# REAL-WORLD EVALUATION OF AN AI AGENT DRAFTING TRANSLATIONAL IMPACT SUMMARIES

A PREPRINT

Mohammad Arvan*, Amber E. Osterholt, Bailee Rue, Yuvaneswaren Ramakrishnan Sureshbabu, Krishna Riteshkumar Patel, Rebecca T. Feinstein, Bethany C. Bray, Niranjan S. Karnik

University of Illinois Chicago, Chicago, IL, USA

**ABSTRACT**

**Introduction.** Clinical and Translational Science Award (CTSA) programs must document their scholars' research impact, but assembling each scholar's record by hand takes staff an estimated 15 hours and does not scale to a full cohort. An artificial intelligence (AI) agent could serve as a tool to gather scholar data across platforms and disciplines.

**Methods.** We built a human-in-the-loop AI agent that assembles a dossier of sourced evidence for each scholar and drafts one-sentence Translational Science Benefits Model (TSBM) impact summaries for staff review. We evaluated it in the impact-reporting workflow of one CTSA hub across 10 career-development (KL2/K12) scholars. Two evaluation staff independently coded all 507 findings as accept, edit, or reject; the primary measure was the unanimous usable rate, defined as the share both accepted or edited.

**Results.** Both reviewers accepted or edited 81.7% of the agent's findings. Reviewers each spent a median of 14 minutes per scholar, replacing an estimated 15 hours of manual assembly. Inter-rater agreement was moderate (Cohen's kappa 0.43 on the usable-versus-reject decision). A profile discovery study found the agent's recall close to human search. The agent's impact evidence spanned all four TSBM domains, and about a third of the reviewed findings fell in non-scholarly categories that routine processes tend to miss. Reviewers rated synthesis accuracy 4.5 and usefulness 4.8 on a 5-point scale.

**Conclusions.** A human-in-the-loop AI agent can serve as the first-pass author of a scholar's impact record, shifting staff from collecting and writing to reviewing, and making cohort-scale impact reporting feasible.

***Keywords*** human-AI collaboration; generative AI agent; large language models; workflow integration; research impact assessment

## 1 Introduction

National Institutes of Health (NIH) career-development (K) awards support the move from trainee to independent investigator, providing protected time, mentorship, and training [1]. K awardees reach independent (R01-equivalent) funding at higher rates than their non-K peers [2]. The Clinical and Translational Science Award (CTSA) program extends that support to the translational science workforce through institutional career-development funding, awarded first under the KL2 mechanism and more recently as a K12 program [3]. Unlike individual K awards, these mechanisms are institutional grants through which a single hub supports a cohort of scholars.

CTSA KL2 scholars go on to productive translational careers. About 78% remain at CTSA institutions after their award [4]. At five years, they hold a publication advantage over other mentored K awardees [5], and their work is cited about twice as often as the typical NIH-funded study [6]. Policy documents indexed by Overton have also cited more than 3,600 KL2-scholar publications since 2006 [6]. Yet a scholar's impact goes beyond publications and grants, and no routine tracking follows the rest.

*Corresponding author: marvan3@uic.edu

Assessing whether a scholar becomes an independent investigator and moves research toward real-world benefit requires a complete record. Progress toward independence leaves an indexed trail of grants, publications, and promotions that tools such as FlightTracker already ingest [7]. Real-world impact leaves no such trail and takes many forms. The Translational Science Benefits Model (TSBM) groups this impact into clinical, community, economic, and policy benefits [8]. Non-scholarly evidence such as clinical-program leadership, community engagement, cost or adoption data, and policy uptake sits outside every system that current tools draw on.

Keeping that record up to date manually does not scale. Scholars have the most direct knowledge of their own outcomes, but response rates to information requests are low (11.9% within our hub) and decline over time. Scholars also move institutions, within and beyond the CTSA network, and each move makes the scholar harder to track and reach. The non-scholarly evidence adds open-web searching for every scholar, on top of the funder, bibliographic, and institutional systems staff already query. What surfaces must then be written into one-sentence TSBM impact summaries, each stating a finding's translational benefit. Writing them involves reading across publications, grants, and all four TSBM domains, and no tool performs that synthesis. The effort grows with every new cohort, because tracking continues for years after each award ends.

An artificial intelligence (AI) agent can search and draft material that is usually gathered through manual searches and scholar information requests. The agent searches databases and the open web, follows scholars who move, and drafts each summary based on the collected evidence. None of this depends on a scholar answering a request. Yet the promise often outruns the demonstration. Few AI systems are evaluated inside the workflows they serve, so clear outcomes there are rarely demonstrated [9, 10]. A first draft helps only if the staff who own the reporting can verify and correct it with less effort than it would take to assemble the record themselves. Whether that holds is a property of the workflow rather than the model alone.

We built such an agent and evaluated it inside one CTSA hub's impact-reporting workflow across a cohort of 10 KL2/K12 scholars. The agent starts with the hub's own records, assembles a dossier of sourced evidence for each scholar, and drafts a one-sentence TSBM impact summary for each impact finding. Staff review every finding before it enters the record. In the evaluation, two staff members independently reviewed everything the agent proposed. A *profile discovery study* also checked the agent's retrieval against human search. We report what the staff kept, what they corrected, and how long the review took, the outcomes the workflow cares about. If that division of labor holds, staff effort shifts from an estimated 15 hours of assembling each scholar's record to minutes of reviewing it.

## 2 Materials and Methods

For each scholar, our artificial intelligence (AI) agent assembles an *evidence dossier* and drafts impact summaries from it. Together, the dossier and its summaries form the first draft of the scholar's impact record. The dossier is a structured collection of candidate findings, each tied to a source and gathered by searching scholarly databases and the open web. It covers several areas, including the scholar's identity, career, publications, grants, and any mentions in the news or media. Using the evidence dossier, the agent surfaces non-scholarly impact evidence, such as clinical-program leadership, community engagement, cost or adoption data, and policy citations. Each impact finding also includes an *impact summary*, a single sentence that states its translational benefit, framed by the Translational Science Benefits Model (TSBM) [8].

The agent and a reviewer form a human-AI collaboration, a division of labor in a real workflow (Figure 1). The agent does the two costly tasks for the reviewer: gathering and drafting. For each finding, the reviewer accepts, edits, or rejects it, and can add findings that the agent missed. Accepting or rejecting is quick, editing takes more effort, and adding is the most demanding. The agent is therefore tuned for recall. It proposes generously, so the reviewer rejects the excess but rarely has to add. For each finding type, the agent cites a preferred authoritative source, so verification starts from that source.

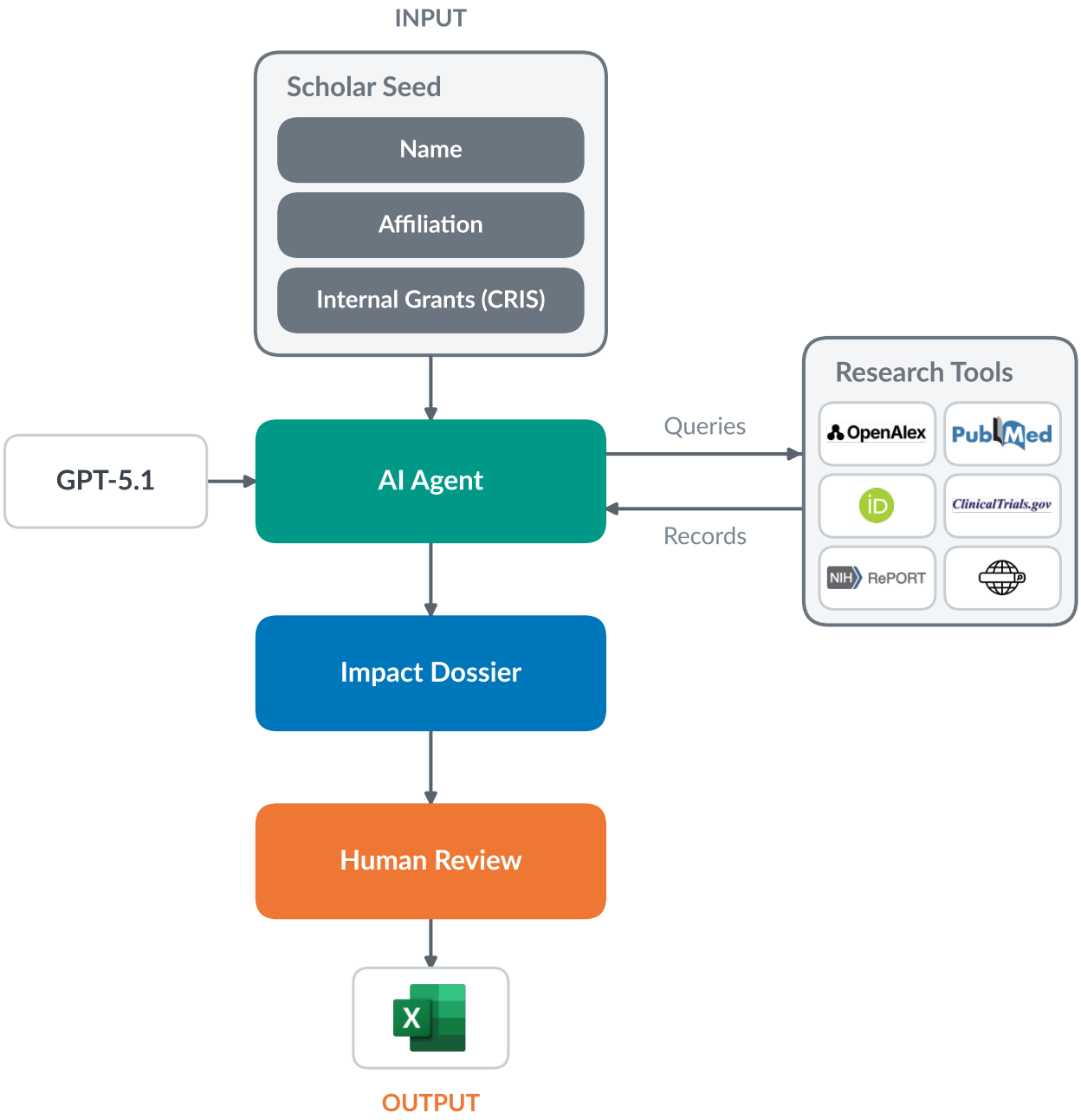


**Figure 1. How the agent gathers evidence.** From a *scholar seed* (the name, institution, and any grant records the hub already holds), the agent queries public scholarly and grant databases (OpenAlex, PubMed, ORCID, ClinicalTrials.gov, and the National Institutes of Health (NIH) RePORTER) and the open web, reading records back to assemble an impact dossier of evidence and one-sentence benefit statements. A reviewer then accepts, edits, rejects, or adds each finding.

We evaluated the agent in two studies at the University of Illinois Chicago's Center for Clinical and Translational Science, a Clinical and Translational Science Award (CTSA) hub. The first, the *profile discovery study*, checks the agent's retrieval on its own, comparing its profile recall against *human-only* and *AI-guided* search. The second, the *review study*, checks the agent in use, with staff working from its drafts in their real reporting workflow, and scores their review by *usable rate* and effort. Both studies drew on the same 10 KL2/K12 scholars. In the *review study*, both reviewers independently scored every finding to measure their agreement.

### 2.1 Agent Design

An agent is a system in which a language model dynamically directs its own process and tool use, maintaining control over how it accomplishes its task [11]. The agent decides which searches to run, which tools to call, and when its research is complete. The dynamic form fits this task because impact discovery is open-ended, with each scholar's record calling for different searches.

The agent starts from a *scholar seed* containing the name, institution, and any grant records the hub already holds. It organizes its work as a *research run* (Figure 2), repeated up to three times. A run has three stages. The *gather* stage collects the evidence, the *assemble* stage writes it into dossier sections and impact summaries, and the *critique* stage flags what to repair. The agent returns the first draft of the impact record, a structured dossier plus an impact summary for each impact finding.

We cap the *gather* stage at 25 turns, each planning the next objective and searching for it; the stage ends earlier when the agent finishes its coverage. It queries five research databases directly through their data interfaces, namely OpenAlex, PubMed, ORCID, ClinicalTrials.gov, and NIH RePORTER. It routes to open-web search for evidence those databases do not hold, such as local media and policy citations. To finish, the agent files a *coverage report* marking every part of the dossier found, confirmed absent, or searched and empty; the report is the agent's own account of its coverage.

A two-model design processes what the *gather* stage retrieves. A large model (gpt-5.1) plans the searches, reasons over the evidence, and drafts the impact summaries. A small model (gpt-5.4-mini) handles the two high-volume reading subtasks. It distills each retrieved web page to the facts the agent asked for, and it trims each publication abstract to its outcome. From every other tool result, the system keeps only the fields the dossier needs. Only this trimmed content enters the large model's context. The agent also manages its own context, condensing accumulated findings into the dossier as it works and keeping the context small even for a scholar with a long record.

The distillation also limits context poisoning, since adversarial text on a page stays confined to the small model, a mitigation against indirect prompt injection [12]. Every stage that reads untrusted content fences it as data, following the spotlighting family of defenses [13].

We defined what counts as an authoritative source for each finding type. For a grant, the preferred source is the hub's own record or the NIH RePORTER project page for NIH awards. For any finding grounded in a paper, it is the paper's digital object identifier (DOI). For a trial, it is the ClinicalTrials.gov record. For a profile or media mention, it is the scholar's own page or a news outlet's own coverage. These source preferences guide every stage of the pipeline.

The *assemble* stage then turns the gathered evidence into the dossier, one section at a time. It writes the four impact sections last, so each impact can cite evidence already in the dossier. For each impact finding, it then drafts a one-sentence impact summary from that evidence. The drafting prompt includes few-shot examples [14] that separate the impact from the publication describing it.

An independent *critique* closes each run. It reads the assembled dossier and the run's action log, and flags gaps, duplicates, contradictions, and weak sources. The action log lets it tell a genuine gap from a search that already came back empty. The agent repairs the flagged issues in the next *research run*. The repeated runs follow an evaluator-optimizer-style pattern for agentic systems [11]; the *critique* is adversarial, so runs typically use the full budget rather than converging early.

A *register-revise* stage measures each drafted summary against a linguistic signature built from 76 source-grounded, human-written benefit statements drawn from published CTSA impact profiles and the TSBM literature [8]. It revises the drafts that drift from this signature, using interpretable features such as sentence length to guide the edit. We cap the stage at three turns.

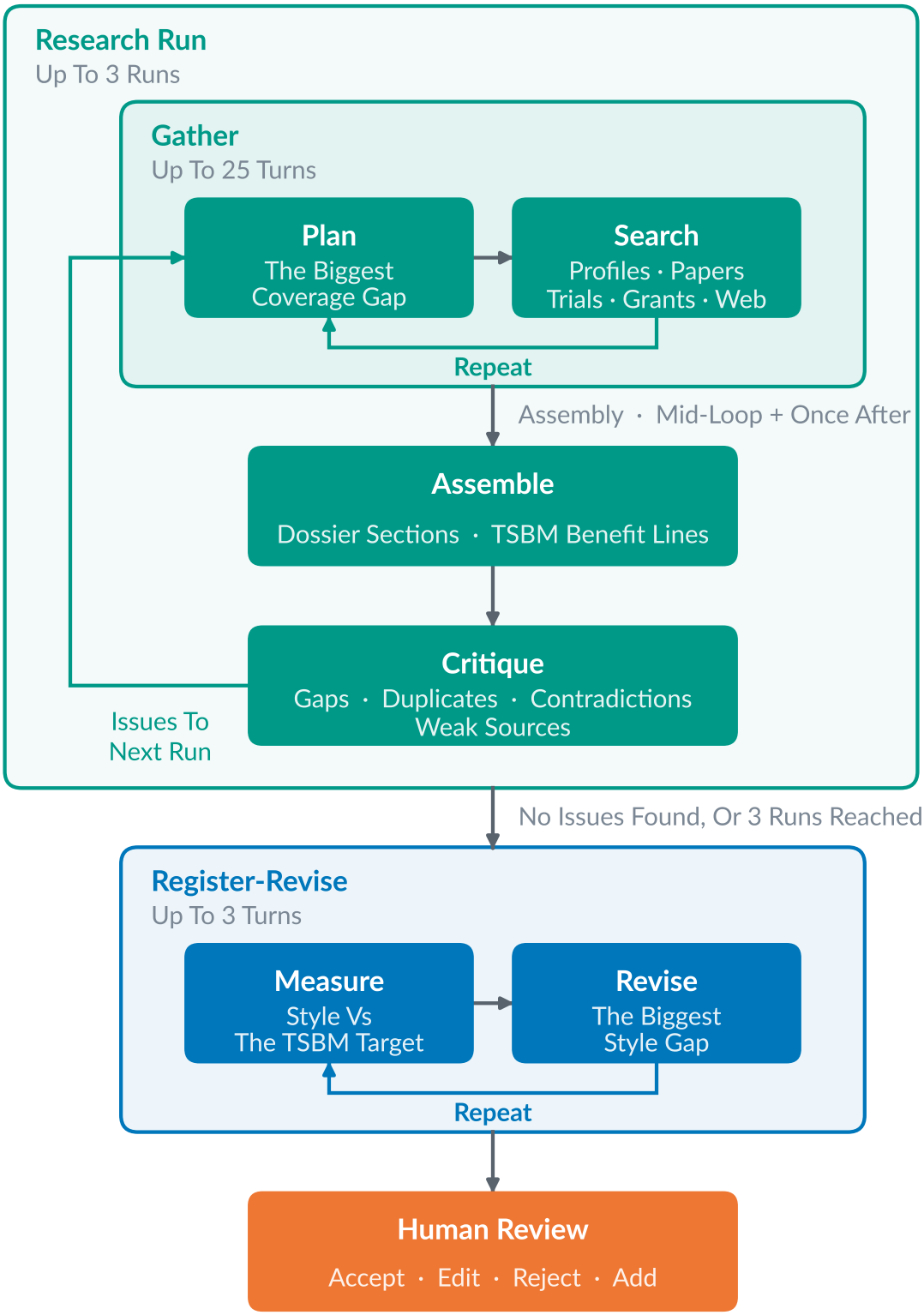


**Figure 2. The agent's nested stages.** The *gather* stage alternates between planning the next objective and searching five research databases and the web (up to 25 turns). The agent assembles the dossier from the running transcript, and an isolated *critique* flags gaps, duplicates, contradictions, and weak sources. The run repeats up to three times. A *register-revise* stage (up to three turns) then tightens the drafted one-sentence impact summaries toward the Translational Science Benefits Model (TSBM) target style before a human accepts, edits, rejects, or adds.

### 2.2 Profile Discovery Study

We evaluated discovery on profiles, the pages that identify a scholar (institutional homepage, Google Scholar, ORCID, LinkedIn). Profiles are foundational, because the profile set fixes which scholar the dossier describes. A profile error cascades to every publication, grant, impact, and summary. They are also the easiest to verify, since a reviewer confirms a canonical page at a glance. Two non-overlapping five-scholar subsets of the cohort were each searched by the agent and by one human arm, *human-only* or *AI-guided*. In the *human-only* arm, a person searched the open web and scholarly databases without AI tools. In the *AI-guided* arm, a person used a general-purpose AI chat assistant with web search (LibreChat) and verified each suggestion against its source. We pooled the agent's and the human arms' profile findings and used human review to verify each against its source [15]. Recall is therefore relative to this pool, so a profile every search missed is invisible to it. We compared the agent to *human-only* search on one subset and to *AI-guided* search on the other.

### 2.3 Review Study

The *review study* asked how the agent performs as a first-pass author inside the hub's real impact-reporting workflow. The synthesis accuracy of the impact summaries was a secondary check. Because the *profile discovery study* provided the retrieval comparison, the *review study* was a formative single-group study of the human-AI collaboration with no comparison arm.

Two evaluation staff who own the hub's impact-reporting workflow (the reviewers) independently reviewed the agent-generated dossier and impact summaries for each scholar using a structured workbook. The workbook was kept lean on purpose. It showed each finding, its source, and the agent's reasoning, and withheld detail that would add load without aiding judgment. Publications were excluded from row-by-row review. The agent pulls them directly from the bibliographic databases (OpenAlex and PubMed) for the matched author, so each publication is already an indexed record. The remaining check is whether the matched author is the right scholar, which the profile review covers. The workbook covered the eight remaining sections. The agent generated each scholar's dossier and summaries once, and both reviewers independently reviewed all 10 scholars, without seeing each other's ratings. The review workbook held 507 findings (median 44.5 per scholar, range 36 to 85).

The reviewer opened each finding's source before recording an action. Edits were tagged by type (fix value, overstated, fix identity, recategorize, fix source, or other). Rejections were tagged by reason (wrong person, unsupported by source, stale, duplicate, broken link, or other). The tags do not isolate hallucination as its own category. A claim that fails the reviewer's check against its source is rejected as unsupported by source, whether it was invented or overstated. A fabricated or mismatched URL is rejected as a broken link. A broken link can also be a real page that has died since the agent read it. A finding the agent missed could be added as an inserted row. After reviewing the evidence, the reviewer rated synthesis accuracy and usefulness on five-point scales, and whether they would use the dossier.

The primary review measure was the *usable rate*, the share of the agent's findings that both reviewers accepted or edited, reported overall and by section. The unanimous requirement makes this rate a conservative floor. A finding is *not usable* when at least one reviewer rejected it. We coded each *not-usable* finding to a primary cause from the rejecting reviewer's reason and note; when both reviewers rejected with different reasons, we adjudicated to a single cause. We also report each reviewer's individual rate, review time per scholar, the reasons for rejecting a finding, and whether reviewers added anything the agent missed. We computed the inter-rater agreement (Cohen's kappa [16]) on the action coding. The analysis was descriptive. No human-first-draft baseline exists for this workflow, so the *usable rate* stands alone rather than against a comparator.

### 2.4 Reproducibility

The agent runs on OpenAI large language models deployed via Azure OpenAI. Exact versions are in the released configuration. We used AI assistants (based on Anthropic and OpenAI models) to help develop the tool. We reviewed and verified all code and materials. To support transparency and reproducibility, we release the system's full code, prompts, and configuration at `https://github.com/mo-arvan/scholar-dossier-agent` so it can be reproduced. The analysis reported here drew on the hub's internal grant records, which remain private, so the analysis itself is not reproducible by others. The hosted models run in the university's Azure environment under the institution's agreements, and the internal grant records stay within that boundary.

## 3 Results

Both reviewers accepted or edited 81.7% of the agent's findings (Table 1); the dossiers and summaries reached review exactly as the agent drafted them. Most of the agent's output was usable as a first draft. Edits were rare, with 75% of all findings accepted unchanged by both reviewers and each reviewer editing fewer than 5%. Review took each reviewer a median of 14 minutes per scholar. Generating each scholar's draft used a median of about 1.5 million input and 119,000 output tokens. The agent issued 9 to 25 distinct search queries per scholar and made 413 tool calls in aggregate.

**Table 1. Review actions by section.** A finding is *usable* when both reviewers accepted or edited it; split means exactly one reviewer rejected it (507 findings, both reviewers).

| Section | Findings | Both usable | Split | Both reject | Usable % |
|---|---|---|---|---|---|
| Profiles | 93 | 80 | 7 | 6 | 86.0 |
| Career trajectory | 92 | 89 | 3 | 0 | 96.7 |
| Media | 55 | 44 | 6 | 5 | 80.0 |
| Grants | 93 | 80 | 12 | 1 | 86.0 |
| Clinical impact | 77 | 57 | 12 | 8 | 74.0 |
| Community impact | 57 | 37 | 13 | 7 | 64.9 |
| Economic impact | 20 | 14 | 6 | 0 | 70.0 |
| Policy impact | 20 | 13 | 3 | 4 | 65.0 |
| **Overall** | **507** | **414** | **62** | **31** | **81.7** |

*Usable rates* by section ranged from 96.7% (career trajectory) down to community (64.9%) and policy (65.0%) (Table 1). Community and policy both scored low because the finding fell short of a completed impact, though the shortfall looked different in each. In community, these were real activities such as providing clinical care, offered as impacts. In policy, they were claims stronger than their evidence, such as an op-ed treated as expert testimony. Each fell short of the reviewer's threshold for a demonstrated benefit, a line the recall-oriented design leaves for the reviewer to draw. Grants (86.0%) involved a different judgment. The rejected awards were self-reported, but the reviewer judged the scholar's role below principal or co-investigator, such as a consultant. The agent surfaced these deliberately. It kept such an award even when the scholar was not the listed principal investigator, because the funder's principal-investigator field is often blank or names only the contact principal investigator. Gating on that field would drop a scholar's co-investigator and team-science contributions. The reviewer then decided which to keep. The same attribution question explains a few impact findings drawn from those grants and trials. Per-scholar *usable rates* ranged from 66.7% to 93.3%, with eight of ten above 70%.

We coded each of the 93 *not-usable* findings to a primary cause (Figure 3). Sourcing was the most common problem. The largest group cited a non-authoritative source, such as a provider-directory page offered as a profile (weak source, 33). A further eight gave no source at all (empty source), and four pointed to a broken or mismatched link. Those four links were the closest the agent came to hallucination. Next were findings that fell short of a completed impact (not an impact, 22), an activity such as providing clinical care, or a claim stronger than its evidence. Another group was awards or impacts on which the scholar's role was too peripheral (wrong attribution, 16). A smaller group was filed under the wrong category (8), and two were entirely irrelevant to the scholar's work. The reviewer's practical work was therefore curation. The reviewer judged whether each finding cleared the impact bar, then confirmed that it was attributed to the right person, sourced to the right record, and filed under the right category.

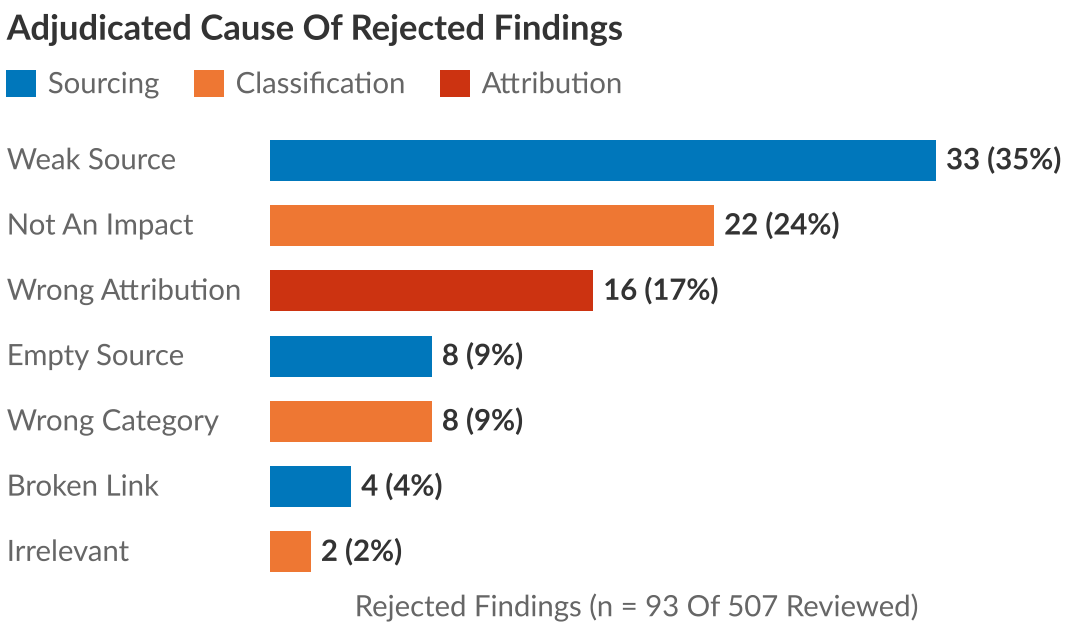


**Figure 3. Why findings were not usable.** Each of the 93 *not-usable* findings (of 507, where at least one reviewer rejected), coded to a primary cause. A weak or non-authoritative source is the largest share, followed by impact-threshold and attribution judgments.

Each impact summary was a single sentence, averaging 25.4 words (SD 6.2). Four reviewer-accepted examples illustrate the register. A clinical example described a multidrug-resistant organism alert tool that raised facility registrations from 40% to 89% across 121 Veterans Affairs facilities. A community example reported a program that lifted colorectal cancer screening from 13% to 20% among 1,252 patients at three Federally Qualified Health Centers. An economic example described a sickle-cell transplant program that cut annual per-patient health-care costs from $64,634 to $16,281 for 16 adults. A policy example credited advocacy that helped Illinois pass a 2018 law letting schools stock albuterol inhalers for asthma emergencies. The agent found impact across all four Translational Science Benefits Model (TSBM) domains, concentrated in clinical and community.

The sources behind the impact evidence varied by domain and by type (Figure 4). Clinical impact drew mostly on journals and trial registries. Community and policy impact drew more on institutional pages and news coverage. No domain drew on a single source type. Each mixed indexed literature with institutional and open-web pages. Even so, 84% of all findings cited a funder record, a digital object identifier (DOI) or PubMed record, a trial registry, or the scholar's own institutional page. The same held for 75% of impact findings.

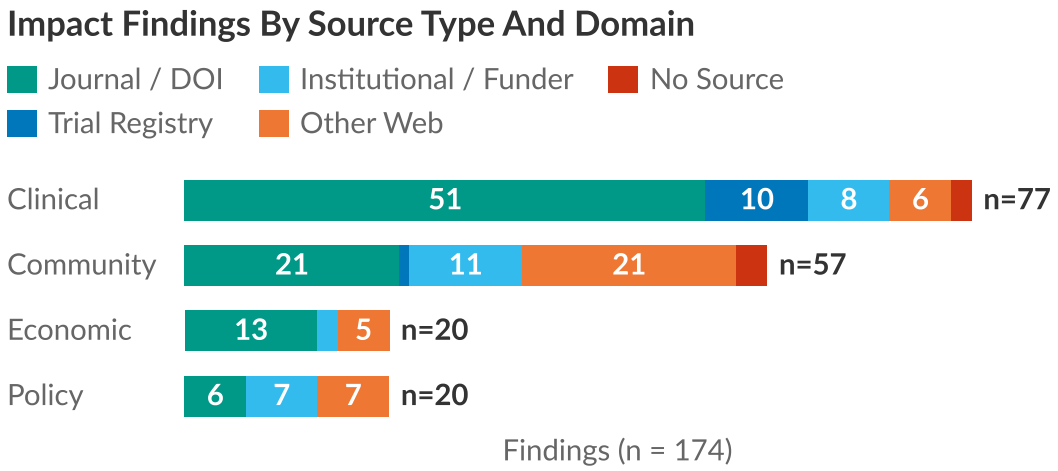


**Figure 4. Where the agent's impact evidence comes from.** The 174 impact findings across the 10 dossiers, by Translational Science Benefits Model (TSBM) domain and source type. Most trace to a journal or digital object identifier (DOI), a trial registry, or an institutional page.

As a secondary check, reviewers rated each dossier. Synthesis accuracy averaged 4.5 of 5 and usefulness 4.8 of 5, and both reviewers said they would use every dossier they rated (10 of 10 each).

Inter-rater agreement was fair to moderate. The usable-versus-reject split is the agreement that matters here, because the accept-versus-edit boundary is a minor rewrite rather than a disagreement about the finding. On that split, Cohen's kappa was 0.43, in the moderate range (Landis-Koch bands [17]). Kappa was lower on the finer-grained schemes, 0.35 on the three-way accept/edit/reject scheme, and 0.40 on the binary accept-versus-any-change decision. The residual disagreement came from the subjective threshold each reviewer draws for a completed impact and from grant attribution. On grants, the two reviewers often handled the same finding differently, one editing the source and the other rejecting it. Taken individually, one reviewer rated 89.9% of findings usable and the other 85.6%; the unanimous rate above counts only the findings both kept.

The reviewers added no findings of their own, consistent with the recall-oriented design, which asks reviewers to prune an over-inclusive list rather than fill gaps. Because they judged the agent's own list, the review reveals what the agent proposed but not what it missed. Discovery, therefore, needs its own check.

On profile discovery, the agent's recall was close to human search on both subsets (Table 2). For the *human-only* search, recall was 0.82 for both the human and the agent across 33 profiles. For the *AI-guided* search, recall was 0.85 for the human versus 0.76 for the agent across 46 profiles. The two human arms ran on non-overlapping subsets, so each comparison is agent versus that arm, and recall is relative to the pooled reference. The gap in the *AI-guided* comparison comes mostly from Google Scholar profiles, which the agent cannot query.

**Table 2. Profile discovery recall, human search versus the agent.** Two non-overlapping five-scholar subsets (*human-only* and *AI-guided* search). Agent recall was close to human search in both comparisons.

| Comparison (same scholars) | Profiles | Human | Agent |
|---|---|---|---|
| Human-only search (n=5) | 33 | 0.82 | 0.82 |
| AI-guided search (n=5) | 46 | 0.85 | 0.76 |

## 4 Discussion

A human-in-the-loop artificial intelligence (AI) agent can serve as a first-pass author of Clinical and Translational Science Award (CTSA) KL2/K12 scholars' impact summaries. In the primary *review study*, both reviewers accepted or edited 81.7% of the agent's findings. Inter-rater agreement was fair to moderate. The agent's profile recall was close to human search. Together these results support a working division of labor between the agent and its reviewers.

The main implication is a relocation of effort. Assembling a scholar's impact evidence by hand is a bottleneck in the translational pathway. Staff estimate roughly 15 hours per scholar, most of it spent collecting evidence and writing

summaries. That magnitude is consistent with published experience. Another CTSA hub co-creates Translational Science Benefits Model (TSBM) impact profiles one research project at a time [18], and a scholar's record spans multiple projects. That effort does not scale to a career-development program's full cohort of scholars. With the agent drafting, reviewing a finding replaces searching for it, verifying it, and writing it from scratch (Figure 5). Both reviewers kept 414 of 507 findings, and at least one rejected the other 93, so confirming what the agent assembled became the bulk of the task. The confirmation work is bounded and schedulable, each reviewer spending a median of 14 minutes per scholar (range 8.5 to 25), so one reviewer covers the cohort in a few hours. A workload measured in minutes is one that a program can plan against a reporting deadline. The intended cadence is one refreshed snapshot per reporting cycle. Each summary is a self-contained benefit statement, so staff can compose the accepted summaries into the impact profiles and narratives their reporting already uses.

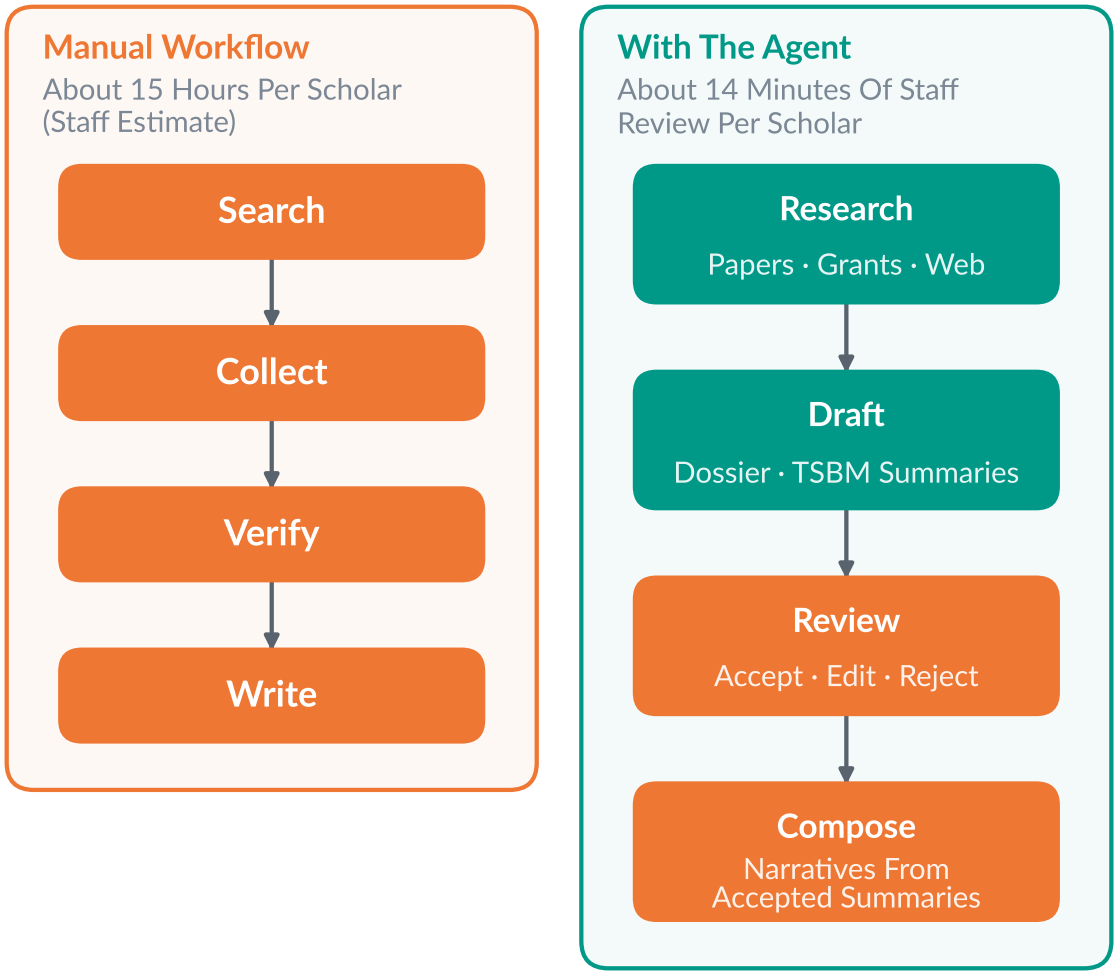


**Figure 5. Division of labor, manual versus with the agent.** In the manual workflow, staff search, collect, verify, and write, an estimated 15 hours per scholar. With the agent, drafting moves to the agent; staff review the drafted findings and compose narratives from the accepted summaries, each reviewer a median of 14 minutes of review per scholar (range 8.5 to 25).

Human judgment remains necessary within the process. The evidence is often ambiguous and current models are imperfect. The agent surfaced an irrelevant wedding announcement despite instructions to include only relevant evidence, the kind of error a person catches at a glance. Both reviewers rejected it. The judgment calls are domain questions, such as whether a grant's attribution holds, whether a policy claim is overstated, or whether a page is really a profile. That judgment is the evaluation staff's existing expertise, and the tool concentrates their time on exactly those calls.

Sourcing failures dominated the rejections. Whether a source is authoritative is a judgment call that stays with the reviewer. After the review round, we built a rule-based filter for the failures that need no judgment. Retrospectively, it flags 21 of the 507 findings, including every rejected finding whose source field was empty. The planned deployment runs the filter as each section's sources are written, so flagged sources are resolved before the draft reaches review. A denylist of known non-authoritative domains would extend it to screen out the provider directories that reviewers repeatedly rejected. Together, the two fixes would resolve a large share of the sourcing failures the reviewers identified.

Prior approaches to research-impact assessment divide into automated pipelines over closed sources and hands-on processes that spend expert hours. The closest prior work applies topic modeling across a historical CTSA corpus and maps topics to TSBM benefit categories with subject-matter-expert validation [19]. An evaluation by the National Center for Advancing Translational Sciences (NCATS) uses a large language model to draft TSBM impact profiles for 55 awards from the awardees' own progress reports, validating a 10% sample with subject-matter experts [20]. Current research information systems (CRIS) serve the same reporting need in production, auto-ingesting publications and grants into a system of record. All three draw only on closed sources, whether indexed, identifier-keyed databases or the investigators' own reports. None reaches the non-scholarly categories of translational impact that this study targets, namely clinical-program leadership, community engagement, cost or adoption data, and policy uptake.

Hands-on processes do reach that evidence, but at a documented cost. The CTSA hub cited earlier co-creates its TSBM impact profiles with the investigators through a four-phase process. Each profile takes roughly 9 hours, and 8 of the 20 investigators it invited never responded [18]. A TSBM self-report survey at a third hub drew completed responses from 67% of invited investigators [21]. Any process that spends hours per profile or needs the scholar's own time inherits a scale and participation ceiling. The agent avoids both costs. It retrieves evidence by open-web search as well as from the indexed databases, works one scholar at a time without the scholar's involvement, and puts every finding through review. As Figure 4 shows, the impact findings trace to institutional pages, news coverage, and the open web alongside journals and registries. The four impact sections hold about a third of the 507 reviewed findings. The agent surfaces evidence that closed sources miss and hands-on processes cannot afford, and staff review it before it enters the system of record.

Coverage is not the only gap this study addresses. Evaluations of AI systems inside the workflows they serve are rare. One structured survey estimated real-world impact evaluations at 0.1% of computational linguistics papers [9]. A systematic review of language models in clinical workflows found only 4 of 288 screened records to be original studies of a live deployment [10]. Our *review study* is one such evaluation. The agent ran inside the real reporting workflow, and we measured the outcomes that workflow cares about, review time per scholar, and the action each finding needed. In-workflow evidence is what a program's adoption decision weighs.

Once an agent writes the first draft, review becomes the bottleneck, so the review stage has to be designed deliberately. We treated review as part of the contribution and designed it accordingly. The workbook showed only what aids the judgment, each finding with its source and the agent's reasoning. Each finding took one action, with the source opened first and the reason tagged. A shared, self-contained codebook aligned the two reviewers. Neighboring domains already measure this review stage in live deployments. AI-drafted data extractions are revised inside live systematic reviews [22], edit reasons are coded into taxonomies for clinical notes [23], edit magnitude and draft uptake are tracked on patient messages [24, 25], and task completion is measured at enterprise scale for code assistants [26]. Research-impact assessment has had no equivalent measurement. This study supplies one.

Coverage across the four domains was strongest for clinical and community impact and sparsest for economic and policy. Much of the imbalance likely reflects the underlying distribution of impacts. The NCATS evaluation found benefits concentrated in the clinical and medical domain [20]. Corpus mining likewise tags economic and policy benefits rarest across its publication corpus [19]. Economic and policy impacts also arrive late in a translational timeline that these early-career scholars have only begun to traverse. Retrieval may add to the imbalance. The agent queries PubMed and ClinicalTrials.gov for publications and trials but has no equivalent interface yet for economic and policy impact, so it falls back on less reliable open-web search. The study cannot separate the two explanations, because the completeness of the impact evidence is unmeasured.

Closing the retrievable part of the gap means wiring in databases that already exist. Patent records are the clearest addition. Prior TSBM mining already pulls them from the United States Patent and Trademark Office by identifier [19]. They would slot into the same data interfaces the agent uses for trials and publications. Policy-citation indexes such as Overton [27], legislative databases, and cost or adoption data would ground the policy and economic domains the same way. Overton already indexes policy references to 13% of CTSA-supported publications [28]. Wiring in these interfaces would extend the agent's retrieval into the economic and policy domains, where this study's coverage was thinnest.

Several limitations bound these claims. The sample is small, 10 scholars at one hub. The pipeline's design is not site-specific. Only the internal grant seed is local, and the rest draws on public sources. Whether performance replicates at another hub is untested. The time saved rests on a staff estimate of manual effort rather than a measured comparison within this study, although the estimate is consistent with published per-profile effort at another hub [18]. The *profile discovery study* checks retrieval of profiles, and the *review study* judges only what the agent proposed, so the completeness of the impact evidence is not established.

The system itself carries limits. It runs on hosted proprietary language models, so exact reproduction depends on model availability and version; the released configuration pins the versions we used. We evaluated OpenAI models only. Finally, the fencing against prompt injection is containment rather than elimination, since the small model still ingests raw page text. Three bounds contain that exposure. The agent's capabilities are limited to the search, database, and page-reading calls described; the work is capped at every level (turns per run, runs per scholar, and context size); and human review of every finding remains the backstop.

## 5 Conclusions

A human-in-the-loop artificial intelligence (AI) agent can serve as the first-pass author of a scholar's impact record, turning source evidence into one-sentence impact summaries that staff review into Translational Science Benefits Model (TSBM) benefit statements. In a real reporting workflow, staff found most of the output usable. The effort shifted from collecting and writing to reviewing. The evaluation is as much a contribution as the tool. We measured the agent in the workflow it serves, by the outcomes that workflow cares about. Future work includes multi-site evaluation, strategies to improve economic and policy impact coverage, and efficiency optimizations that keep each scholar's record current at lower cost.

## 6 Acknowledgments

The authors thank Charles Frisbie, Medical Device Labs Director at the University of Illinois Chicago Innovation Center, for his support of this work.

The authors used AI assistants based on Anthropic and OpenAI models to help draft and edit this manuscript. They accessed these tools through the providers' standard commercial interfaces. The authors reviewed and revised all output and are entirely responsible for the scientific content of the paper.

## 7 Financial Support

This work was supported by the AI.Health4All Center at the University of Illinois Chicago.

The project described was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UM1TR005438. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

## 8 Competing Interests

The authors declare none.

## References

[1] Marisa L. Conte and M. Bishr Omary. NIH Career Development Awards: conversion to research grants and regional distribution. *The Journal of Clinical Investigation*, 128(12):5187–5190, December 2018. ISSN 0021-9738. doi: 10.1172/JCI123875. URL https://www.jci.org/articles/view/123875.

[2] Silda Nikaj and P Kay Lund. The Impact of Individual Mentored Career Development (K) Awards on the Research Trajectories of Early-Career Scientists. *Academic Medicine*, 94(5):708–714, May 2019. ISSN 1040-2446. doi: 10.1097/ACM.0000000000002543. URL https://doi.org/10.1097/ACM.0000000000002543.

[3] 12.2.5 Institutional Scientist Development Programs, March 2026. URL https://grants.nih.gov/grants/policy/nihgps/html5/section_12/12.2.5_institutional_scientist_development_programs.htm.

[4] Christine A. Sorkness, Linda Scholl, Alecia M. Fair, and Jason G. Umans. KL2 mentored career development programs at clinical and translational science award hubs: Practices and outcomes. *Journal of Clinical and Translational Science*, 4(1):43–52, February 2020. ISSN 2059-8661. doi: 10.1017/cts.2019.424. URL https://www.cambridge.org/core/journals/journal-of-clinical-and-translational-science/article/kl2-mentored-career-development-programs-at-clinical-and-translational-science-award-hubs-practices-and-outcomes/DF613F08826E1613CE8039E71C7CC219.

[5] Kelli Qua, Fei Yu, Tanha Patel, Gaurav Dave, Katherine Cornelius, and Clara M. Pelfrey. Scholarly Productivity Evaluation of KL2 Scholars Using Bibliometrics and Federal Follow-on Funding: Cross-Institution Study. *Journal of Medical Internet Research*, 23(9):e29239, September 2021. doi: 10.2196/29239. URL https://www.jmir.org/2021/9/e29239.

[6] Eric J. Nehl, Clara M. Pelfrey, Deborah DiazGranados, Gaurav Dave, and Nicole M. Llewellyn. Academic influencers: Clinical and Translational Science scholars and trainees at the intersection of influential scholarship and public attention. *Journal of Clinical and Translational Science*, 9(1):e130, January 2025. ISSN 2059-8661. doi: 10.1017/cts.2025.10067. URL https://www.cambridge.org/core/journals/journal-of-clinical-and-translational-science/article/academic-influencers-clinical-and-translational-science-scholars-and-trainees-at-the-intersection-of-influential-scholarship-and-public-attention/5E3CF02A329F85F8A8E37D972AFEFCE3.

[7] Rebecca Helton, Scott Pearson, and Katherine Hartmann. 112 Flight Tracker: A REDCap Tool to Streamline Career Development Grant Preparation and Reporting. *Journal of Clinical and Translational Science*, 8(s1):32, April 2024. ISSN 2059-8661. doi: 10.1017/cts.2024.110. URL https://www.cambridge.org/core/journals/journal-of-clinical-and-translational-science/article/112-flight-tracker-a-redcap-tool-to-streamline-career-development-grant-preparation-and-reporting/D42C583CF9E0D6F6188177E2362D9E03.

[8] Douglas A. Luke, Cathy C. Sarli, Amy M. Suiter, Bobbi J. Carothers, Todd B. Combs, Jae L. Allen, Courtney E. Beers, and Bradley A. Evanoff. The Translational Science Benefits Model: A New Framework for Assessing the Health and Societal Benefits of Clinical and Translational Sciences. *Clinical and Translational Science*, 11(1): 77–84, January 2018. ISSN 1752-8054. doi: 10.1111/cts.12495. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5759746/.

[9] Ehud Reiter. We Should Evaluate Real-World Impact. *Computational Linguistics*, 51(4):1419–1431, December 2025. ISSN 0891-2017. doi: 10.1162/COLI.a.18. URL https://doi.org/10.1162/COLI.a.18.

[10] Yaara Artsi, Vera Sorin, Benjamin S. Glicksberg, Panagiotis Korfiatis, Girish N. Nadkarni, and Eyal Klang. Large language models in real-world clinical workflows: a systematic review of applications and implementation. *Frontiers in Digital Health*, 7:1659134, 2025. ISSN 2673-253X. doi: 10.3389/fdgth.2025.1659134. URL https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1659134.

[11] Anthropic. Building Effective AI Agents, 2024. URL https://www.anthropic.com/engineering/building-effective-agents.

[12] Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In Maura Pintor, Xinyun Chen, and Florian Tramèr, editors, *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, AISec 2023, Copenhagen, Denmark, 30 November 2023*, pages 79–90. ACM, 2023. doi: 10.1145/3605764.3623985.

[13] Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending Against Indirect Prompt Injection Attacks With Spotlighting. In Rachel Allen, Sagar Samtani, Edward Raff, and Ethan M. Rudd, editors, *Proceedings of the Conference on Applied Machine Learning in Information Security (CAMLIS 2024), Arlington, Virginia, USA, October 24-25, 2024*, volume 3920 of *CEUR Workshop Proceedings*, pages 48–62. CEUR-WS.org, 2024. URL https://ceur-ws.org/Vol-3920/paper03.pdf.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[15] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press, August 2005. ISBN 978-0-262-22073-6.

[16] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. ISSN 0013-1644. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

[17] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174, March 1977. ISSN 0006-341X.

[18] Kera Swanson, Nicole A. Stadnick, Isaac Bouchard, Zeying Du, Lauren Brookman-Frazee, Gregory A. A. Aarons, Emily Treichler, Maryam Gholami, and Borsika A. A. Rabin. An iterative approach to evaluating impact of CTSA projects using the translational science benefits model. *Frontiers in Health Services*, 5:1535693, 2025. ISSN 2813-0146. doi: 10.3389/frhs.2025.1535693. URL https://www.frontiersin.org/journals/health-services/articles/10.3389/frhs.2025.1535693.

[19] Tejaswini Manjunath, Eline Appelmans, Sinem Balta, Dominick DiMercurio, Claudia Avalos, and Karen Stark. Topic analysis on publications and patents toward fully automated translational science benefits model impact extraction. *Frontiers in Research Metrics and Analytics*, 10:1596687, September 2025. ISSN 2504-0537. doi: 10.3389/frma.2025.1596687. URL https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2025.1596687/full.

[20] Kristopher Bough, Francisco Leyva, Monica Donerson, and Soju Chang. 243 An AI-driven, Translational Science Benefits Model (TSBM) approach to assess the real-world impacts of NCATS' CTSA Collaborative and Innovative Acceleration Award Initiative. *Journal of Clinical and Translational Science*, 10(s1):81–82, April 2026. ISSN 2059-8661. doi: 10.1017/cts.2026.10445. URL https://www.cambridge.org/core/journals/journal-of-clinical-and-translational-science/article/243-an-aidriven-translational-science-benefits-model-tsbm-approach-to-assess-the-realworld-impacts-of-ncats-ctsa-collaborative-and-innovative-acceleration-award-initiative/71AFFBF1BE517C4C3150C42E4E9FB5C9.

[21] Andrea Molzhon, Pamela M. Dillon, and Deborah DiazGranados. Leveraging the translational science benefits model to enhance planning and evaluation of impact in CTSA hub-supported research. *Frontiers in Public Health*, 13:1593920, June 2025. ISSN 2296-2565. doi: 10.3389/fpubh.2025.1593920. URL https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2025.1593920/full.

[22] Gerald Gartlehner, Shannon Kugley, Karen Crotty, Meera Viswanathan, Andreea Dobrescu, Barbara Nussbaumer-Streit, Graham Booth, Jonathan R. Treadwell, Jung Min Han, Jesse Wagner, Eric A. Apaydin, Erin L. Coppola, Margaret Maglione, Rainer Hilscher, Robert Chew, Meagan Pilar, Bryan Swanton, and Leila C. Kahwati. Artificial Intelligence–Assisted Data Extraction With a Large Language Model: A Study Within Reviews. *Annals of Internal Medicine*, 178(12):1763–1771, December 2025. ISSN 0003-4819. doi: 10.7326/ANNALS-25-00739. URL https://www.acpjournals.org/doi/10.7326/ANNALS-25-00739.

[23] Yawen Guo, Di Hu, Ziqi Yang, Seungjun Kim, Brian Tran, Jamie Lee, Sitha Vallabhaneni, Rachael Zehrung, Sairam Sutari, Steven Tam, Emilie Chow, Danielle Perret, Deepti Pandita, and Kai Zheng. What do clinicians edit in ambient AI-drafted clinical documentation? A qualitative content analysis. *Journal of the American Medical Informatics Association*, page ocag073, May 2026. ISSN 1527-974X. doi: 10.1093/jamia/ocag073. URL https://doi.org/10.1093/jamia/ocag073.

[24] Charlotte M. H. H. T. Bootsma-Robroeks, Jessica D. Workum, Stephanie C. E. Schuit, Anne Hoekman, Tarannom Mehri, Job N. Doornberg, Tom P. van der Laan, and Rosanne C. Schoonbeek. AI-generated draft replies to patient messages: exploring effects of implementation. *Frontiers in Digital Health*, 7:1588143, June 2025. ISSN 2673-253X. doi: 10.3389/fdgth.2025.1588143. URL https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1588143/full.

[25] Soumik Mandal, Batia M. Wiesenfeld, Adam C. Szerencsy, William R. Small, Vincent Major, Safiya Richardson, Antoinette Schoenthaler, Devin Mann, and Oded Nov. Utilization of Generative AI-drafted Responses for Managing Patient-Provider Communication. *npj Digital Medicine*, 8(1):591, October 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01972-w. URL https://www.nature.com/articles/s41746-025-01972-w.

[26] Zheyuan (Kevin) Cui, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers, August 2025. URL https://papers.ssrn.com/abstract=4945566.

[27] Martin Szomszor and Euan Adie. Overton: A bibliometric database of policy document citations. *Quantitative Science Studies*, 3(3):624–650, November 2022. ISSN 2641-3337. doi: 10.1162/qss_a_00204. URL https://doi.org/10.1162/qss_a_00204.

[28] Nicole M Llewellyn, Amber A Weber, Clara M Pelfrey, Deborah DiazGranados, and Eric J Nehl. Translating Scientific Discovery Into Health Policy Impact: Innovative Bibliometrics Bridge Translational Research

Publications to Policy Literature. *Academic Medicine*, 98(8):896–903, August 2023. ISSN 1040-2446. doi: 10.1097/ACM.0000000000005225. URL https://doi.org/10.1097/ACM.0000000000005225.