OncoCITE: AI-Driven Genomic Evidence Curation with Multi-Agent Natural Language Processing

Supplementary Information

Mujahid Quidwai[1], Santiago Thibaud[2], Dennis Shasha[3], Sundar Jagannath[2], Samir Parekh[2,4] and Alessandro Laganà[1,4,5,*]

[1]Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY
[2]Division of Hematology and Medical Oncology, Icahn School of Medicine at Mount Sinai, New York, NY
[3]Courant Institute of Mathematical Sciences, New York University, New York, NY
[4]Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY
[5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY

[*]Corresponding author: alessandro.lagana@mssm.edu

Contents

Supplementary Figures

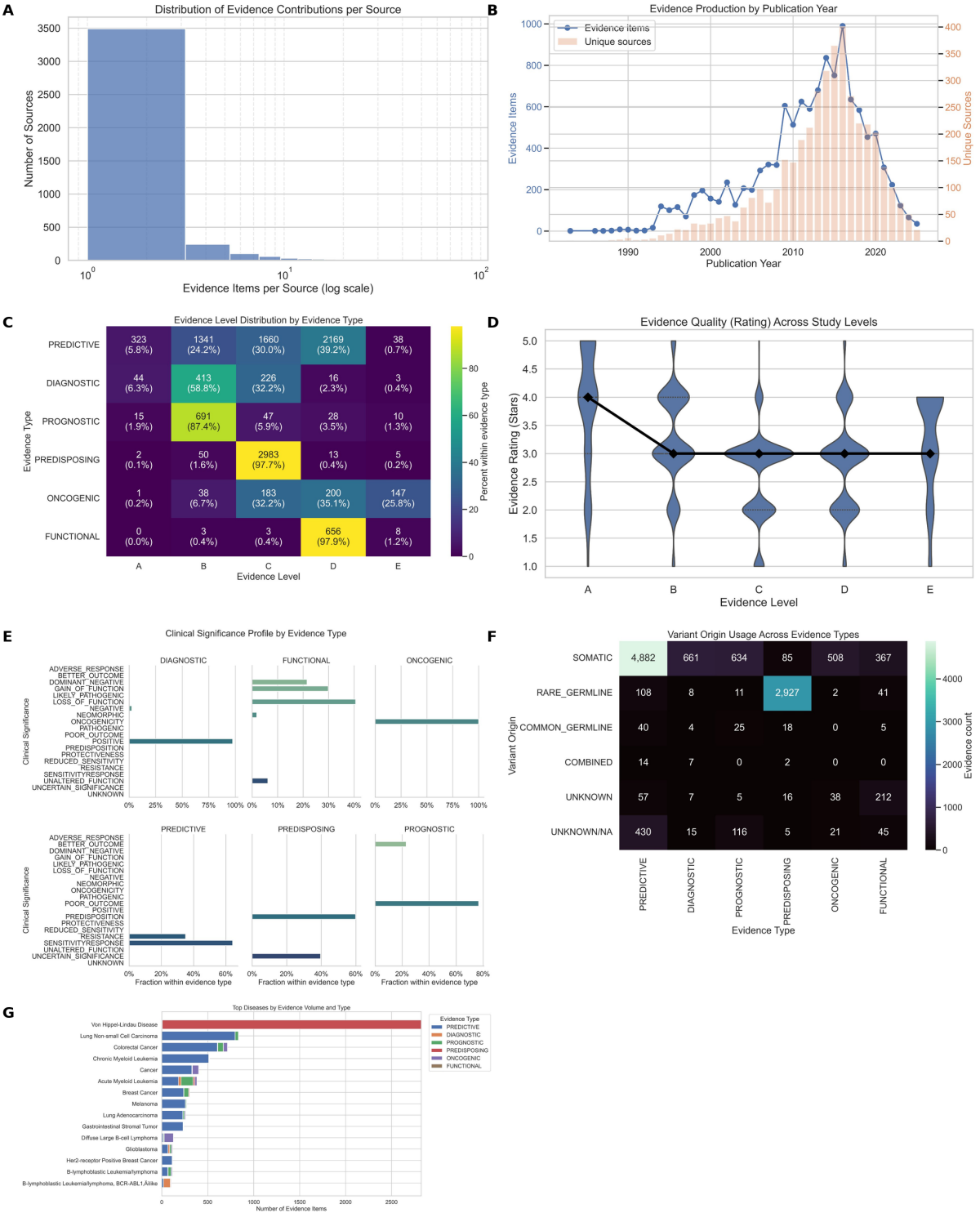

Figure 1: **Supplementary Figure S1. Extended CIViC Database Analysis.** Comprehensive analysis of 11,312 evidence items from 3,083 source publications [1, 2]. **(A)** Long-tail source distribution showing a median contribution of 1 evidence item per source (log scale); most sources contribute only a single item, demonstrating low curation yield per paper. **(B)** Temporal analysis of evidence production by publication year (1985–2025), with a clear surge during 2010–2020 followed by decline; evidence production has outpaced curation capacity. **(C)** Evidence level distribution heatmap by evidence type, showing concentration of Predictive evidence at Level D (39.2% preclinical) with only 5.8% at Level A (validated practice standard), and Predisposing evidence concentrated at Level C (97.7%). **(D)** Trust rating violin plots across evidence levels, demonstrating wide variance at all levels with median ratings around 3 stars; no clear quality stratification emerges from trust ratings alone. **(E)** Clinical significance profiles by evidence type, revealing underrepresentation of resistance in Predictive evidence (35% vs 64% sensitivity)

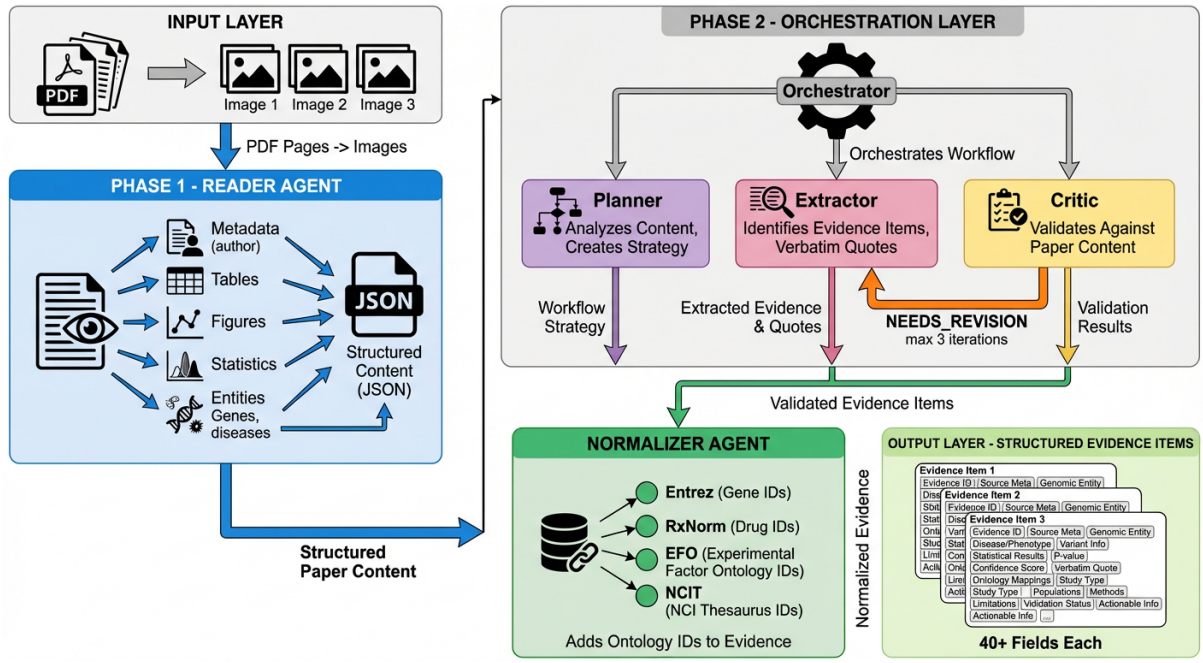

Figure 2: **Supplementary Figure S2. Multi-Agent System Architecture.** The OncoCITE "Reader-First" architecture decomposes evidence extraction into three sequential phases. Phase 1 (Reader): Converts all PDF pages to images, performs comprehensive visual extraction of metadata, text, tables, figures, and statistics, and outputs structured JSON as a single source of truth. Phase 2 (Orchestrator + Subagents): Planner analyzes content and generates an extraction strategy; Extractor identifies evidence items with mandatory verbatim quotes and ontology normalization; Critic validates against source text by checking field completeness, type-specific rules, statistical accuracy, and entity grounding; iterative refinement loop (maximum of three iterations) corrects errors before approval. Phase 3 (Normalizer): Enriches approved items with standardized identifiers from external ontologies (Entrez, RxNorm, EFO, NCIt, DOID). All agent interactions are mediated through a Model Context Protocol (MCP) server with 22 specialized tools, ensuring strict separation of reasoning and execution with complete audit trails.

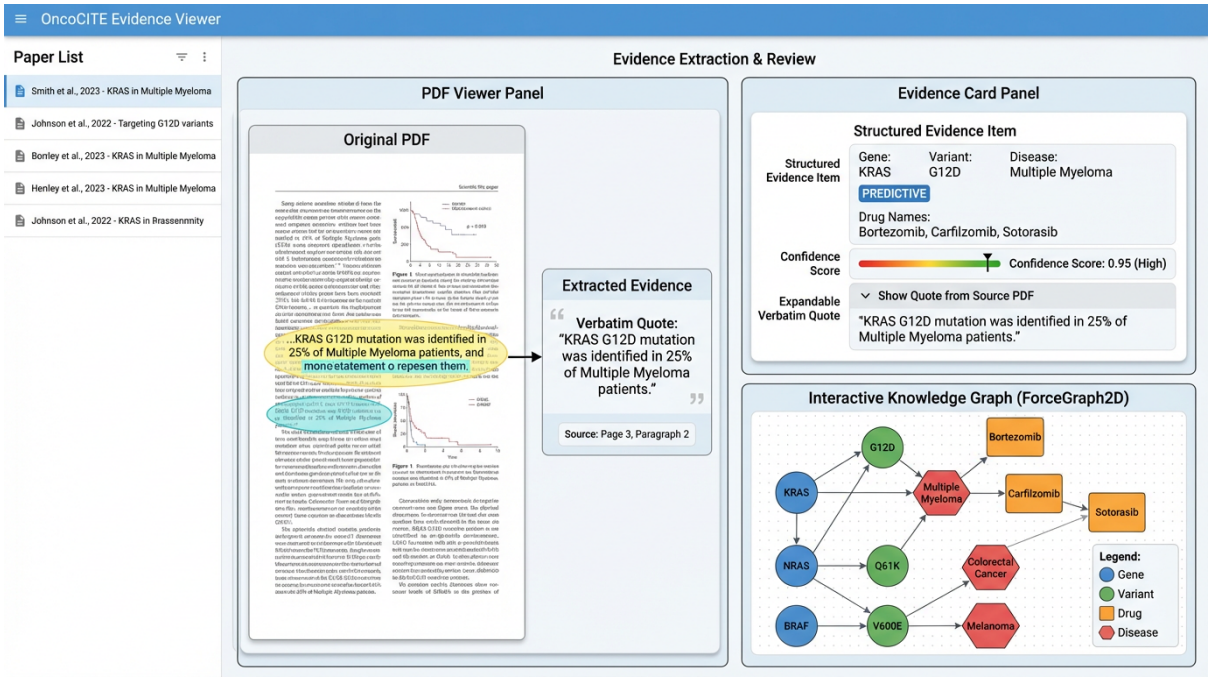

Figure 3: **Supplementary Figure S3. Web Interface for Evidence Verification.** React-based single-page application enabling real-time verification and knowledge graph exploration. Left panel: PDF viewer (PDF.js) with text layer for searchability and highlighting of verbatim quote locations. Right panel: Evidence cards displaying gene/variant (with Entrez IDs), disease (with EFO ID), evidence type badge, confidence score meter, therapy names (with RxNorm/NCIt IDs), and expandable verbatim quote section. "View in PDF" buttons scroll and highlight source locations, enabling rapid verification without full-text review. Bottom panel: Interactive knowledge graph (ForceGraph2D) visualizing relationships between genes (blue circles), variants (green diamonds), drugs (orange squares), and diseases (red hexagons) with edge thickness encoding evidence weight (confidence $\times$ evidence level). Users can filter by evidence type, confidence threshold, or specific entities.

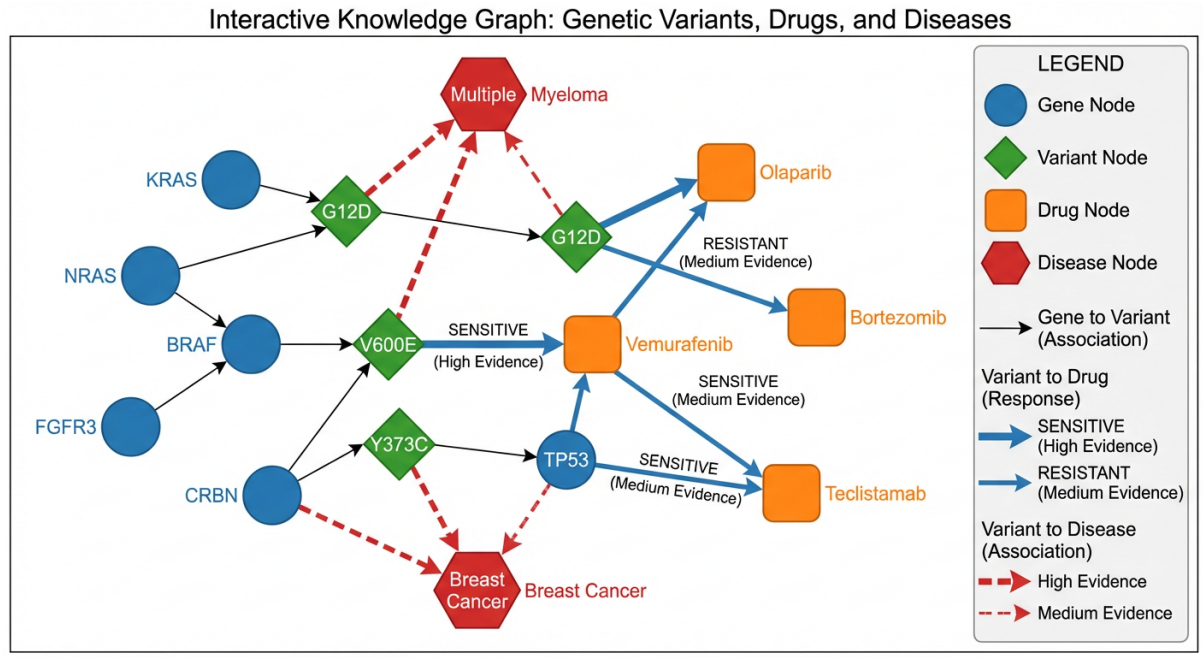

Figure 4: **Supplementary Figure S4. Interactive Knowledge Graph Visualization.** Force-directed graph representation of extracted evidence relationships across the validation corpus. Nodes represent biological entities: genes (blue circles), variants (green diamonds), diseases (red hexagons), and therapies (orange squares). Edges encode clinical relationships with directionality (SENSITIVE, RESISTANT, POOR_OUTCOME) and thickness proportional to evidence weight (confidence score × evidence level). The graph enables exploration of therapeutic networks, identification of resistance mechanisms, and discovery of cross-variant patterns. Users can filter by evidence type, confidence threshold, or specific entities, with real-time updates as new extractions complete. This visualization demonstrates the system's ability to synthesize complex gene-variant-drug-disease relationships into navigable clinical knowledge structures.

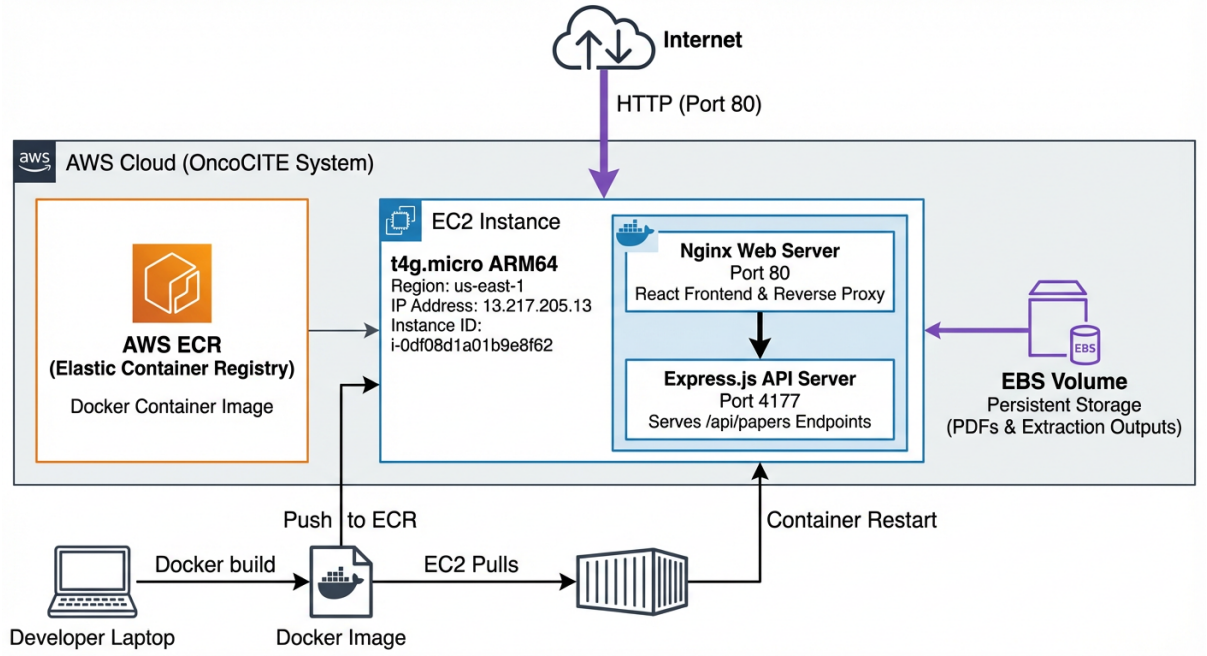

Figure 5: **Supplementary Figure S5. Cloud Deployment Architecture.** AWS infrastructure diagram showing the production deployment configuration for OncoCITE. The system runs on an EC2 t4g.micro instance (ARM64 Graviton processor, 2 vCPU, 1 GB RAM) providing cost-effective hosting suitable for academic research budgets. The Docker container packages the complete stack: Python 3.11 backend with Claude Agent SDK, Node.js 18 Express API server, Nginx reverse proxy, and pre-built React frontend. An EBS volume (20 GB) provides persistent storage for PDFs, extraction outputs, and checkpoint files. The architecture supports horizontal scaling through Application Load Balancer for high-availability deployments, with optional S3 integration for large-scale PDF corpus storage and CloudFront CDN for global distribution.

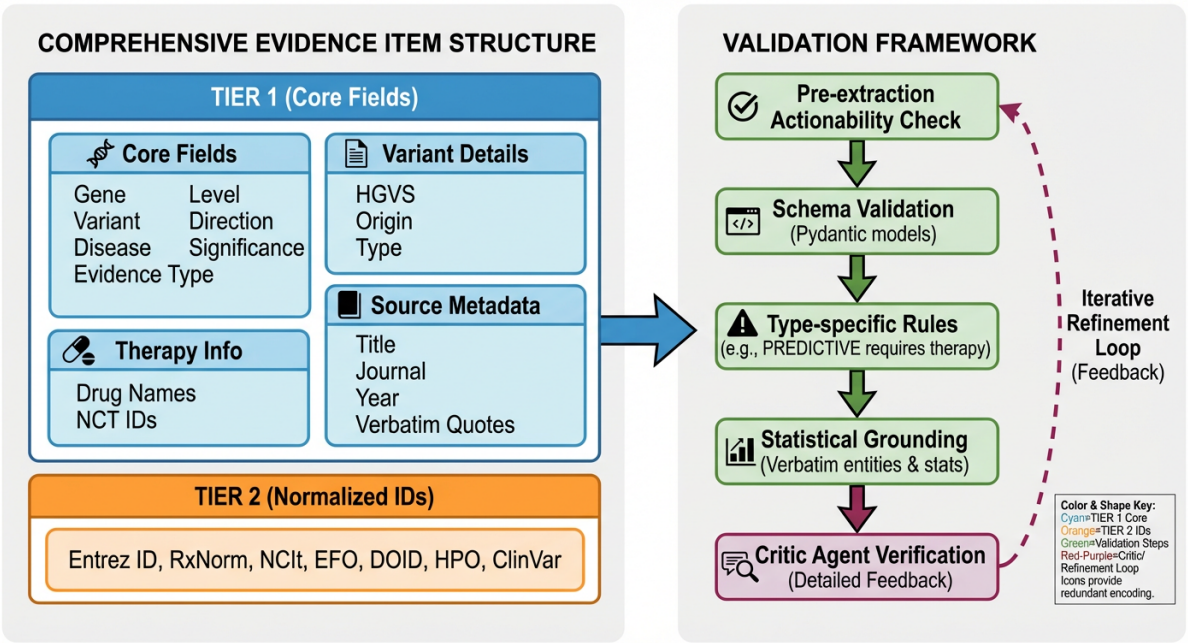

Figure 6: **Supplementary Figure S6. Three-Way Validation Framework.** Schematic representation of the validation methodology employed to evaluate both AI extraction accuracy and ground truth quality. Unlike traditional approaches that treat curated databases as infallible ground truth, this framework uses the original scientific publication as the sole source of truth. For each paper, the system performs parallel comparison: (1) CIViC ground truth items are validated against source text, (2) OncoCITE extractions are validated against source text, and (3) discrepancies between systems are flagged for domain expert adjudication. This bidirectional approach enables detection of errors in both AI extraction (hallucinations, missed evidence) and human curation (over-specification, therapy misattribution, evidence type misclassification), providing a more rigorous evaluation than single-reference validation.

Supplementary Note 0: Detailed Database Comparison

S0.1 Comprehensive Analysis of Oncology Knowledge Base Limitations

Table S1 presents a detailed comparative analysis of the five major oncology variant knowledge bases referenced in the main text (CIViC [1, 2], OncoKB [3], COSMIC [4], ClinVar [5], and JAX-CKB [6]). This analysis highlights systematic barriers across access, functionality, and integration that collectively limit the clinical utility of precision oncology genomic data (Supplementary Figure S1; Figure 1).

Table 1: **Table S1.** Detailed comparative analysis of oncology variant knowledge bases. Features critical for clinical integration are highlighted. Commercial pricing is license-dependent and was obtained via institutional quotes where not publicly available.

Feature	CIViC	OncoKB	COSMIC	ClinVar	JAX-CKB
Access & Cost					
Access model	Open	Commercial	Commercial	Open	Commercial
Monthly cost	Free	By quote[a]	By quote[a]	Free	By quote[a]
API rate limits	Rate-limited[b]	Per license	Per license	Per NCBI policy[c]	Per license
Data export	JSON/TSV	License only	License only	XML/VCF	License only
Evidence Standards					
Grading system	A–E + stars	AMP/ASCO/CAP L1–4	Proprietary	ACMG/AMP	Proprietary
Evidence levels	5 levels + 5-star trust rating	4 levels + R1/R2	N/A	5-tier	4-tier
Curation model	Community	Expert panel	Automated	Mixed	Expert
Update frequency	Weekly	Monthly	Quarterly	Daily	Monthly
Content Coverage					
Gene coverage	470+	680+	20,000+	50,000+	400+
Variant count	3,200+	5,000+	1M+	2M+	3,500+
Drug associations	6,300+	4,500+	Limited	N/A	2,000+
Clinical trials	Limited	Extensive	None	None	Extensive
Usability & Integration					
Natural language	Partial	No	No	No	No
EHR integration	Limited	Partial	None	Partial	Partial
Cross-references	Yes	Partial	Partial	Yes	Partial
Quality control	Peer review	Expert panel	Algorithmic	Variable	Expert

[a]Commercial databases require institutional licensing; pricing varies by institution size and use case.
[b]CIViC API documentation describes throttling behavior; see `https://docs.civicdb.org`.
[c]NCBI E-utilities usage policies apply; API key recommended for programmatic access.

Key Observations.

Access Barriers. While CIViC and ClinVar provide full open access, OncoKB, COSMIC, and JAX-CKB impose commercial licensing barriers that disadvantage institutions with limited resources. This creates a two-tiered system in which well-funded cancer centers have access to comprehensive genomic interpretation resources, while community oncology practices—where most cancer patients receive care—must rely on incomplete free resources.

Evidence Grading Heterogeneity. No two resources use an identical evidence classification system:

- **OncoKB**: Levels 1–4 aligned to AMP/ASCO/CAP guidelines, plus R1/R2 for resistance variants
- **CIViC**: Dual system with A–E evidence levels (study quality) and a 1–5-star trust rating (curation confidence)
- **ClinVar**: ACMG/AMP 5-tier pathogenicity classification (pathogenic, likely pathogenic,

VUS, likely benign, benign)

- **COSMIC**: Proprietary scoring without published methodology
- **JAX-CKB**: 4-tier actionability levels specific to their platform

As a result, the same variant may be classified differently across resources, and clinicians must understand multiple systems to interpret results consistently.

Coverage Gaps. No single database captures all actionable alterations. COSMIC and ClinVar have the broadest variant coverage but limited therapeutic annotations. OncoKB and JAX-CKB have extensive drug associations but cover fewer genes. CIViC provides the richest evidence linking but has the smallest variant catalog. For comprehensive interpretation, users must query multiple databases—each with different interfaces, query syntaxes, and output formats.

Integration Limitations. All databases show limited EHR integration capability, with no support for natural language queries that would enable clinicians to ask questions in plain English rather than constructing structured queries with precise nomenclature. This forces reliance on bioinformatics specialists or specialized training that many clinical personnel lack.

S0.2 The Six Barriers and How OncoCITE Addresses Them

Table S2 maps the six systematic barriers identified in the main text to the specific OncoCITE features that address each limitation.

Table 2: **Table S2.** Mapping of systematic barriers to OncoCITE solutions with implementation details.

Barrier	Problem Description	OncoCITE Solution	Implementation
(i) Rigid search interfaces	Requires exact gene symbols (e.g., "ERBB2" not "HER2"), specific variant notation (HGVS), and Boolean operators	Natural language understanding; automatic gene alias resolution	Gene synonym lookup via NCBI; fuzzy matching algorithms
(ii) Heterogeneous schemas	Each database uses different field names, data types, and relationship structures	Unified 45-field JSON schema compatible with CIViC	Schema validation at extraction time; Critic agent enforcement
(iii) Divergent evidence grading	AMP/ASCO/CAP, ACMG/AMP, A–E levels, and trust ratings—no interoperability	Standardized evidence types (PREDICTIVE, PROGNOSTIC, DIAGNOSTIC) with explicit 0–1 confidence scores	Planner agent classification; confidence calibration
(iv) Commercial paywalls	$2,500–$10,000+/month licenses exclude smaller institutions	Fully open-source under MIT license; extracts from open-access literature	GitHub repository; containerized deployment
(v) Coverage gaps	2–3 year lag between publication and database inclusion for emerging targets	Real-time extraction from PDFs within minutes of publication	Direct PDF processing; no manual curation queue
(vi) Limited EHR integration	No natural language APIs; minimal structured output options	RESTful API; JSON/TSV export; confidence thresholds for clinical filtering	Express.js backend; React frontend; Docker deployment

Table 3: **Table S2b.** OncoCITE addresses each systematic limitation identified through comprehensive analysis of the CIViC database (11,312 evidence items, 3,083 sources). Quantified findings reference Figure 2 and Supplementary Figure S1; technical specifications detailed in Supplementary Tables S17–S21.

EDA Problem	**Quantified Finding**	**OncoCITE Solution**	**Component**
Curation latency	Median 31 days, P90 >21 months (Fig. 2B)	Minutes per paper	Multi-agent pipeline
Long-tail scalability	Median 1 item/source (Fig. 2A; Supp. Fig. S1A)	Automated batch processing	Reader + Extractor
Literature growth	2010–2020 surge (Supp. Fig. S1B)	Real-time extraction	Full pipeline
Pre-clinical skew	39.2% Predictive Level D (Supp. Fig. S1C)	Explicit level classification	Planner agent
Trust variance	Wide distribution (Supp. Fig. S1D)	Confidence scores (0-1)	Critic agent
Resistance gap	Underrepresented (Supp. Fig. S1E)	Comprehensive extraction	Extractor agent
Conflicting evidence	238 combinations (Fig. 2C)	Verbatim provenance	Critic + source attribution
Somatic bias	4,882 vs 108 (Supp. Fig. S1F)	Origin-agnostic extraction	Extractor agent
Disease concentration	VHL: 2,800+ (Supp. Fig. S1G)	Disease-agnostic processing	Full pipeline
Emerging target gap	GPRC5D: 0 items (Fig. 2D)	Immediate extraction	Real-time processing
Schema heterogeneity	Database-specific	Unified 45-field JSON	Schema validation
Missing ontology IDs	Inconsistent	Standardized IDs	Normalizer agent

Supplementary Note 1: Multiple Myeloma Evidence Extraction — Validation Against CIViC Ground Truth

S1.1 Overview and Rationale

This case study provides a rigorous evaluation of the multi-agent extraction system against expert-curated ground truth from the Clinical Interpretation of Variants in Cancer (CIViC) database [1, 2]. CIViC is a community-curated knowledge base in which evidence items undergo expert review before inclusion, making it a suitable benchmark for assessing extraction accuracy. We selected Multiple Myeloma as the evaluation disease based on domain expertise, enabling thorough manual verification of extracted evidence against source publications.

Corpus Selection. Using the CIViC disease filter for "Multiple Myeloma," we identified 10 publications containing expert-curated evidence items. This corpus spans 2000–2015 and represents the complete set of Multiple Myeloma papers with curated evidence in CIViC at the time of evaluation (CIViC v2 API, accessed December 15, 2024). The papers cover five major target categories: (1) RAS family mutations (KRAS, NRAS) and chemotherapy response; (2) FGFR3 alterations and response to tyrosine kinase inhibitors; (3) BRAF V600E and response to BRAF/MEK inhibitors; (4) CRBN mutations and immunomodulatory drug resistance; and (5) cancer-testis antigens (NY-ESO-1, LAGE-1) and TCR-based immunotherapy response.

Evaluation Objectives. This case study addresses three key questions: (1) Can the system recover evidence deemed clinically significant by expert curators? (2) Does the system identify valid evidence missed by curators, and at what precision? (3) Does the system produce clinically dangerous errors that would affect patient care? These questions correspond to the three primary metrics: ground truth recovery rate, novel discovery precision, and critical error rate.

S1.2 Evaluation Metrics: Definitions and Rationale

We designed the evaluation metrics to capture distinct aspects of extraction quality relevant to clinical knowledge base curation. Each metric addresses a specific performance question:

Ground Truth Recovery Rate measures the proportion of expert-curated evidence items that the system successfully identifies. This metric asks: "Does the system find what human experts found?" A high recovery rate indicates that the system achieves comparable coverage to expert curation. We report this as: Recovery Rate = (GT items matched by system) / (Total valid GT items) $\times 100\%$. Matching employed hierarchical criteria: Level 1 (Exact) requires all core fields match; Level 2 (Core) requires gene + disease + evidence direction match; Level 3 (Subsumption) allows codon-level variants to match specific amino acid changes (e.g., "CODON 12 MUTATION" matches G12A, G12C, G12D, G12S). Standard gene aliases (CTAG1B=NY-ESO-1, CTAG2=LAGE-1) and drug name equivalences (PKC412=Midostaurin, PLX4032=Vemurafenib) were applied.

Novel Discovery Precision measures the proportion of system-extracted items not present in the ground truth that are scientifically valid upon manual verification. This metric asks: "When the system extracts evidence beyond what curators found, is it correct?" A high novel discovery precision indicates that expanded extraction does not sacrifice accuracy. We report this as: Discovery Precision = (Verified novel items) / (Evaluable novel items) $\times 100\%$. Of 48 novel items, 3 were excluded as near-duplicates; 44/45 evaluable items verified as valid. Each novel discovery includes an extraction confidence score (0–1), enabling users to prioritize high-confidence items for downstream use.

Critical Error Rate measures the proportion of extracted items containing errors that would

affect clinical interpretation. This metric asks: "Does the system make dangerous mistakes?" Critical errors include wrong gene attribution (e.g., claiming KRAS when a paper discusses NRAS), incorrect variant position (e.g., Y375C instead of Y373C), reversed therapeutic direction (sensitivity reported as resistance or vice versa), or fabricated data not present in the source. We report this as: Error Rate = (Items with critical errors) / (Total extracted items) ×100%. This is the most important safety metric—even a system with high recovery and discovery rates would be unsuitable for clinical use if it produces errors that could affect patient care.

S1.3 Primary Validation Results

Table S3 presents the primary validation metrics with 95% confidence intervals. These results demonstrate strong performance across all three evaluation dimensions: recovery of expert-curated evidence, precision on novel discoveries, and absence of clinically significant errors.

Table 4: **Table S3.** Primary validation metrics with 95% confidence intervals. Critical error CI calculated for the 69-item retrospective corpus; for the combined 108-item corpus, CI is 0–3.4%.

Metric	Result	95% CI	Interpretation
Ground Truth Recovery	**84.0% (21/25)**	65.3–93.6%	Strong expert concordance
Novel Discovery Precision	**97.8% (44/45)**[a]	88.4–99.9%	High precision on new findings
Critical Error Rate	**0% (0/69)**	0–5.2%	No clinically dangerous errors
GT Curation Errors Detected	**24.2% (8/33)**	12.4–41.0%	QA capability demonstrated

[a]Of 48 novel items, 3 excluded as near-duplicates; 44/45 evaluable items verified as valid.

Interpretation of Results. The 84% ground truth recovery rate indicates that the system identifies most evidence deemed clinically significant by expert curators. The 97.8% novel discovery precision indicates that evidence extracted beyond the ground truth is almost always scientifically valid, suggesting that expanded extraction does not come at the cost of reliability. The 0% critical error rate is the most clinically important finding: across 69 extracted items, none contained errors that would affect clinical interpretation (e.g., wrong gene attribution, incorrect variant position, or reversed sensitivity/resistance direction). Finally, the detection of curation errors in 24.2% of ground truth items highlights that even expert-curated databases can contain errors, and that AI systems with explicit source attribution can help identify discrepancies for re-review.

S1.4 Per-Paper Performance Analysis

Table S4 presents extraction performance for each of the 10 CIViC-indexed publications. This breakdown reveals consistent performance across papers with varying characteristics, including publication year, evidence types, and therapeutic targets. Two papers contained verified ground truth curation errors and were excluded from recovery calculations.

Table 5: **Table S4.** Per-paper extraction performance. GT = ground truth items; Novel = verified novel discoveries (excludes 3 near-duplicates from total 48 novel items); Errors = critical errors. Papers with verified ground truth curation errors are excluded from recovery calculations.

PMID	Year	GT	System	Recovery	Novel	Errors	GT Quality	Primary Targets
11050000	2000	4	8	100% (4/4)	4	0	Valid	KRAS, TP53, NRAS
12483530	2002	4	6	N/A	6	0	GT Error	KRAS, NRAS
15339850	2005	4	4	N/A	4	0	GT Error	KRAS, NRAS
16091734	2005	4	12	100% (4/4)	8	0	Valid	FGFR3, ETV6-FGFR3
18528420	2008	5	8	80% (4/5)	3	0	Valid	KRAS, NRAS
19381019	2009	6	12	100% (6/6)	6	0	Valid	FGFR3
23480694	2013	1	4	100% (1/1)	3	0	Valid	CRBN, PSMG2, NR3C1
23612012	2013	1	3	100% (1/1)	2	0	Valid	BRAF V600E
24997557	2014	2	3	100% (2/2)	1	0	Valid	BRAF V600E
26193344	2015	2	9	100% (2/2)	7	0	Valid	NY-ESO-1, LAGE-1, HLA-A
TOTAL	—	**33**	**69**	**84% (21/25)**	**44**[a]	**0**	—	—

[a]Total novel items = 48; 3 excluded as near-duplicates; 44 verified as valid out of 45 evaluable.

Key Observations. The system achieved 100% recovery on 7 of 8 papers with valid ground truth, with the single partial recovery (80% for PMID 18528420) reflecting a minor classification difference (DIAGNOSTIC vs PROGNOSTIC) rather than a substantive extraction failure. The ratio of system items to ground truth items ranged from 1.0× to 4.5×, indicating consistent discovery of valid evidence beyond expert curation across all papers. Notably, the papers with the highest discovery ratios (PMIDs 16091734, 23480694, 26193344) contained novel fusion genes, resistance mutations, and immunotherapy biomarkers that were scientifically valid but not captured in CIViC curation.

S1.5 Ground Truth Curation Errors Identified

A significant finding of this evaluation was the identification of curation errors in the expert ground truth itself. Manual verification of all 33 ground truth items against source PDFs revealed that 8 items (24.2%) from 2 papers contained systematic errors. This finding has important implications: even community-curated, expert-reviewed databases can contain errors, and AI systems with explicit source attribution and confidence scoring can support quality assurance by flagging discrepancies for re-review. Tables S5 and S6 document these errors in detail.

PMID 12483530 — Therapy Misattribution from Citation (4 items affected)

The CIViC ground truth asserts that KRAS G12A/C/D/S mutations confer resistance to Melphalan in Multiple Myeloma. However, examination of the source PDF indicates that this paper does not experimentally test Melphalan. Fig. 5 tests dexamethasone-induced apoptosis; Fig. 6 tests doxorubicin-induced apoptosis. The mention of Melphalan appears in the Introduction as a citation to a different paper (Rowley et al., 2000; PMID 11050000), not as new experimental data from this study.

Table 6: **Table S5.** PMID 12483530 ground truth error: therapy misattribution from citation.

Field	CIViC Ground Truth	Actual Paper Content
Therapy	Melphalan	Dexamethasone (Fig. 5), Doxorubicin (Fig. 6)
Evidence Source	This paper	Citation to PMID 11050000
System Extraction	—	Correctly identified Dexamethasone, Doxorubicin

Paper Quote (Introduction): *"ras mutations have been shown to protect myeloma cell lines from apoptosis induced by dexamethasone, doxorubicin, or melphalan (Rowley et al., 2000)"* — The parenthetical citation indicates this is referenced work, not new data from this study.

PMID 15339850 — Evidence Type Misclassification (4 items affected)

The CIViC ground truth classifies this paper as PREDICTIVE evidence for Melphalan resistance. However, this paper is a mutation frequency study comparing RAS mutation prevalence between MGUS (5%) and Multiple Myeloma (31%) patients. It contains no therapeutic response data—no drugs are tested and no treatment outcomes are reported. This is purely PROGNOSTIC evidence (mutation associated with disease stage), but it is classified as PREDICTIVE evidence (mutation associated with drug response) in the ground truth.

Table 7: **Table S6.** PMID 15339850 ground truth error: evidence type misclassification.

Field	CIViC Ground Truth	Actual Paper Content
Evidence Type	PREDICTIVE	PROGNOSTIC
Therapy	Melphalan	None (no drugs tested)
Study Design	Drug response study	Cross-sectional mutation frequency
System Extraction	—	Correctly classified as PROGNOSTIC, no therapy

Paper Quote (Abstract): *"One RAS mutation was identified in 20 MGUS tumors (5%), in contrast to a much higher prevalence in MM (31%)"* — This describes mutation frequency by disease stage, not drug response.

Implications for Ground Truth Quality. These errors represent two common curation failure modes: (1) citation confusion, where findings from referenced papers are incorrectly attributed to the current paper; and (2) evidence type misclassification, where observational studies are incorrectly labeled as interventional. Both error types would be difficult to detect through traditional database quality checks but are readily identified by AI systems that maintain explicit links between extracted claims and source text. This suggests a potential quality assurance application for multi-agent extraction systems as verification overlays for existing curated databases.

S1.6 Novel Discoveries with Clinical Significance

The system identified 48 evidence items not present in the CIViC ground truth. Of these, 3 were excluded as near-duplicates of other extracted items, leaving 45 evaluable items. Manual verification against source PDFs confirmed that 97.8% (44/45) were scientifically valid. Table S7 presents selected discoveries with immediate clinical relevance. These findings demonstrate that expert curation, while valuable, is not exhaustive—automated systems can identify clinically actionable evidence that is missed by manual curation.

Table 8: **Table S7.** Selected novel discoveries with clinical significance. All findings verified against source PDFs. Each item includes confidence score enabling prioritization for downstream use.

PMID	Gene	Variant	Therapy	Type	Clinical Implication
11050000	TP53	Wild-type	Doxorubicin	PREDICTIVE	p53 status as drug biomarker
11050000	TP53	Wild-type	Melphalan	PREDICTIVE	p53 status as drug biomarker
11050000	NRAS	Codon 12 mut	Dexamethasone	PREDICTIVE	Expands RAS beyond KRAS
16091734	ETV6-FGFR3	TEL-FGFR3	PKC412	PREDICTIVE	Novel fusion target
16091734	ZNF198-FGFR1	Fusion	PKC412	PREDICTIVE	Additional fusion target
23480694	PSMG2	E171K	Bortezomib	PREDICTIVE	Novel resistance gene
23480694	NR3C1	G369A	Dexamethasone	PREDICTIVE	GR mutation → steroid resistance
26193344	HLA-A	A*0201+	TCR T-cell	PREDICTIVE	Patient selection biomarker
26193344	NY-ESO-1	Loss	TCR T-cell	PREDICTIVE	Resistance mechanism

Clinical Significance of Key Discoveries. Several novel discoveries have immediate therapeutic implications: (1) **PSMG2 E171K** and **NR3C1 G369A** mutations (PMID 23480694) represent novel resistance mechanisms to bortezomib and dexamethasone, respectively—two backbone agents in Multiple Myeloma therapy; (2) **HLA-A*0201 status** (PMID 26193344) is a patient-selection biomarker for TCR-based immunotherapy and is essential for clinical trial eligibility; (3) **NY-ESO-1/LAGE-1 antigen loss** (PMID 26193344) represents an acquired resistance mechanism to antigen-directed therapy, informing treatment sequencing decisions.

S1.7 Evidence Quality and Annotation Completeness

Beyond extraction volume, the system provides substantially richer annotation than the CIViC ground truth. Table S8 compares annotation completeness across key fields. These additional fields serve specific purposes: verbatim quotes enable rapid human verification; page references support source lookup; confidence scores enable risk-stratified review; and ontology identifiers enable cross-database integration.

Table 9: **Table S8.** Annotation completeness comparison. pp = percentage points. Novel items: 48 total, 3 excluded as near-duplicates, 44/45 evaluable verified as valid (97.8%).

Annotation Feature	Ground Truth (n=33)	System (n=69)	Δ
Total evidence items	33	**69**	+109%
Fields per item	11	**45**	+309%
Verbatim supporting quote	0 (0%)	**65 (95%)**	+95pp
Source page/figure reference	0 (0%)	**65 (95%)**	+95pp
Extraction confidence score	0 (0%)	**69 (100%)**	+100pp
Entrez Gene ID	0 (0%)	**69 (100%)**	+100pp
Disease ontology ID (EFO)	0 (0%)	**69 (100%)**	+100pp
Drug ontology ID (RxNorm/NCIt)	0 (0%)	59 (85%)	+85pp

Value of Enhanced Annotation. The 95% verbatim quote rate is particularly valuable for clinical knowledge bases because it enables rapid human verification of extracted claims without requiring full-text review. The 100% confidence score coverage supports risk-stratified workflows in which high-confidence items can proceed to direct use while lower-confidence items are flagged for expert review. The 100% ontology identifier coverage enables seamless integration with downstream applications including variant annotation pipelines, clinical decision support systems, and cross-database queries.

S1.8 Prospective Application: Extraction from Uncurated Recent Literature

To demonstrate that the system generalizes beyond retrospective validation against curated databases, we applied it to five recent publications (2022–2024) that were not yet indexed in CIViC. These papers were selected based on (1) publication in high-impact journals (*Nature Medicine*, *Nature Cancer*, *Blood Neoplasia*, *JCO Precision Oncology*); (2) clinically significant biomarkers for emerging immunotherapy modalities; and (3) direct relevance to contemporary therapeutic decision-making in Multiple Myeloma. This prospective application reflects the primary use case for automated extraction: accelerating knowledge base coverage in rapidly evolving therapeutic areas where curated resources inherently lag behind the primary literature.

Table 10: **Table S9.** Prospective extraction corpus: recent immunotherapy-resistance publications (0% coverage in CIViC as of December 2024).

Publication	**Journal**	**Evidence items**	**Key targets and focus**
Da Và et al., 2023	Nature Medicine	9	TNFRSF17 (BCMA): bispecific antibodies/CAR-T resistance
Derrien et al., 2023	Nature Cancer	8	GPRC5D: Talquetamab resistance mutations
Dutta et al., 2024	Blood Neoplasia	11	BCL2, MCL1, CDK7: Venetoclax response biomarkers
Restrepo et al., 2022	JCO Precision Oncology	5	WNT10A, DUSP1, ETV7: Selinexor response signature
Elnaggar et al., 2022	Journal of Hematology & Oncology	6	BRAF, KRAS: MAPK inhibitor combinations
TOTAL	—	**39**	16 unique genes; 0 items in CIViC

Rationale for Paper Selection. These five papers were chosen because they report resistance mechanisms and biomarkers for therapeutic modalities that have transformed Multiple Myeloma treatment in the past 2–3 years: (1) bispecific antibodies targeting BCMA (teclistamab, elranatamab) and GPRC5D (talquetamab); (2) CAR-T cell therapy (idecabtagene vicleucel); (3) BCL2 inhibitors (venetoclax); and (4) XPO1 inhibitors (selinexor). Curated knowledge bases inherently lag behind the literature for emerging therapies, creating gaps that affect clinical decision-making. Automated extraction can accelerate knowledge base coverage in these rapidly evolving areas.

S1.9 BCMA (TNFRSF17) Resistance Mechanisms

From Da Và et al., Nature Medicine, 2023. BCMA-targeting therapies (bispecific antibodies and CAR-T cells) have achieved unprecedented response rates in relapsed/refractory Multiple Myeloma, but acquired resistance is an emerging clinical challenge. This study reports BCMA mutations and deletions that confer resistance to multiple BCMA-targeting agents. These findings have immediate implications for patient monitoring and treatment sequencing.

Table 11: **Table S10.** BCMA variants associated with resistance to bispecific antibodies and CAR-T therapy. These alterations affect the BCMA extracellular domain and disrupt antibody/CAR binding.

Gene	Variant	Affected therapies	Mechanism
TNFRSF17	Biallelic deletion	Idecabtagene vicleucel, Teclistamab	Complete antigen loss
TNFRSF17	p.Arg27Pro	Teclistamab, Elranatamab, Alnuctamab	Epitope disruption
TNFRSF17	p.Pro34del	Teclistamab, Elranatamab, Alnuctamab	Epitope disruption
TNFRSF17	p.Ser30del	Teclistamab, Elranatamab, Alnuctamab	Epitope disruption

Clinical Implications. These findings have immediate relevance for clinical practice: (1) patients progressing on BCMA-targeting therapy should be evaluated for BCMA mutations/deletions; (2) the specific epitope mutations (p.Arg27Pro, p.Pro34del, p.Ser30del) affect bispecific antibodies but may not affect all CAR-T constructs depending on epitope targeting; and (3) patients with BCMA loss may be candidates for alternative targets such as GPRC5D-directed therapy.

S1.10 GPRC5D Resistance Mutations

From Derrien et al., Nature Cancer, 2023. GPRC5D is an emerging target for bispecific antibody therapy (Talquetamab) in Multiple Myeloma. This study provides a comprehensive catalog of GPRC5D alterations associated with Talquetamab resistance, encompassing multiple resistance mechanisms including frameshift and nonsense variants, in-frame deletions, structural alterations, and copy-number loss.

Table 12: **Table S11.** GPRC5D variants associated with Talquetamab resistance. Mechanisms include frameshift/nonsense (protein loss), in-frame deletion (epitope disruption), structural alteration (transcriptional silencing), and copy-number loss (complete antigen loss).

Gene	Variant	Variant type	Therapy	Predicted effect
GPRC5D	E27fs	Frameshift	Talquetamab	Truncation; loss of function
GPRC5D	S125fs	Frameshift	Talquetamab	Truncation; loss of function
GPRC5D	F158fs	Frameshift	Talquetamab	Truncation; loss of function
GPRC5D	W217*	Nonsense	Talquetamab	Premature stop; loss of function
GPRC5D	W237*	Nonsense	Talquetamab	Premature stop; loss of function
GPRC5D	G97_F100del	In-frame deletion	Talquetamab	Epitope disruption
GPRC5D	TSS deletion	Structural	Talquetamab	Transcriptional silencing
GPRC5D	LOSS	Copy number	Talquetamab	Complete antigen loss

S1.11 Venetoclax Response Biomarkers

From Dutta et al., Blood Neoplasia, 2024. Venetoclax is a BCL2 inhibitor with activity in Multiple Myeloma, particularly in patients with t(11;14) translocation. This study reports a biomarker panel for Venetoclax response prediction, including established markers and a novel gene signature that may expand patient selection beyond t(11;14) status.

Table 13: **Table S12.** Biomarkers associated with Venetoclax response. A novel 6-gene signature (ATP1B3, MYL2, CNR1, FKBPL, LRRK2-DT, LINC02541) may expand patient selection beyond t(11;14).

Gene(s)	**Biomarker**	**Direction**	**Combination**	**Clinical utility**
CCND1	t(11;14)	Sensitivity	Venetoclax monotherapy	Patient selection
MCL1	Gain/Amplification	Resistance	Venetoclax	Resistance mechanism
BCL2	High expression	Sensitivity	Venetoclax	Response prediction
BCL2/MCL1	High ratio	Sensitivity	Venetoclax	Response prediction
CDK7	Expression	Sensitivity	Mevociclib + Venetoclax	Synergy biomarker
ATP1B3/MYL2/CNR1	6-gene signature	Sensitivity	Venetoclax	Novel biomarker panel

S1.12 Combined Extraction Metrics

Table S13 summarizes performance across both the retrospective validation corpus (CIViC-indexed publications) and the prospective corpus (recent, uncatalogued literature). Consistent performance across both corpora—including 0% critical errors in each—supports generalizability beyond the specific papers and evidence types used for retrospective validation.

Table 14: **Table S13.** Combined extraction metrics across retrospective and prospective corpora. For the combined 108-item corpus, the 95% confidence interval for the critical error rate is 0–3.4%.

Metric	**CIViC validation (n=10)**	**Prospective (n=5)**	**Combined**
Total evidence items	69	39	**108**
Items per paper (mean)	6.9	7.8	7.2
Unique genes	13	16	**26**
Unique variants	25	23	**45**
Unique therapies	12	16	**25**
Critical errors	0 (0%)	0 (0%)	**0 (0%)**[a]
Items with confidence score	69 (100%)	39 (100%)	108 (100%)
Items with verbatim quote	65 (95%)	39 (100%)	104 (96%)

[a]For the combined 108-item corpus, 95% CI: 0–3.4% (for the 69-item retrospective corpus alone, 95% CI: 0–5.2%).

Summary. Across 15 papers (10 CIViC-indexed and 5 prospective), the system extracted 108 evidence items spanning 26 unique genes, 45 unique variants, and 25 unique therapies, with a 0% critical error rate (95% CI: 0–3.4%). All items include a confidence score for risk-stratified downstream use, and 96% include verbatim supporting quotes for rapid verification. The prospective corpus highlights extraction from emerging therapeutic areas (bispecific antibodies, CAR-T, and targeted combinations) where curated knowledge bases have not yet achieved coverage.

Supplementary Note 2: Code Availability and Technical Documentation

S2.1 Repository Information

The complete OncoCITE source code is publicly available in two implementations:

- **Claude Agent SDK implementation (repository):** `https://github.com/Ali-Maq/civic-extraction-agent`
- **LangChain implementation (repository):** `https://github.com/Ali-Maq/oncocite-langchain`
- **License:** MIT License (both repositories)
- **Languages:** Python 3.11+ and Node.js 18+
- **Frameworks:** Claude Agent SDK or LangChain; React 18; Express.js

The LangChain implementation was additionally validated using GLM-4 (9B parameters) accessed via Fireworks AI, showing comparable qualitative extraction behavior on the Multiple Myeloma case study and suggesting that the overall system performance is driven primarily by the architecture rather than model-specific capabilities.

S2.2 System Requirements

Table 15: **Table S14.** Minimum system requirements for OncoCITE deployment.

Component	Requirement
Python	3.11+
Node.js	18.x+
RAM	1 GB minimum (API mode)[a]; 4–8 GB for local processing
Storage	10 GB for base installation
API access	Anthropic API key (Claude SDK) or any LangChain-compatible provider

[a]API mode: LLM inference via external API; local instance hosts only the web interface and the MCP server.

S2.3 Installation

```
# Option A: Claude Agent SDK implementation

git clone https://github.com/Ali-Maq/civic-extraction-agent.git
cd civic-extraction-agent

pip install -r requirements.txt

cd frontend && npm install && cd ..

export ANTHROPIC_API_KEY="your-api-key"
```

```
python run_extraction.py --input paper.pdf --output results/


# Option B: LangChain implementation

git clone https://github.com/Ali-Maq/oncocite-langchain.git
cd oncocite-langchain

pip install -r requirements.txt

cd frontend && npm install && cd ..

# Example provider configuration
export FIREWORKS_API_KEY="your-api-key"

python run_extraction.py --input paper.pdf --output results/
```

S2.4 Docker Deployment

```
# Build container image

docker build -t oncocite:latest .

# Run with API key

docker run -d -p 80:80 \
  -e ANTHROPIC_API_KEY="your-api-key" \
  -v $(pwd)/data:/app/data \
  oncocite:latest
```

S2.5 AWS Deployment

The system is deployed on AWS infrastructure:

- **Compute:** EC2 t4g.micro (ARM64, 2 vCPU, 1 GB RAM)
- **Container registry:** Amazon ECR
- **Storage:** EBS 20 GB gp3
- **Architecture note:** The EC2 instance hosts the web interface, API server, and MCP server wrapper; computationally intensive LLM inference occurs via external API calls to Anthropic's cloud infrastructure (or other LangChain-compatible providers).
- **Estimated cost:** $7–10/month for hosting infrastructure (excludes API usage costs)

A live demonstration instance is described in the repository README.

S2.6 MCP Tool Reference

Table S15 lists all 22 Model Context Protocol (MCP) tools available to the multi-agent system.

Table 16: **Table S15.** Complete MCP tool reference.

Tool Name	Agent(s)	Function
save_paper_content	Reader	Persist extracted paper content
get_paper_content	All	Retrieve paper content
save_extraction_plan	Planner	Save extraction strategy
get_extraction_plan	Extractor	Retrieve extraction strategy
check_actionability	Extractor	Check clinical relevance
validate_evidence_item	Extractor	Check schema compliance
save_evidence_items	Extractor	Save validated items
get_evidence_items	Critic, Normalizer	Retrieve current items
save_critique	Critic	Save validation results
increment_iteration	Orchestrator	Increment refinement iteration
lookup_gene_entrez	Normalizer	Look up NCBI Gene ID
lookup_rxnorm	Normalizer	Look up RxNorm ID
lookup_therapy_ncit	Normalizer	Look up NCIt therapy code
lookup_efo	Normalizer	Look up EFO disease ID
lookup_disease_doid	Normalizer	Look up DOID
lookup_clinical_trial	Normalizer	Verify NCT identifier
lookup_variant_info	Normalizer	Look up genomic coordinates/HGVS
save_final_output	Orchestrator	Save final extraction output
get_workflow_status	All	Retrieve workflow status
log_agent_action	All	Log agent actions (audit trail)
save_checkpoint	Orchestrator	Save intermediate checkpoint
restore_checkpoint	Orchestrator	Restore from checkpoint

Supplementary Note 3: Detailed Technical Methods

This supplementary note provides comprehensive technical specifications for the OncoCITE multi-agent extraction pipeline, enabling full reproducibility of the system. All specifications are derived directly from the production codebase available in the GitHub repository.

S3.1 Vision-Based PDF Processing Pipeline

The Reader agent implements a vision-first approach to PDF processing, treating each page as an image rather than extracting text. This design choice enables accurate extraction from complex layouts including multi-column text, embedded tables, figures with annotations, and supplementary materials that often defeat text-based PDF parsers.

PDF Rendering Configuration

Table 17: **Table S16.** PDF rendering parameters for vision-based extraction.

Parameter	Value	Rationale
Resolution	300 DPI	Balances text legibility with file size
Output format	JPEG	Compressed format reduces API payload
Library	PyMuPDF (fitz)	Fast rendering with accurate layout preservation
Color space	RGB	Required for figure interpretation

Context Window Management

Claude 3.5 Sonnet provides a 200,000-token context window, but processing high-resolution images of multi-page papers requires careful management to avoid context overflow. The system implements a chunking strategy:

- **Images per turn:** 2 pages maximum per API call
- **Sequential processing:** For a paper with N pages, the Reader issues $\lceil N/2 \rceil$ sequential API calls
- **Cumulative extraction:** Each turn builds upon previously extracted content, with the Reader maintaining running lists of genes, variants, diseases, and therapies
- **Final consolidation:** After all pages are processed, the Reader produces a unified structured JSON output

S3.2 Evidence Schema Specification

The evidence schema defines 45 total fields organized into two tiers: 25 Tier-1 extraction fields populated directly from paper content, and 20 Tier-2 normalization fields populated via external API lookups.

Tier-1 Extraction Fields (25 fields)

Table 18: **Table S17.** Tier-1 extraction fields with descriptions and examples.

Field Name	Description	Example
Core Required Fields (8)		
`feature_names`	Gene symbol(s)	BRAF
`variant_names`	Variant designation	V600E
`disease_name`	Disease name	Melanoma
`evidence_type`	PREDICTIVE, PROGNOSTIC, DIAGNOSTIC, PREDISPOSING, ONCOGENIC, FUNCTIONAL	PREDICTIVE
`evidence_level`	A (validated), B (clinical), C (case), D (preclinical), E (inferential)	B
`evidence_direction`	SUPPORTS or DOES_NOT_SUPPORT	SUPPORTS
`evidence_significance`	Type-specific significance value	SENSITIZES
`evidence_description`	1–3 sentence summary with statistics	"V600E mutation..."
Variant Fields (6)		
`variant_origin`	SOMATIC, GERMLINE, N/A	SOMATIC
`variant_type_names`	Mutation type	Missense
`variant_hgvs_descriptions`	HGVS notation	p.V600E
`molecular_profile_name`	Complex alteration description	BRAF V600E
`fusion_five_prime_gene_names`	5' fusion partner	EML4
`fusion_three_prime_gene_names`	3' fusion partner	ALK
Feature Fields (2)		
`feature_full_names`	Full gene name	B-Raf proto-oncogene
`feature_types`	GENE or FACTOR	GENE
Disease Fields (1)		
`disease_display_name`	Display-friendly disease name	Non-Small Cell Lung Cancer
Therapy Fields (2)		
`therapy_names`	Drug name(s)	Vemurafenib
`therapy_interaction_type`	COMBINATION, SEQUENTIAL, SUBSTITUTES	COMBINATION
Source Fields (3)		
`source_title`	Paper title	"Targeting BRAF..."
`source_publication_year`	Publication year	2013
`source_journal`	Journal name	Cancer Discovery
Clinical Trial Fields (2)		
`clinical_trial_nct_ids`	NCT identifier(s)	NCT01234567
`clinical_trial_names`	Trial name(s)	BRAF V600E Study
Phenotype Field (1)		
`phenotype_names`	Associated phenotype	Drug resistance

Tier-2 Normalization Fields (20 fields)

Table 19: **Table S18.** Tier-2 normalization fields populated via external API lookups. Database-scale normalization of all 11,312 CIViC items achieved 83.12% overall success rate across these fields.

Field Name	Description	Source API
Ontology Identifiers (5)		
`disease_doid`	Disease Ontology ID	OLS/DOID
`gene_entrez_ids`	NCBI Entrez Gene ID	MyGene.info
`therapy_ncit_ids`	NCI Thesaurus drug ID	OLS/NCIt
`factor_ncit_ids`	Factor NCIt ID	OLS/NCIt
`variant_type_soids`	Sequence Ontology ID	SO mappings
Variant Identifiers (4)		
`variant_clinvar_ids`	ClinVar accession	MyVariant.info
`variant_allele_registry_ids`	ClinGen registry ID	MyVariant.info
`variant_mane_select_transcripts`	MANE transcript	MyVariant.info
`variant_rsid`	dbSNP rsID	MyVariant.info
Phenotype Identifiers (2)		
`phenotype_ids`	General phenotype ID	OLS
`phenotype_hpo_ids`	Human Phenotype Ontology ID	OLS/HPO
Source Identifiers (2)		
`source_citation_id`	PubMed ID (PMID)	PubMed E-utilities
`source_pmcid`	PubMed Central ID	NCBI ID Converter
Genomic Coordinates (7)		
`chromosome`	Chromosome number	MyVariant.info
`start_position`	Genomic start position	MyVariant.info
`stop_position`	Genomic stop position	MyVariant.info
`reference_build`	Reference genome (GRCh37/38)	MyVariant.info
`representative_transcript`	Canonical transcript	MyVariant.info
`reference_bases`	Reference allele	MyVariant.info
`variant_bases`	Alternate allele	MyVariant.info

Evidence Type and Significance Matrix

Each evidence type permits only specific significance values, enforced by the Critic agent:

Table 20: **Table S19.** Valid evidence significance values by evidence type.

Evidence Type	Valid Significance Values
PREDICTIVE	SENSITIZES, RESISTANCE, REDUCED_SENSITIVITY, N/A
PROGNOSTIC	BETTER_OUTCOME, POOR_OUTCOME, N/A
DIAGNOSTIC	POSITIVE, NEGATIVE
PREDISPOSING	PREDISPOSITION, PROTECTIVENESS
ONCOGENIC	ONCOGENICITY, PROTECTIVENESS
FUNCTIONAL	GAIN_OF_FUNCTION, LOSS_OF_FUNCTION, NEOMORPHIC, UNALTERED_FUNCTION, UNKNOWN, DOMINANT_NEGATIVE

S3.3 Multi-Agent Orchestration

The system employs five extraction and validation agents plus a Normalizer agent (six total), coordinated through explicit handoff protocols. Each agent has bounded responsibilities and communicates exclusively through the Model Context Protocol (MCP) tool ecosystem.

Agent Roles and Responsibilities

Table 21: **Table S20.** Agent roles, primary tools, and outputs. Tool names correspond to MCP tools listed in Table S15.

Agent	Primary Responsibility	Key Tools	Output
Reader	PDF visual extraction	`save_paper_content`	Structured JSON
Orchestrator	Workflow coordination, iteration control	`increment_iteration`, `get_workflow_status`, `save_final_output`	Delegation decisions
Planner	Extraction strategy	`get_paper_content`, `save_extraction_plan`	Strategy JSON
Extractor	Evidence identification	`check_actionability`, `validate_evidence_item`, `save_evidence_items`	Draft evidence items
Critic	Validation and QA	`get_evidence_items`, `save_critique`	`APPROVE/ NEEDS_REVISION/ REJECT`
Normalizer	Ontology enrichment	`lookup_*` tools[a]	Final evidence items

[a]Includes `lookup_gene_entrez`, `lookup_rxnorm`, `lookup_therapy_ncit`, `lookup_efo`, `lookup_disease_doid`, `lookup_variant_info`.

Iterative Refinement Loop

The Orchestrator implements a feedback loop capped at 3 iterations:

1. **Initial extraction:** Extractor produces draft evidence items
2. **Critic review:** Critic validates against source, returns assessment
3. **Decision point:**
   - APPROVE → Proceed to normalization
   - NEEDS_REVISION & iterations < 3 → Increment iteration, return to Extractor
   - NEEDS_REVISION & iterations $= 3$ → Finalize with current items
   - REJECT → Finalize with empty/minimal items

Empirical analysis shows 60% of papers complete in one iteration, 30% require two iterations, and 10% use all three iterations.

S3.4 Model Inference Settings and Validation Algorithm

Model Inference Configuration

All extraction and validation agents used Claude 3.5 Sonnet with the following deterministic inference settings to ensure reproducibility:

Table 22: **Table S20b.** Model inference settings for reproducible extraction.

Parameter	Value	Rationale
Model	claude-3-5-sonnet-20241022	Latest Sonnet with vision capabilities
Temperature	0.0	Deterministic output generation
Top_p	1.0	No nucleus sampling truncation
Max_tokens	200,000	Full context window utilization
Seed	Fixed per run	Enables exact reproducibility
API access	Anthropic API	December 2024 – January 2025

For the LangChain implementation, validation was performed using GLM-4 (9B parameters) accessed via Fireworks AI with equivalent settings, showing comparable qualitative behavior.

S3.5 Ontology Normalization Pipeline

The Normalizer agent enriches evidence items with standardized identifiers by querying external ontology services. Database-scale normalization of all 11,312 CIViC evidence items achieved an overall item-level success rate of 83.12%. Table S21 documents all API endpoints and their usage.

Table 23: **Table S21.** External API endpoints for ontology normalization.

Service	API Endpoint	Fields Populated
MyGene.info	`https://mygene.info/v3/query`	`gene_entrez_ids`
MyVariant.info	`https://myvariant.info/v1/query`	Genomic coordinates, ClinVar ID, rsID, HGVS
OLS/DOID	`https://www.ebi.ac.uk/ols/api/search` (`ontology=doid`)	`disease_doid`
OLS/NCIt	`https://www.ebi.ac.uk/ols/api/search` (`ontology=ncit`)	`therapy_ncit_ids`, `factor_ncit_ids`
OLS/EFO	`https://www.ebi.ac.uk/ols/api/search` (`ontology=efo`)	Disease EFO ID
OLS/HPO	`https://www.ebi.ac.uk/ols/api/search` (`ontology=hp`)	`phenotype_hpo_ids`
RxNorm	`https://rxnav.nlm.nih.gov/REST/approximateTerm.json`	Drug RXCUI
OpenFDA/FAERS	`https://api.fda.gov/drug/event.json`	Adverse event data
ClinicalTrials.gov	`https://clinicaltrials.gov/api/v2/studies/{nct_id}`	Trial status, phases
NCBI ID Converter	`https://pmc.ncbi.nlm.nih.gov/tools/idconv/api/v1/articles/`	`source_pmcid`

S3.6 Performance Characteristics

Table 24: **Table S23.** System performance metrics from validation corpus (n=15 papers).

Metric	Value
Processing Time	
Average time per paper	3–5 minutes
Reader phase	40–60% of total time
Normalization phase	20–30% of total time
Iteration Distribution	
Papers completing in 1 iteration	60%
Papers requiring 2 iterations	30%
Papers requiring 3 iterations	10%
Normalization Success (Database-Scale)	
Overall success rate (11,312 items)	83.12%
Gene Entrez ID lookup success	95%+
Disease DOID lookup success	90%+
Therapy NCIt lookup success	85%+
Field Coverage	
Average Tier-1 field coverage	85–95%
Average Tier-2 field coverage	60–70%
Verbatim quote attachment rate	96%

Table 25: **Table S24.** Normalization performance stratified by field category across all 11,312 CIViC evidence items (excluding variant coordinate fields, which are under active reconciliation). Success rate = proportion of field-level lookups returning a valid ontology identifier. Primary failure modes reflect limitations in source record nomenclature rather than system errors. The 83.12% item-level resolution rate reported in the main text accounts for all 20 Tier-2 fields.

Field Category	Lookups	Success (%)	Primary Failure Mode
Gene identifiers (Entrez)	11,312	95.2	Legacy symbols, non-gene features
Disease ontology (DOID/EFO)	22,624	89.7	Rare/historical disease names
Drug identifiers (RxNorm/NCIt)	22,624	82.4	Experimental compounds, brand names
Phenotype (HPO)	22,624	78.6	Broad phenotype descriptions
Source identifiers (PMID/PMC)	22,624	97.8	Preprints, non-indexed sources

Supplementary Note 4: Web Interface and Visualization

The OncoCITE web interface provides real-time evidence verification and knowledge graph exploration (Supplementary Figure S3), with an example of the standalone graph visualization shown in Supplementary Figure S4. Built with React 18.2 [7] and Vite 4.0 [8], the single-page application connects to a Node.js 18 [9] Express API server [10] (port 4177) serving three core endpoints: `GET /api/papers` (list available papers with status flags), `GET /api/papers/:id/pdf` (stream PDF with HTTP range request support for large files), and `GET /api/papers/:id/extractions` (return final evidence items and intermediate checkpoints).

PDF Evidence Verification. The PDF viewer panel implements side-by-side verification: the left pane displays the original paper using PDF.js 3.0 [11], with text layer enabled for searchability, while the right pane shows extracted evidence cards. When a user selects an evidence item, the system highlights the corresponding verbatim quote in the PDF if source page numbers are available, enabling rapid verification of extraction accuracy. This design directly addresses the "black box" problem of neural IE systems by making source attribution transparent and immediately verifiable.

Interactive Knowledge Graph. The knowledge graph panel uses ForceGraph2D (react-force-graph 1.43 [12], D3.js-based force-directed layout [13]) to visualize relationships between genes, variants, diseases, and therapies across all extracted evidence. Node types receive distinct visual encodings: genes (blue circles), variants (green diamonds), diseases (red hexagons), and drugs (orange squares). Edge thickness encodes evidence weight (confidence $\times$ evidence level), where evidence levels are mapped numerically as A=5, B=4, C=3, D=2, E=1, with labels indicating relationship direction (SENSITIVE, RESISTANT, POOR_OUTCOME). Display labels are derived from schema significance values by deterministic mapping: SENSITIZES $\rightarrow$ SENSITIVE, RESISTANCE $\rightarrow$ RESISTANT, POOR_OUTCOME $\rightarrow$ POOR_OUTCOME. Users can filter by evidence type, confidence threshold, or specific entities, enabling exploration of therapeutic networks and resistance mechanisms. The graph updates in real-time as new extractions complete.

Supplementary Note 5: Deployment Architecture

OncoCITE deploys as a Docker container [14] on AWS infrastructure (Supplementary Figure S5), enabling reproducible deployment and version control of the complete extraction pipeline.

Containerization. The Docker image packages the full stack: (1) Python 3.11 environment with Claude Agent SDK, Pydantic schemas, and biomedical NLP libraries; (2) Node.js 18 runtime for the Express API server; (3) Nginx 1.25 [15] as reverse proxy and static file server; (4) Pre-built React frontend in `/app/dist`; (5) PDF corpus and extraction outputs in `/app/data` and `/app/outputs`. The image builds for ARM64 architecture (linux/arm64) to target AWS Graviton processors, reducing compute costs compared to x86-based instances.

AWS Infrastructure. The container runs on an EC2 t4g.micro instance (2 vCPU, 1 GB RAM, ARM64) in us-east-1, suitable for demonstration and moderate usage (<100 concurrent users). The Docker image is stored in Elastic Container Registry (ECR) with tag-based versioning, enabling atomic updates via `docker pull` and `docker restart`. An EBS volume (20 GB) attaches to the instance for persistent storage of logs, PDFs, and extraction checkpoints across container restarts.

Model Context Protocol (MCP) Server. The extraction and normalization pipeline is exposed as an MCP server implementing the Anthropic Model Context Protocol (MCP) specification [16]. The server registers all 22 specialized tools (Supplementary Table S15) with JSON Schema parameter definitions, enabling Claude instances or compatible LLM agents to invoke OncoCITE capabilities through natural language requests. The MCP server runs as a persistent process alongside the main application, communicating via stdio transport.

Skills-Based Integration. A skills file encapsulating OncoCITE capabilities is provided for direct integration with Claude-based workflows. The skills definition includes tool descriptions, parameter specifications, and usage examples enabling any Claude instance to perform evidence extraction without infrastructure deployment.

Genomic Pipeline Integration. Wrapper modules for Nextflow and Snakemake enable incorporation of OncoCITE as a pipeline stage. The modules accept variant call files (VCF) as input, extract gene and variant identifiers, query the normalized evidence database, and output annotated evidence items in JSON or TSV format compatible with downstream analysis tools.

Cost and Scalability. Current single-instance deployment demonstrates feasibility for academic research budgets with minimal cloud infrastructure costs (specific pricing examples provided in Supplementary Note S2.5). For production deployment, the architecture supports horizontal scaling: an Application Load Balancer for high availability, S3 with CloudFront CDN for PDF storage, and RDS PostgreSQL for persistent evidence storage enabling complex queries across the full corpus.

References

References

[1] Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, Erica K Barnell, Alex H Wagner, Zachary L Skidmore, Amber Wollam, Connor J Liu, Martin R Jones, Rachel L Bilby, Robert Fenberg, Malik G Richters, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170–174, 2017.

[2] Alex H Wagner, Kilannin Krysiak, Adam C Coffman, Arpad M Danos, Erica K Barnell, Joshua F McMichael, Susanna Kiwala, Nicholas C Spies, Zachary L Skidmore, Cody A Ramirez, Lynzey Kujan, et al. CIViCdb 2022: evolution of an open-access cancer variant interpretation knowledgebase. *Nucleic Acids Research*, 51(D1):D1230–D1241, 2023.

[3] Debyani Chakravarty, Jianjiong Gao, Sarah M Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E Rudolph, Rona Yaeger, Tara Soumerai, Moriah H Nissan, Matthew T Chang, Sarat Chandarlapaty, Tiffany A Traina, Paul K Paik, Alan L Ho, Feras M Hantash, Andrew Grupe, Shrujal S Baxi, Margaret K Callahan, Alexandra Snyder, Ping Chi, Daniel Danila, Mrinal Gounder, James J Harding, Matthew D Hellmann, Gopa Iyer, Yelena Janjigian, Thomas Kaley, Douglas A Levine, Maeve Lowery, Antonio Omuro, Michael A Postow, Dana Rathkopf, Alexander N Shoushtari, Neerav Shukla, Martin Voss, Ederlinda Paraiso, Ahmet Zehir, Michael F Berger, Barry S Taylor, Leonard B Saltz, Gregory J Riely, Marc Ladanyi, David M Hyman, José Baselga, Paul Sabbatini, David B Solit, and Nikolaus Schultz. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, 1:1–16, 2017.

[4] John G Tate, Sally Bamford, Harry C Jubb, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2019.

[5] Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, et al. ClinVar: improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844, 2020.

[6] Sara E Patterson, Rangjiao Liu, Cara M Statz, et al. The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Human Genomics*, 10:4, 2016.

[7] Meta Open Source. React Documentation. Online documentation, 2024. Accessed 2026-01-26.

[8] VoidZero Inc. Vite Documentation. Online documentation, 2024. Accessed 2026-01-26.

[9] OpenJS Foundation. Node.js Documentation. Online documentation, 2024. Accessed 2026-01-26.

[10] OpenJS Foundation. Express: Fast, unopinionated, minimalist web framework for Node.js. Project documentation and source code, 2024. Accessed 2026-01-26.

[11] Mozilla. PDF.js. Project documentation and source code, 2023. Accessed 2026-01-26.

[12] Vasturiano. react-force-graph. Project documentation and source code, 2024. Accessed 2026-01-26.

[13] Bostock, Mike and others. D3.js: Data-Driven Documents. Project documentation and source code, 2024. Accessed 2026-01-26.

[14] Docker, Inc. Docker Documentation. Online documentation, 2024. Accessed 2026-01-26.

[15] F5, Inc. NGINX Documentation. Online documentation, 2024. Accessed 2026-01-26.

[16] Anthropic. Model Context Protocol (MCP) Specification. Online documentation, 2024. Accessed 2026-01-26.