EVISTREAMS: Human-in-the-Loop AI Data Extraction for Systematic Reviews in Medicine*

Sai Karthik Kosuri[1] **Ankita Shashikant Bhosale**[1] **Michael Glick**[1]
Alonso Carrasco-Labra[1] **Chris Callison-Burch**[2]

[1]Center for Integrative Global Oral Health (CIGOH),
Penn Dental Medicine, University of Pennsylvania
[2]Department of Computer and Information Science, University of Pennsylvania

Corresponding author: karthik9@upenn.edu

Abstract

Systematic reviews underpin clinical guidelines, yet their data-extraction step is a major expert-labor bottleneck bound by a protocolized workflow: two reviewers extract each study independently, an adjudicator resolves disagreements, and the team keeps an auditable record of how every value was produced. Large language models can assist with extraction, but that assistance must fit established review protocols and preserve reproducibility. We present EVISTREAMS, a live, open-source, no-code web platform that puts review teams in control of AI-assisted extraction at three key stages: *program design* (a structured decomposition approved before any code runs), *field specification* (typed field definitions calibrated from a pilot), and *extracted predictions* (reviewer-blinded dual review with adjudication). Working through a form builder, a domain expert defines typed fields rather than prompts, runs extraction over uploaded PDFs, inspects every value alongside the supporting passage it came from, and resolves a reviewer-blinded dual review into an auditable consensus export. An evaluation across four clinical corpora and three frontier model families, released with the system, shows that extraction quality is shaped far more by the field specification than by the choice of model. EVISTREAMS is live at https://evistreams.com/demo and released under Apache-2.0.

1 Introduction

Systematic reviews inform clinical and public health practice guidelines, health-technology assessments, and payer and regulatory decisions. Their authority rests on a demanding protocol: two reviewers extract each study independently, an adjudicator resolves disagreements, blinding guards against anchoring, and the team keeps an auditable record of how every value was produced. Data extraction (the transfer of numerical and categorical findings into structured tables) is among the most labor-intensive steps of a systematic review (Higgins et al., 2023). Large language models can fill in much of an extraction form, though how well varies sharply from field to field (Simmons et al., 2025), and an extraction error can be catastrophic: a single mis-extracted sample size or effect estimate can propagate into a pooled estimate and a clinical recommendation. Therefore we are motivated to design a human-in-the-loop system for systematic reviews that lets humans oversee the review, trace each value, and verify the outputs. We investigate three key stages that put humans in control of AI-generated extraction while meeting the evidentiary standards of the discipline.

Our main finding is that extraction quality is governed far less by *which* model reads a paper than by *how* the task is specified. This echoes the growing observation in clinical information extraction that "success may increasingly hinge on the clear articulation of objectives, rather than on singular workflow methodologies" (Hein et al., 2025), and that customized, domain-tuned prompts dominate generic ones (Li et al., 2026), which we elevate into an architectural commitment. We study extraction across four clinical corpora. Choosing the model or changing the extraction pipeline changes performance by only a few F1 points, with mixed direction across corpora. The field specification has the largest effect: a richer, more detailed specification produces better results.

We present EVISTREAMS (short for evidence streams), an open-source, live platform running at https://evistreams.com/demo that puts domain experts in control of AI-assisted extraction through a no-code interface. Human review enters at three key stages: (1) *program*, where the system proposes a plan for extracting each field from the PDF and a reviewer approves, edits, or rejects it before any code runs; (2) *specification*, where each field's definition is tested on a handful of papers

*Live demo: https://evistreams.com/demo

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 759–770
October 24-29, 2026 ©2026 Association for Computational Linguistics

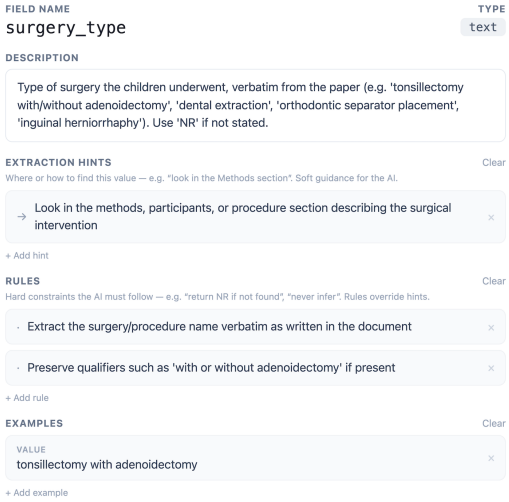

Figure 1: A structured field in the form builder (`surgery_type`): description, extraction hints, rules, and examples, edited directly by the reviewer.

and the reviewer refines it based on what comes back; and (3) *predictions*, where every extracted value is checked by two independent reviewers, and a third resolves any disagreement. Figure 1 shows how the domain expert controls extraction by adding hints, rules, and examples.

Every AI-extracted value is paired with the verbatim quote it was extracted from, which reviewers can inspect. Unlike prior work, which lets users correct only the final output, EVISTREAMS gives many ways to control how the system behaves: from the extraction plan to the field specification to the final adjudication step.

2 The EVISTREAMS System

We designed EVISTREAMS for the ADA Living Guidelines Program, a collaboration between the American Dental Association and our lab, the Center for Integrative Global Oral Health (CIGOH), that keeps oral health guidelines updated as new evidence comes in (Carrasco-Labra et al., 2026). Our team uses EVISTREAMS end to end: a methodologist designs a structured extraction form and approves the pipeline the system proposes for it, then reviewers run extraction over the uploaded PDFs and adjudicate the results, with human review entering at three key stages (Figure 5).

2.1 Extraction pipeline

A methodologist designs a *form*, giving only a name, a description, and optionally an example for each field. The system groups related fields into stages and proposes this grouping as a plan (Figure 2) for the reviewer to approve: never executable code. Once approved, EVISTREAMS[1] adds hints and rules to each field and turns the plan into extraction code automatically: a fixed compiler maps the approved plan to predefined DSPy components; no LLM writes this code. Separately, uploaded PDFs are stored in S3 and converted to Markdown using Datalab (Datalab, 2024).

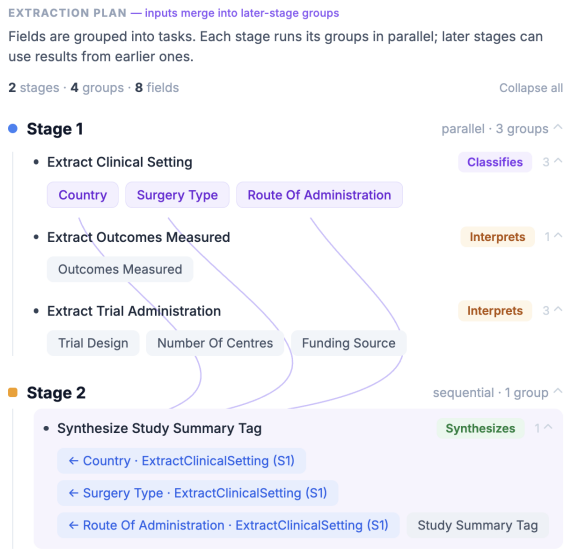

Figure 2: An extraction plan: fields are grouped into tasks, and each stage runs its groups in parallel; later stages can use results from earlier ones. Here Stage 1 extracts clinical setting, outcomes, and trial administration fields in parallel; Stage 2 synthesizes a study summary tag from three of Stage 1's outputs.

Once fields are enriched, each group from the plan becomes a DSPy signature (Khattab et al., 2024): fields in the same group share one Chain-of-Thought signature (Wei et al., 2022), and independent groups run at the same time. A DSPy module then runs each signature to extract the data. We skip DSPy's automatic prompt optimizers (Opsahl-Ong et al., 2024; Agrawal et al., 2026): at our scale (5–20 example papers), we prioritize direct expert editing over automatic prompt optimization (Li et al., 2026). Every value comes back with its source text, in the same format across every model (Figure 3).

2.2 Structured field specification

Each field's specification has four elements (Figure 1, from the introduction): a **description** of

[1] Backend: `https://github.com/karthik-strikes/evistream_backend`. Frontend: `https://github.com/karthik-strikes/evistream_frontend`. Released under Apache-2.0.

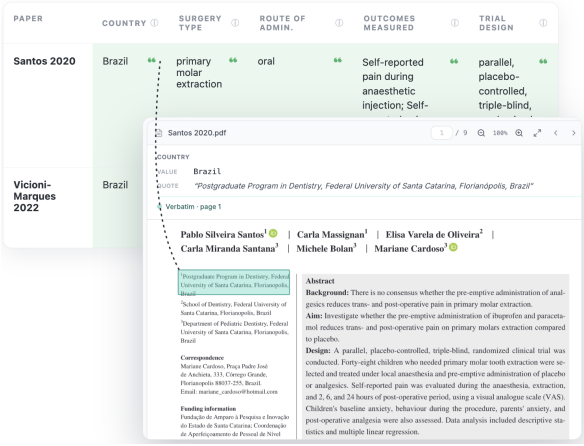

Figure 3: Run: every extracted value can be expanded to its verbatim source quote and page in the original PDF, so a reviewer can verify a value (here `country`: "Brazil") against exactly the sentence it came from.

the concept (always present, user-owned and never rewritten); **hints** locating the evidence within a paper; **rules** imposing hard output constraints (e.g. "return the analyzed N, not the enrolled N"); and **examples** anchoring the expected value shape. Hints and rules are filled in automatically: a LangGraph (LangChain, 2024) chain runs a prompt over the field's description and any user-given examples to generate them, consistent with what the user wrote.

Splitting *what* a field means from *how* to find it lets a methodologist localize an error to just one element. These elements are also what pilot calibration edits later (§2.4); the reviewer refines them after seeing pilot results.

2.3 Table extraction

Medical-study tables record one row per **arm** (the different drugs or interventions a study tests for a condition, e.g., ibuprofen versus an alternative drug), or one row per (arm, follow-up, outcome) triplet. Naive LLM extraction on these tables often truncates the table, drifts in meaning, or hallucinates columns that were never asked for.

We avoid this by decomposing the extraction (Khot et al., 2023): instead of reading a table positionally, we first extract only a few *anchor columns* that name each row (e.g., drug name and dose), then fan out one call per row to fill in the rest. For example, a 10-column table repeated across 400 papers easily loses columns under naive extraction; anchoring each row to its drug and dose first, then filling in the remaining cells per row, is designed to reduce row and column drift, since each call already knows which row it belongs to.

2.4 Human review at three key stages

Program review (§2.1, Figure 2) is implemented as a LangGraph (LangChain, 2024) workflow: the proposed plan is a directed acyclic graph (DAG) of extraction sub-tasks with field-level dependencies, validated mechanically for cycles and for missing or duplicated field assignments (regenerating with structured feedback on failure), and paused at an `interrupt_before` checkpoint until the reviewer approves, edits, or rejects it.

Specification calibration occurs after a form is active: the reviewer tests the form on a handful of papers (typically 3–5), sees the results, and re-edits the fields (Figure 1) based on what comes back. These corrections update the corresponding signatures at runtime: any change to a field's hints, rules, or description is reflected directly in its prompt.

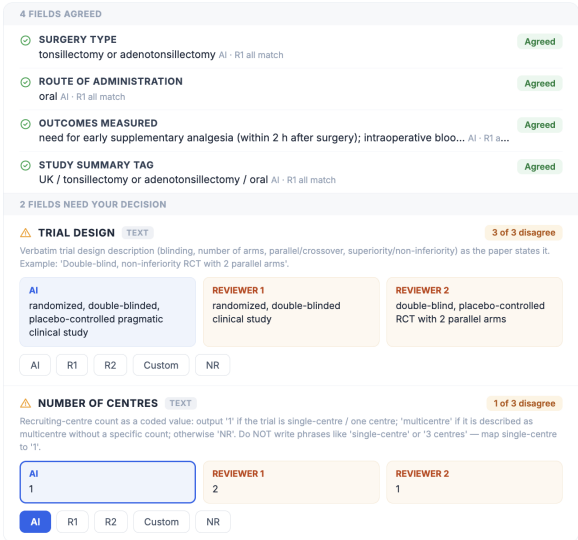

Figure 4: Adjudication: fields where the AI and both reviewers already agree are marked *Agreed*; fields with a disagreement (here `trial_design` and `number_of_centres`) show the AI's and both reviewers' entries side by side for the adjudicator to pick or correct.

Adjudication[2] occurs after extraction: two reviewers independently complete the same form on the same paper under reviewer blinding (each sees the AI pre-fill but not the other's entries). Field-level disagreements are computed on demand, and an adjudicator resolves each conflict (Figure 4). For every field, the system links the AI's pre-fill, both

[2] Screencast ($\leq$2.5 min): https://www.youtube.com/watch?v=j1Gxyu0KitU.

reviewers' entries, and the adjudicator's decision together, so the full history of how a value moved from AI guess to final answer can be looked up later. When a reviewer changes an AI value they attach their own supporting quote, so every value still shows where in the paper it came from: the system keeps the AI's value, the reviewer's replacement, and a pointer into the PDF for each. Reviewer blinding keeps one reviewer's entry from anchoring the other, preserving the independent review that systematic-review protocols require; it does not blind either reviewer to the AI pre-fill itself.

Implementation. EVISTREAMS is a Next.js/FastAPI stack with Celery workers, Redis-backed live WebSocket progress, Supabase (Postgres) storage, Markdown ingestion (Datalab), and a multi-model fallback chain over the three frontier families.

3 Evaluation

We evaluate EVISTREAMS on four clinical datasets, measure the differences among three frontier models, and analyze the errors that occur. The choice of frontier model does not dramatically affect quality: all three perform similarly. The field specification drives the largest variance in quality.

Setup. We score results with a field-typed harness, released publicly alongside EVISTREAMS and implemented independently from the extraction pipeline to reduce evaluation coupling. We use four clinical corpora whose ground truth was drawn, as reported, from completed, peer-reviewed systematic reviews in which two review authors extracted every study independently and resolved disagreements by consensus, the same protocol EVISTREAMS supports: **oral cancer** (three reviews of oral cancer diagnostic adjuncts (Verdugo-Paiva et al., 2026; Bhosale et al., 2026; Urquhart et al., 2026); 43 diagnostic-accuracy studies, four forms, 47 fields), **antibiotic prophylaxis** (CD010266 (Brignardello-Petersen et al., 2015); 10 trials), **periodontitis** (CD004714 (Simpson et al., 2022); 29 trials, outcomes per arm $\times$ time point), and **ibuprofen** (CD015432 (Pessano et al., 2024); 43 trials); all three model families were run on every corpus (Appendix A)[3].

A field's initial definition comes from the review protocol, which sets out the PICO definitions, the eligibility criteria and the outcome definitions before anyone extracts anything, and which the reviewers who produced our reference data worked from as well. Its hints, rules and examples were then refined on pilot papers drawn from the same corpus we later score against.

Each field is scored by a type-matched strategy, and record-level forms are aligned by a globally optimal Hungarian assignment over row-identity keys before scoring, so a missed arm counts against recall and a spurious one is an over-extraction. We report micro- and macro-F1, which agree within $\sim$0.02 on most forms (up to $\sim$0.06 on the hardest oral-cancer forms).

We test the model, the pipeline decomposition, and the field specification one at a time, holding the other two fixed. We score against the reference data, though this ground truth can itself contain reviewer conventions or occasional errors. The free-text judge (Claude Sonnet 4.6) is itself one of the three extraction models, a potential circularity readers should weigh when interpreting free-text scores. That judge decides 34% of scored comparisons; the rest are settled by deterministic numeric or normalised-string matching.

3.1 Production quality and robustness

Table 1 reports overall micro-F1 per corpus for all three frontier families through the identical pipeline and specifications (per-form detail in Appendix B). Quality varies little across the families tested (Claude and Gemini within 0.03 micro-F1 on every corpus, GPT trailing by up to $\sim$0.05 on antibiotic and periodontitis), and is bounded far more by the form and its source text than by the model (the between-corpus range, 0.741–0.898, dwarfs the between-model one). Oral cancer, our most complex corpus (four forms, 47 fields), is consistently hardest across all three models; antibiotic prophylaxis is consistently easiest, tracking form complexity rather than any one model's weakness. With specification and pipeline fixed, the provider can thus be chosen on cost, latency, or

[3]Extraction models: Claude Sonnet 4.6 (Anthropic, 2026), GPT-5.5 (OpenAI, 2026), and Gemini 3.1 Pro (Google DeepMind, 2026) (preview); the free-text judge is Claude Sonnet 4.6. We score every study whose full-text PDF was available: 10 of 11 antibiotic-prophylaxis trials and 29 of 35 periodontitis studies (the rest were excluded for PDF unavailability); ibuprofen is complete at 43 of 43.

Corpus	Studies	Gemini	GPT	Claude
Ibuprofen	43	0.842	0.845	**0.851**
Oral cancer	43	0.741	0.758	**0.763**
Antibiotic	10	0.896	0.851	**0.898**
Periodontitis	29	0.827	0.799	**0.846**

Table 1: Overall micro-F1 per corpus, three frontier model families through the same pipeline with identical form specifications (best per row in bold). Macro- and micro-F1 differ by at most 0.03 at the corpus level; perform macro-F1 detail in Appendix B.

governance rather than accuracy. Collapsing the multi-signature pipeline into a single hand-tuned prompt changes overall F1 by at most 0.036 with mixed direction (the one sizable swing, antibiotic tables, is GPT-specific: $\Delta = -0.096$ vs. -0.005 for Claude and Gemini), so decomposition provides governable structure, not consistent accuracy gains (Appendix D).

3.2 Field specification is the largest observed lever

Across the four calibrated corpora and three tested model families, richer field specifications improved macro-F1 in 40 of 51 model–form comparisons (78%), with a mean ΔF_1 of 0.049 and a median of 0.014 (full per-form ablation in Appendix C). The mean is strongly influenced by the oral-cancer `index_test` form: removing hints, rules, and examples reduced its macro-F1 from 0.724 to 0.247 for Claude, with similarly large declines for Gemini and GPT. Excluding this form, the mean improvement was 0.023, and the full specification remained better in 37 of 48 comparisons (77%). Thus, richer specifications usually provide modest gains but can become decisive for fields requiring disambiguation, source localization, or non-obvious extraction conventions. Model and decomposition effects, reported above, were comparatively small and mixed in direction.

3.3 Error profile

We reviewed and classified every scored discrepancy in 52 studies (all 10 antibiotic trials; 15/43 oral-cancer, 15/43 ibuprofen, 12/29 periodontitis, each of the latter three by set-cover over the lowest-scoring fields): an LLM-assisted first-pass classification, followed by manual verification of every label by an expert systematic-review methodologist. Table 2 summarizes the discrepancy taxonomy and prevalence: of 1,003 discrepancies, two-thirds are

Discrepancy type	Share
Methodological (ground-truth/protocol side)	68%
Genuine model errors (30%)	
Missing record or outcome row	7%
Stated-value omission (`false NR`)	6%
Wrong arm, table, or time point	6%
Cohort or adjacent-column confusion	5%
Denominator or document-context error	4%
Arm or regimen conflation	1%
Uncertain/unadjudicable	2%

Table 2: Taxonomy of 1,003 reviewed discrepancies from an error-enriched sample of 52 studies. Percentages are shares of reviewed discrepancies (rounded; subtypes need not sum to the group total), not extraction error rates. Detailed definitions and causes appear in Appendix E (Table 7).

methodological (ground-truth, convention, ambiguity, or scoring-side), not model failures; under a third are **genuine model errors**, concentrated in dense multi-arm, multi-table reporting; the rest are unadjudicable. Among genuine model errors, omissions and context-assignment mistakes were both common. Missing rows and false-NR omissions accounted for 138 discrepancies, while 160 involved incorrect arms, time points, cohorts, denominators, or regimens. Pilot calibration is intended to reduce specification-driven misses, while dual review and adjudication provide a final check on residual errors. Detailed definitions and causes for each subtype are provided in Appendix E.

4 Demonstration

Demo walkthrough. In the reviewer demo, a methodologist first uploads a batch of PDFs, which EVISTREAMS converts to Markdown via Datalab (Datalab, 2024) in the background while the methodologist keeps working. The methodologist then creates a form, naming each field with a description and an optional example; EVISTREAMS proposes a decomposition that groups related fields into parallel extraction stages (Figure 2), and the methodologist reviews and approves the plan before any code runs.

EVISTREAMS then runs a pilot study, extracting the form over 3–5 papers so the methodologist can check each AI value against the expected value and refine hints, rules, or examples on any field that came back wrong; edits save live and are reflected in the next pilot run.

Once the pilot output matches expectations, the

methodologist runs the approved form over the remaining documents, and every extracted value is shown alongside the supporting passage it was drawn from (Figure 3). Two reviewers then carry out manual extraction: each independently completes the same form on the same paper under reviewer blinding, accepting or correcting the AI pre-fill. EVISTREAMS surfaces their field-level disagreements, and an adjudicator resolves each conflict into the consensus export (Figure 4), which records the AI suggestion, both reviewer entries, the adjudicated value, and inter-rater agreement.

Target audience. Systematic-review teams: clinical experts who extract data, and methodologists who need dual review, adjudication, and inter-rater statistics for Cochrane-style protocols (Higgins et al., 2023). A secondary audience is NLP researchers, for whom the per-field decision chain (AI pre-fill, R1, R2, adjudicator) is a reusable data asset.

5 Related Work

Systems for evidence synthesis. AUTOFOREST (Pronesti et al., 2026) targets a different endpoint (Cochrane forest-plot generation from PDFs) with interactive correction of ICO elements, extracted values, and synthesis outputs. SCHEMATIQ (Levy et al., 2026) discovers a schema from a research question plus a document collection and supports expert revision of the schema and the extracted table; it is our closest general-purpose sibling.

Established review tools (Marshall et al., 2016; Sun et al., 2025) and annotation platforms (Pei et al., 2022; Klie et al., 2018) place the human loop at the prediction level only, correcting outputs rather than the program that produces them.

Table 3 compares the extraction systems directly. Several let a user define the extraction schema, and most let a reviewer correct extracted values, and the commercial platforms carry blinded dual review with adjudication. EVISTREAMS differs in one place outright, review at the program level before any extraction runs, and in combining that with specification calibration and blinded dual review in a single workflow.

Specification, decomposition, and data extraction. Our thesis turns a growing observation (that objective articulation can matter more than model choice (Hein et al., 2025; Li et al., 2026)) into a structured, editable control surface tested across models. Concurrent clinical information-extraction systems have also explored modular or agent-based decomposition (Aal Abdulsalam et al., 2026; Hart and Bergamaschi, 2026).

On the data side, VAXSCOPE (İlgen et al., 2026) benchmarks review-level extraction with a fixed taxonomy, whereas EVISTREAMS exposes user-defined structured fields in a live loop. We adopt a variant of the error taxonomy of Simmons et al. (2025); the per-field extractor is a Chain-of-Thought module (Wei et al., 2022) compiled through DSPy (Khattab et al., 2024), and our anchor-column table extraction (§2.3) is a domain instance of decomposed prompting (Khot et al., 2023; Liu et al., 2025).

We therefore do not claim novelty in pipeline

System	UDS	AIX	SRC	OUT	PRG	CAL	DRA
SCHEMATIQ	●	●	●	●	—	●	—
AUTOFOREST	◗	●	◗	●	—	—	—
ROBOTO2	—	●	●	●	—	—	—
Elicit	●	●	●	●	—	◗	◗
Covidence	●	◗	◗	●	—	—	●
DistillerSR	●	●	●	●	—	—	●
EVISTREAMS	●	●	●	●	●	●	●

Table 3: Extraction-assistance capabilities across systems: SCHEMATIQ (Levy et al., 2026), AUTOFOREST (Pronesti et al., 2026), ROBOTO2 (Hevia et al., 2025), and the commercial platforms Elicit, Covidence and DistillerSR, whose capabilities were verified against their official documentation (Elicit, 2026; Covidence, 2026a,b; DistillerSR, 2026, 2016); TRIALSTREAMER (Nye et al., 2020) and CLINICALTRIALSHUB (Park et al., 2026) extract fixed elements rather than a review-defined form and are not compared here. **UDS**: user-defined schema; **AIX**: AI-generated pre-fill; **SRC**: value-level source grounding (each value linked to its verbatim passage); **OUT**: reviewer correction of extracted outputs; **PRG**: program-level review (extraction plan approved before it runs); **CAL**: specification calibration loop (pilot results → spec edits with immediate effect); **DRA**: reviewer-blinded dual review with adjudication carried to the export. ● supported; ◗ partial (AUTOFOREST: control over ICO elements only, and free-text extraction rationales without links to source locations; Covidence: field suggestions rather than form pre-fill, and a supporting quote per suggestion without a link to its location; Elicit: iteration over free-text column prompts rather than structured field elements, and dual review described for screening rather than extraction); — not described in the cited work or documentation. Systems target different endpoints and corpora, and no shared extraction benchmark exists; we therefore compare capabilities, and release our evaluation harness and four corpora to enable future head-to-head comparison.

decomposition or source-grounding in themselves (both appear in the systems above), but in unifying human review over a structured, editable specification, and in showing empirically that, across our corpora, the field specification is the largest and most consistent control surface for quality.

6 Conclusion

EVISTREAMS is a human-in-the-loop platform for conducting systematic reviews via human–AI collaboration. It is made up of several parts: the extraction *program* (approved by users before it runs), the *specification* (iteratively refined by letting the user run test questions), and the *extractions* (validated through reviewer-blinded dual review). Our evaluation shows, across four clinical corpora and three frontier model families, that extraction quality is robust to model and pipeline decomposition, and that the structured field specification drives the biggest variance in quality. Expert effort is therefore best spent on the specification, with adjudication the second-most-useful place for human intervention to catch residual errors. Because a field's specification can be edited and take effect immediately, without regenerating the pipeline, the lever our evaluation identifies as most valuable is also the cheapest one to pull. An extraction workflow that is expert-controlled and auditable is intended to reduce transcription work and supports systematic-review workflows.

Limitations

Human agreement. Inter-annotator agreement was not measured when the reference data was created (§3).

Specification calibration. We refined specifications using pilot papers from the same reviews used for evaluation. Our results therefore measure the benefit of review-specific, calibrated specifications. Performance when specifications are developed for a new review without prior calibration remains untested.

LLM-judge overlap. The free-text judge shares a provider family with one extraction model (§3), creating potential evaluator dependence that we do not quantify.

Corpus scope. We evaluated on four corpora totalling 125 studies. All four come from completed, peer-reviewed systematic reviews, and we chose them to differ in the kind of question they ask: diagnostic accuracy, surgical prophylaxis, treatment of a chronic disease, and relief of post-operative pain. We are extending the evaluation to reviews from further areas to see whether the field specification is still the biggest lever there.

Ethics Statement

EVISTREAMS is a research tool for evidence synthesis, not a medical device, and its outputs are not clinical advice; this work adheres to the ACL Code of Ethics and the ACM Code of Ethics and Professional Conduct. Its central design property is that no AI-extracted value enters a final dataset without human review: the dual-review and adjudication workflow (§2) enforces verification structurally rather than relying on user diligence.

The evaluation corpora are published, peer-reviewed articles; no patient-level or identifiable data is processed. Users upload PDFs they are licensed to access; uploaded text is sent to commercial LLM APIs (Anthropic, OpenAI, Google) subject to those providers' data-handling terms. The hosted instance restricts project data to invited team members via role-based access control.

Automated extraction errors could propagate into clinical guidance if left unchecked; we report the error analysis openly (§3.3), distinguishing genuine model errors from ground-truth and annotation-convention discrepancies, and the per-field decision chain records every step from AI suggestion to human-approved value.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300) and is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 56000026C0019 (the BENGAL program). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Abdulrahman Aal Abdulsalam, Adhari Al Zaabi, Riham Jeeballah, and Habiba El Keraby. 2026. A multi-agent open-source LLM for structured cancer registry information extraction from pathology and medical reports. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP)*, pages 531–551. Association for Computational Linguistics.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2026. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations (ICLR)*.

Anthropic. 2026. Introducing Claude Sonnet 4.6.

Ankita Shashikant Bhosale, Francisca Verdugo-Paiva, Olivia Urquhart, Carolina Martins-Pfeifer, Matthias Wallach Weinstein, Tanya Walsh, Kelly K. O'Brien, Ana María Rojas-Gómez, Michael Glick, Mark W. Lingen, and Alonso Carrasco-Labra. 2026. Vital staining adjuncts to determine the need for biopsy for early detection of oral squamous cell carcinoma and potentially malignant disorders: An evidence summary of a living systematic review, version 2026 1.0. *The Journal of the American Dental Association*, 157(6):588–601.

Romina Brignardello-Petersen, Alonso Carrasco-Labra, Ignacio Araya, Nicolas Yanine, Luis Cordova Jara, and Julio Villanueva. 2015. Antibiotic prophylaxis for preventing infectious complications in orthognathic surgery. *Cochrane Database of Systematic Reviews*, (1):CD010266.

Alonso Carrasco-Labra, Olivia Urquhart, Ankita Shashikant Bhosale, Francisca Verdugo-Paiva, Carolina Castro Martins-Pfeifer, Krishna Aravamudhan, Marko Vujicic, Tim Wright, and Michael Glick. 2026. American Dental Association living evidence-informed guidelines: shaping the future of clinical and public oral health decision making. *The Journal of the American Dental Association*, 157(3):213–217.

Covidence. 2026a. AI feature: Extraction suggestions. https://support.covidence.org/help/ai-feature-extraction-suggestions. Accessed 30 August 2026.

Covidence. 2026b. How to do comparison and consensus in extraction 1. https://support.covidence.org/help/comparison-and-consensus. Accessed 30 August 2026.

Datalab. 2024. Datalab: Document intelligence API. https://www.datalab.to.

DistillerSR. 2016. Data extraction: Weighing your options. https://www.distillersr.com/resources/blog/data-extraction-weighing-your-options. Published 14 July 2016; accessed 30 August 2026.

DistillerSR. 2026. Purpose-built GenAI for literature reviews. https://www.distillersr.com/resources/guides-white-papers/purpose-built

-genai-for-literature-reviews. Accessed 30 August 2026.

Elicit. 2026. Systematic literature reviews. https://elicit.com/solutions/literature-review. Accessed 30 August 2026.

Google DeepMind. 2026. Gemini 3.1 Pro model card.

Steven N. Hart and Teya S. Bergamaschi. 2026. Agent-based large language model system for extracting structured data from breast cancer synoptic reports: a dual-validation study. *JAMIA Open*, 9(1):ooag016.

David Hein, Alana Christie, Michael Holcomb, Bingqing Xie, AJ Jain, Joseph Vento, Neil Rakheja, Ameer Hamza Shakur, Scott Christley, Lindsay G Cowell, James Brugarolas, Andrew R Jamieson, and Payal Kapur. 2025. Iterative refinement and goal articulation to optimize large language models for clinical information extraction. *npj Digital Medicine*, 8:301.

Anthony Hevia, Sanjana Chintalapati, Veronica Ka Wai Lai, Nguyen Thanh Tam, Wai-Tat Wong, Terry P Klassen, and Lucy Lu Wang. 2025. ROBOTO2: An interactive system and dataset for LLM-assisted clinical trial risk of bias assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–25, Suzhou, China. Association for Computational Linguistics.

Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch, editors. 2023. *Cochrane Handbook for Systematic Reviews of Interventions*, version 6.4 (updated August 2023) edition. Cochrane.

Bahar İlgen, Ebenezer Awotoro, and Georges Hattab. 2026. VaxScope: Document-level structured evidence extraction from immunization systematic reviews. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP)*, pages 853–863. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into self-improving pipelines. In *International Conference on Learning Representations (ICLR)*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *International Conference on Learning Representations (ICLR)*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

LangChain. 2024. LangGraph: A framework for building stateful, multi-actor applications with LLMs. https://langchain-ai.github.io/langgraph/.

Shahar Levy, Eliya Habba, Reshef Mintz, Barak Raveh, Renana Keydar, and Gabriel Stanovsky. 2026. ScheMatiQ: From research question to structured data through interactive schema discovery. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 220–230. Association for Computational Linguistics.

Lingbo Li, Anuradha Mathrani, and Teo Susnjak. 2026. What level of automation is "good enough"? A benchmark of large language models for meta-analysis data extraction. *Research Synthesis Methods*, 17(4):671–692.

Shuodi Liu, Yingzhuo Liu, Zi Wang, Yusheng Wang, Huijia Wu, Liuyu Xiang, and Zhaofeng He. 2025. Select-then-decompose: From empirical analysis to adaptive selection strategy for task decomposition in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5454–5477. Association for Computational Linguistics.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201.

Benjamin Nye, Ani Nenkova, Iain Marshall, and Byron C. Wallace. 2020. Trialstreamer: Mapping and browsing medical evidence in real-time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 63–69. Association for Computational Linguistics.

OpenAI. 2026. Introducing GPT-5.5.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366. Association for Computational Linguistics.

Jiwoo Park, Ruoqi Liu, Avani Jagdale, Andrew Srisuwananukorn, Jing Zhao, Lang Li, Ping Zhang, and Sachin Kumar. 2026. ClinicalTrialsHub: Bridging registries and literature for comprehensive clinical trial access. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Rabat, Morocco. Association for Computational Linguistics.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337. Association for Computational Linguistics.

Sara Pessano, Natasha R. Gloeck, Luca Tancredi, Martin Ringsten, Ameer Hohlfeld, Sumayyah Ebrahim, Martina Albertella, Tamara Kredo, and Matteo Bruschettini. 2024. Ibuprofen for acute postoperative pain in children. *Cochrane Database of Systematic Reviews*, (1):CD015432.

Massimiliano Pronesti, Angelo Miculescu, Mohsin Kapdi, Paul Flanagan, Oisín Redmond, Joao Bettencourt-Silva, Gurdeep S. Mannu, Spiros Denaxas, Rui Bebiano Da Providencia E Costa, Anya Belz, and Yufang Hou. 2026. AutoForest: Automatically generating forest plots from biomedical studies with end-to-end evidence extraction and synthesis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 128–137. Association for Computational Linguistics.

Zalaya Simmons, Beti Evans, Tamsyn Harris, Harry Woolnough, Lauren Dunn, Jonathon Fuller, Kerry Cella, and Daphne Duval. 2025. Assessing the feasibility and acceptability of a bespoke large language model pipeline to extract data from different study designs for public health evidence reviews. *Cochrane Evidence Synthesis and Methods*, 3(6):e70061.

Terry C. Simpson, Janet E. Clarkson, Helen V. Worthington, Laura MacDonald, Jo C. Weldon, Ian Needleman, Zipporah Iheozor-Ejiofor, Sarah H. Wild, Ambrina Qureshi, Andrew Walker, Veena A. Patel, Dwayne Boyers, and Joshua Twigg. 2022. Treatment of periodontitis for glycaemic control in people with diabetes mellitus. *Cochrane Database of Systematic Reviews*, (4):CD004714.

Qiang Sun, Sirui Li, Tingting Bi, Du Q Huynh, Mark Reynolds, Yuanyi Luo, and Wei Liu. 2025. DocSpiral: A platform for integrated assistive document annotation through human-in-the-spiral. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 267–274. Association for Computational Linguistics.

Olivia Urquhart, Francisca Verdugo-Paiva, Carolina Martins-Pfeifer, Ankita Shashikant Bhosale, Carla Massignan, Dalia Meisha, Kelly K. O'Brien, Ana María Rojas-Gómez, Michael Glick, Mark W. Lingen, and Alonso Carrasco-Labra. 2026. Light-based adjuncts to determine the need for biopsy for early detection of oral squamous cell carcinoma and potentially malignant disorders: An evidence summary of a living systematic review, version 2026 1.0. *The Journal of the American Dental Association*.

Francisca Verdugo-Paiva, Carolina Martins-Pfeifer, Ankita Shashikant Bhosale, Olivia Urquhart, Tanya Walsh, Mallika Balagopal, Mehrnaz Zakershahrak, Kelly K. O'Brien, Camíla Ávila-Oliver, Michael Glick, Mark W. Lingen, and Alonso Carrasco-Labra. 2026. Cytology adjuncts to determine the need for biopsy for early detection of oral squamous cell carcinoma and potentially malignant disorders: An evidence summary of a living systematic review, version 2026 1.0. *The Journal of the American Dental Association*, 157(3):235–246.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Ablation and Model Coverage

All three model families (Claude, GPT, Gemini) were run through the production pipeline, the single-prompt arm (decomposition ablation), and the description-only arm (specification ablation) on all four corpora.

B Per-Corpus Production Results

Table 4 gives the per-form macro-F1 for all four corpora behind the main-text corpus-level micro-F1 (Table 1). Interventions and outcomes are record-level (one row per arm, or per arm × time point).

Corpus	Form	Gemini	GPT	Claude
Ibuprofen	study_char.	0.790	0.827	**0.829**
	patient_pop.	0.920	**0.938**	0.933
	interventions	0.892	0.900	**0.915**
	continuous_out.	**0.795**	0.772	0.765
	dichotomous_out.	0.774	0.754	**0.792**
	Overall	0.834	0.838	**0.847**
Oral cancer	study_char.	0.721	**0.736**	0.730
	patient_pop.	0.788	0.788	**0.819**
	reference_std.	0.682	0.687	**0.695**
	index_test	0.679	0.721	**0.724**
	Overall	0.718	0.733	**0.742**
Antibiotic	study_char.	**0.921**	0.907	0.919
	patient_pop.	0.862	**0.880**	**0.880**
	interventions	0.920	0.853	**0.922**
	dichotomous_out.	**0.874**	0.746	0.862
	Overall	0.894	0.847	**0.896**
Periodontitis	study_char.	**0.878**	0.802	0.859
	patient_pop.	0.777	0.787	**0.815**
	interventions	0.819	0.790	**0.881**
	continuous_out.	**0.830**	0.807	0.822
	Overall	0.826	0.797	**0.844**

Table 4: Per-form macro-F1, three model families through the same pipeline with identical form specifications, across all four corpora (best per row in bold; ties bolded on both). Corpus-level micro-F1 is in Table 1. Periodontitis is scored on the canonical 29-field set (OPTIONAL and structural row-matching fields excluded).

C Specification Ablation, Full Detail

Table 5 gives the per-form specification ablation (ΔF_1 = full − description-only) for all three model families, the finding summarized in §3.2. Full-pipeline per-form macro-F1 is in Table 4 (Appendix B). The oral-cancer index_test recall collapse discussed in the main text reproduces across all three models (Claude +0.476, Gemini +0.489, GPT +0.433).

Corpus	Form	Cld.	Gem.	GPT
Antibiotic	study_char.	+0.012	+0.029	+0.014
	patient_pop.	0.000	−0.018	+0.031
	interventions	+0.029	+0.026	+0.005
	dichotomous_out.	+0.014	+0.037	+0.092
Periodont.	study_char.	−0.027	+0.013	+0.036
	patient_pop.	+0.014	+0.002	+0.008
	interventions	+0.065	+0.116	+0.307
	continuous_out.	+0.004	+0.022	+0.058
Ibuprofen	study_char.	−0.011	−0.030	+0.010
	patient_pop.	−0.001	+0.002	+0.010
	interventions	−0.014	+0.041	+0.018
	continuous_out.	−0.033	+0.001	+0.006
	dichotomous_out.	+0.044	0.000	−0.050
Oral canc.	study_char.	−0.010	+0.015	+0.018
	patient_pop.	+0.048	+0.029	+0.005
	reference_std.	+0.049	+0.015	+0.071
	index_test	**+0.476**	**+0.489**	**+0.433**

Table 5: Specification ablation, ΔF_1 = full − description-only, macro-F1, per form across all three model families. Positive favors the full specification; bold marks the oral-cancer index_test collapse, reproduced by all three models.

D Decomposition Ablation

Table 6 gives the full corpus-by-modality breakdown (table vs. scalar forms, averaged over all three model families) behind the robustness finding in §3.1.

Corpus	Table ΔF_1	Scalar ΔF_1
Oral cancer	+0.027	+0.003
Antibiotic	−0.036	−0.012
Periodontitis	−0.002	−0.012
Ibuprofen	0.000	+0.003

Table 6: Decomposition ablation: overall ΔF_1 (full multi-signature pipeline minus single-prompt arm), split by table and scalar forms and averaged over Claude, GPT, and Gemini. Positive favors the full pipeline.

E Detailed Error Analysis

Table 7 expands the taxonomy in Table 2, giving for each discrepancy type what it is and why it occurs.

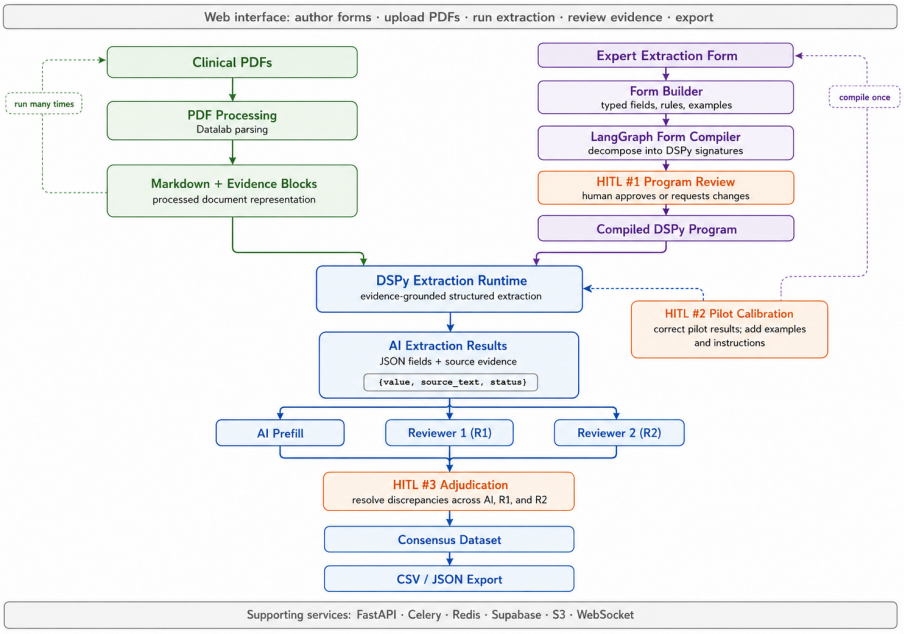

Figure 5: The EVISTREAMS pipeline. Human-in-the-loop (HITL) review enters at three key stages: the extraction *program* (HITL #1), its *specification* (HITL #2), and its *predictions* (HITL #3).

Discrepancy type	What it is	Why it occurs
Methodological or evaluation-side *683/1003 (68%)*	Four patterns: the ground-truth cell is wrong, blank, or falsely coded NR (27%); a report-silent reviewer convention (back-calculated SD, converted dose, figure read off a graph) vs. the literal value (24%); a field with two defensible readings and no multi-arm tie-break (11%); or a surface-form mismatch, e.g. explained vs. bare NR (6%).	Ground truth carries transcription slips and undocumented conventions the extraction is deliberately not allowed to infer; specifications do not disambiguate every multi-arm edge case; and scoring does not normalize every equivalent string.
Genuine model errors 298/1003 (30%)		
Missing record or outcome row *75/1003 (7%)*	A whole row that should appear in a multi-arm outcome table (a subgroup, a secondary time point, or a prose-only outcome) is absent, not merely mis-valued.	Dense multi-arm, multi-time-point tables (and prose-only outcomes) are harder for row enumeration than single-value fields.
Stated-value omission (false NR) *63/1003 (6%)*	NR was coded for an event count or arm size stated plainly in the report text or a results-table cell, needing no calculation.	The value sits inside dense multi-arm results prose or a table cell the extraction did not check before defaulting to NR.
Wrong arm, table, or time point *60/1003 (6%)*	A value was produced but taken from the wrong arm, wrong table, or wrong time-point bin (or a plain transcription slip).	Multiple similar-looking tables, subgroups, and time points sit close together in these results sections.
Cohort or adjacent-column confusion *49/1003 (5%)*	A value taken from the whole-trial or a nested subcohort N instead of the arm's analysed N, or from an adjacent time-point column.	Multi-stage randomization and adjacent per-time-point columns in dense baseline tables invite cohort and column mix-ups.
Denominator or document-context error *37/1003 (4%)*	A stated value was dropped, the wrong patient/lesion denominator was assigned, or generic Discussion text was reported as the study's own Methods.	Diagnostic-accuracy papers report several similar counts (received vs. analysed, per-arm vs. per-technique); citation-heavy Discussion can read like procedure.
Arm or regimen conflation *14/1003 (1%)*	The wrong arm's dose, regimen, or label was copied onto a cell (e.g. the placebo arm described with the active arm's schedule), or a stated fact was dropped.	Multi-arm drug-versus-placebo and short- versus long-course dosing descriptions sit close together in dense prose.
Uncertain/unadjudicable *22/1003 (2%)*	A disagreement the reviewer could not confidently attribute to either a genuine model error or a methodological/ground-truth issue from the paper text alone.	The source paper does not contain enough information to adjudicate the disagreement either way, so it is left unclassified rather than forced into a category.

Table 7: Detailed error definitions and causes for the discrepancy types in Table 2 (counts and shares are given there).