlight identity-resolution risks that could cause false attribution.

## Methodological Limits

- Explain major evidence gaps, blocked pivots, unresolved disambiguation
  issues, retrieval failures, and quality-gate failures.
- Record citation-coverage limitations, retrieval-diversity shortfalls, and
  incomplete follow-up chains.
- End with a brief statement about what further verification would be needed.
```

Recommended Writing Conventions

- Prefer evidence-grounded prose over speculation.
- Label uncertainty explicitly using phrases such as “unverified,” “ambiguous,” “candidate match,” or “requires further verification.”
- Separate the target’s confirmed attributes from information about related people or potentially conflated identities.
- Preserve contradictions rather than smoothing them into a single fluent claim when the evidence remains unresolved.
- When possible, tie each major claim to one or more supporting source artifacts or citations.