nature cancer

Article

https://doi.org/10.1038/s43018-026-01264-6

An agentic AI system enhances clinical detection of immune-related adverse events in a multi-phase validation study

Received: 21 April 2026

Accepted: 28 September 2026

Published online: 29 October 2026

Check for updates

Jack Gallifant[1,2], Shan Chen[1,2], Kee-Young Shin[2,3], Katherine C. Kellogg[4], Patrick F. Doyle[1,3], Joyce Guo[1], Bingyang Ye[1,2], Andrew Warrington[1,3], Bingxue K. Zhai[1,3], Matthew J. Hadfield[5,6], Alexander Gusev[2,3], Biagio Ricciuti[2,3], David C. Christiani[1,2], Hugo JWL Aerts[1,2,7], Benjamin H. Kann[1,2,3], Raymond H. Mak[1,2,3], Tanna L. Nelson[8], Paul Nguyen[1,2,3], Jonathan D. Schoenfeld[1,2,3], Umit Topaloglu[8], Paul Catalano[2,3], Harry Hochheiser[9], Jeremy L. Warner[5,6], Elad Sharon[2,3], David E. Kozono[1,2,3], Guergana K. Savova[2,10] & Danielle S. Bitterman[1,2,3,10]

Immune-related adverse events (irAEs) affect up to 40% of patients receiving immune checkpoint inhibitors, yet identification requires laborious, inconsistent manual chart review. Here we show that agentic artificial intelligence coupled with human verification could enhance efficiency, performance and consistency for irAE assessment. We developed and evaluated an agentic large language model system to extract the presence, temporality, severity grade, attribution and certainty of six irAE types from clinical notes. Retrospectively, the system achieved a macro-averaged F_1 of 0.92 for current-event detection and 0.66 for multi-class severity grading; self-consistency improved F_1 by 0.14. The best-performing configuration cost approximately US$0.02 per note. In prospective silent deployment over 3 months, the binary F_1 was 0.72–0.79. In a randomized crossover study of clinical trial staff, agentic assistance reduced annotation time by 2.71 min ($P < 0.001$), increased complete-match accuracy (odds ratio 1.45; 95% confidence interval 1.01–2.08; $P = 0.045$) and improved interannotator agreement (Krippendorff's α from 0.22–0.51 to 0.82–0.85).

Treatment-emergent adverse events (TEAEs) are under-recognized in clinical trials and existing databases[1–8]. Accurate detection and reporting of TEAEs is foundational to cancer care, pharmacovigilance[9] and survivorship planning. In routine practice, however, many clinically important adverse events are documented primarily in unstructured clinical narratives; structured electronic health record (EHR) fields incompletely capture their occurrence, necessitating laborious, resource-intensive and error-prone manual chart review[10,11]. This manual bottleneck currently limits the capacity and quality of adverse event curation and monitoring for clinical trials, real-world evidence and clinical decision-support at the point of care.

Against this backdrop, immune checkpoint inhibitors (ICIs) have transformed cancer treatment, yet their success comes with an important challenge: up to 40% of patients develop immune-related adverse events (irAEs) that can affect any organ system[12–16]. While timely detection of irAEs saves lives, evidence of these toxicities is often buried within clinical notes, escaping standard surveillance methods[15,17]. Furthermore, addressing the challenge of irAE surveillance and management will become increasingly urgent as the number of cancer survivors grows to an estimated 26 million anticipated by 2040 in the USA alone[18]. Furthermore, the clinical burden is compounded by substantial economic costs, with irAE-related hospitalizations

A full list of affiliations appears at the end of the paper. e-mail: dbitterman@bwh.harvard.edu

https://doi.org/10.1038/s43018-026-01264-6

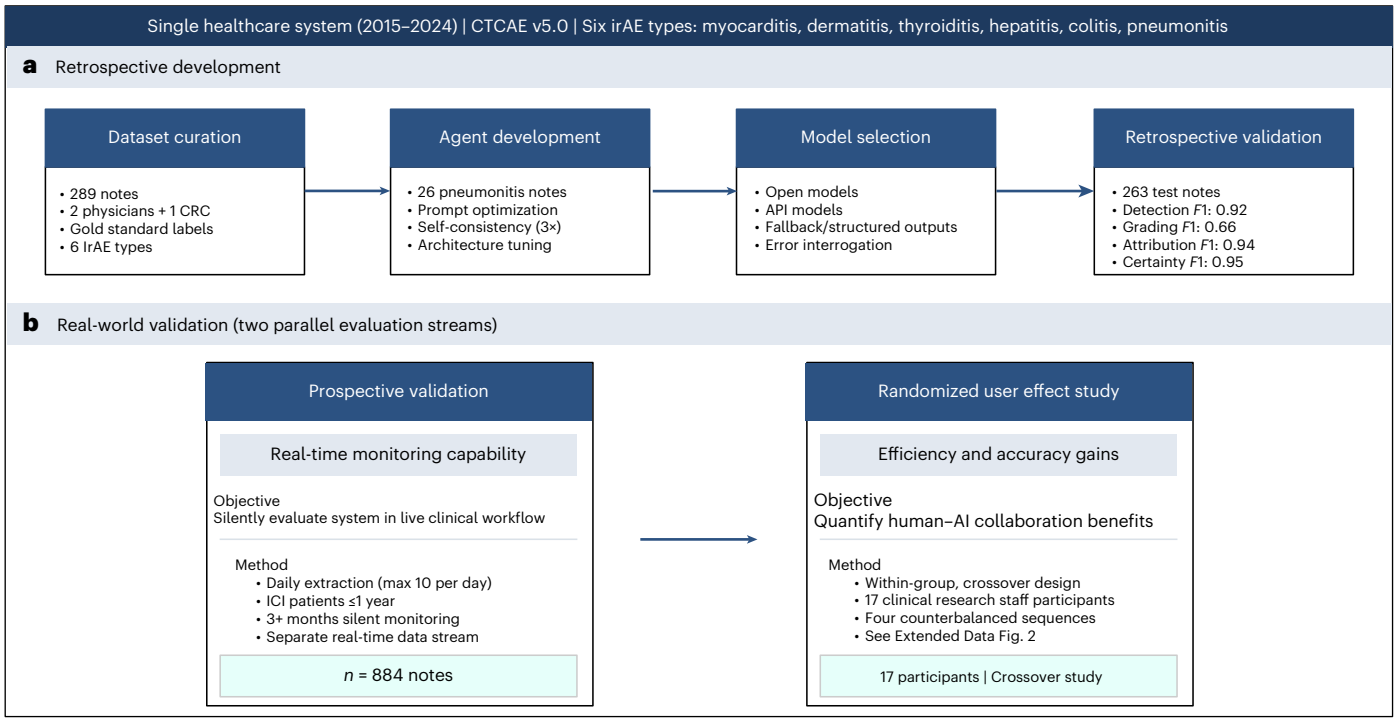


Fig. 1 | Multi-dimensional validation framework for an agentic LLM-based irAE detection system. a, Retrospective development workflow progressing from dataset curation (289 expert-annotated notes) through agent development with prompt optimization on pneumonitis subset ($n = 26$) to validation on held-out test set ($n = 263$), achieving F_1 scores of 0.92 (current-event detection), 0.66 (current CTCAE grading) and 0.94 (temporality-agnostic attribution). The agentic architecture uses six specialized agents with self-consistency mechanisms for robust extraction. **b**, Two complementary validation streams: prospective silent deployment evaluates performance on 884 clinical notes, and a controlled HCI crossover study evaluates annotation efficiency, accuracy and agreement with 20 unique notes and 17 participants (16 CRCs and 1 CRN). BWH, Brigham and Women's Hospital; DFCI, Dana-Farber Cancer Institute.

and treatment adding thousands of dollars per patient and requiring resource-intensive manual chart review for detection[19].

Despite their clinical importance, systematically identifying cancer TEAEs presents a considerable challenge for both clinical care and pharmacovigilance[20]. Directly analyzable structured data fields within EHRs do not reliably capture irAEs, especially those that do not result in hospitalization but still have important impacts on patients' overall morbidity and quality of life[11]. Even of those that result in hospitalization, International Classification of Diseases (ICD) codes have a recall/sensitivity of only approximately 68%[21]. Because of this, manual chart curation is required for irAE reporting, but this suffers from the aforementioned manual curation bottleneck and introduces quality challenges owing to ambiguous definitions and thresholds for irAE occurrence, severity and attribution[22,23].

To address these challenges, natural language processing, which refers to computational analysis of unstructured text and is a central pillar in modern artificial intelligence (AI)[24], has been explored to extract irAEs from EHR documents automatically. Initial efforts using traditional machine learning or small neural language models demonstrated promise but require large annotated datasets for task-specific training and tuning[25]. More recently, large language models (LLMs) have been shown to extract clinical information without extensive fine-tuning in a zero-shot or few-shot scenario through sophisticated prompting. Preliminary studies applying LLMs to irAE identification have demonstrated promise[21,25–28]; however, they have focused on binary classification (that is, irAE present/absent) over an entire inpatient admission record or patient chart. As these approaches operate at the chart or encounter level rather than on individual clinical notes, they are poorly suited for real-time detection and ongoing monitoring. Furthermore, they have not addressed the extraction of irAE clinical attributes necessary to inform clinical decision-making, translational research and regulatory reporting for clinical trials. These critical attributes include the temporal status (distinguishing between current, active events and historical ones), severity grading (aligned with standardized scales), attribution (the clinician's assessment of the likelihood that the event is ICI related) and the certainty of that attribution. Extracting these details accurately from narrative text remains an unmet need to understand and address irAEs and adverse events more broadly.

Concurrently, advanced autonomous prompting strategies and architectures, commonly referred to as 'agentic' systems, aim to further enhance LLM capabilities[29]. These systems decompose complex reasoning tasks into smaller, manageable subproblems, often addressed by multiple specialized LLM-based or utilizing tools (agents), and with mechanisms to evaluate and synthesize the final output[29]. It remains an open question whether such agentic architectures can be effectively leveraged for the complex, multi-faceted task of extracting detailed irAE attributes from clinical notes. Even if technically feasible, the impact of integrating such sophisticated AI assistance into real-world clinical and research workflows, including its effect on efficiency, performance, cognitive load and overall experience, remains unevaluated. Understanding these human–computer interaction (HCI) factors is necessary to translate technological potential into practical, usable and valuable healthcare applications.

Here, we evaluated the impact of real-time agentic LLM assistance on the human curation of detailed irAEs from clinical notes. We developed an agentic pipeline that extracts not only the presence of six irAEs spanning common and lethal events (myocarditis, dermatitis, thyroiditis, hepatitis, colitis and pneumonitis)[30] but also their temporal status, severity grade, ICI attribution and attribution certainty at the encounter level. We integrated this system into an annotation interface and conducted HCI studies to determine whether AI assistance improves reviewer efficiency, accuracy and consistency compared with unaided manual review. We also approximated the costs of computation. We demonstrated the potential of agentic AI to improve both the quality

https://doi.org/10.1038/s43018-026-01264-6

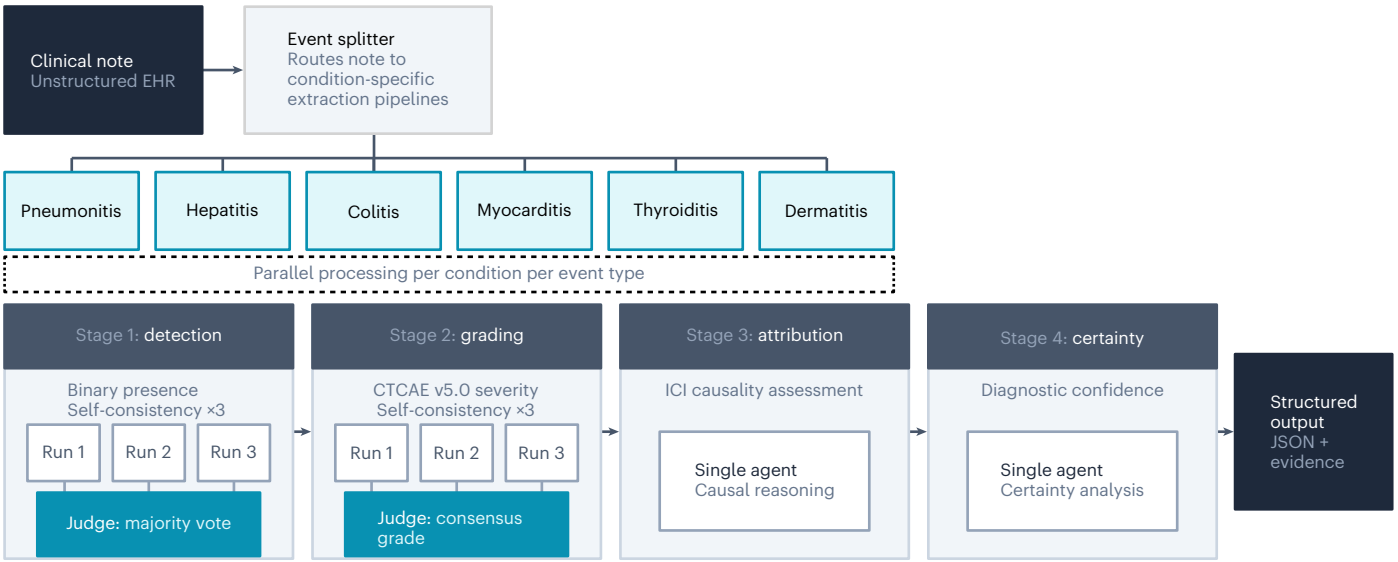


Fig. 2 | Multi-stage agentic architecture for irAE detection, grading, attribution and certainty assessment. Unstructured clinical notes from the EHR are first processed by an Event Splitter module that routes each note to condition-specific extraction pipelines spanning six irAE subtypes (pneumonitis, hepatitis, colitis, myocarditis, thyroiditis and dermatitis), enabling parallel processing across organ systems. The pipeline proceeds through four sequential stages: (1) detection—binary classification of irAE presence using self-consistency decoding (×3 runs) with judge adjudication; (2) grading—severity assignment per CTCAE v5.0 criteria using self-consistency decoding (×3 runs) with consensus-based grade selection; (3) attribution—ICI causality assessment via a single-agent causal reasoning step; and (4) certainty—diagnostic confidence estimation via a single-agent certainty analysis step. Self-consistency is applied to the detection and grading stages, where classification ambiguity is the highest, while attribution and certainty rely on single-pass inference given their more subjective, reasoning-intensive nature. The pipeline produces structured JavaScript Object Notation (JSON) output with supporting evidence traces for each identified event.

and efficiency of irAE curation, which could enhance our understanding and optimization of the benefits and risks of cancer treatment via clinical trials and real-world evidence generation.

Results

This multi-phase study evaluated an agentic LLM system for extracting irAE information from routine clinical documentation, progressing from retrospective development to real-world deployment (Fig. 1). We first developed and retrospectively evaluated a note-level agentic extraction system (phase 1). We then validated prospective performance with the system integrated in the EHR (phase 2) and quantified the effect of human–agentic collaboration on curation efficiency and quality in a randomized user effect study informed by a preparatory field study (phase 3; Methods).

Note-level irAE system development and evaluation (phase 1)

To identify the optimal configuration for note-level irAE extraction (Fig. 1), we systematically benchmarked combinations of proprietary and open-source LLMs across agentic and nonagentic architectures on 263 expert-annotated clinical notes. Two design features are central to interpreting these results. First, each irAE was classified by event temporality, that is, whether the clinical note described it as a currently active event or as a historical (past) event, and performance was evaluated for each timeframe separately. Second, to reduce the variability inherent in single LLM outputs, we implemented a self-consistency mechanism in which specific stages of the agentic process were run three times independently, and a judge agent resolved disagreements before proceeding to the next stage; a single-inference variant that omitted this step served as the primary comparator for ablation analyses (Fig. 2). Among all model–architecture combinations evaluated, GPT-4.1-mini with the default agentic architecture achieved the strongest overall performance (Supplementary Table 1). This configuration incurred an inference cost of US$0.021 per note; costs for all model–architecture combinations are presented in Supplementary Table 1. For note-level irAE detection, the system attained a temporality-agnostic macro-averaged F_1 of 0.93 across the six conditions (Table 1). Performance was consistent when stratified by event temporality: $F_1 = 0.92$ for active (current) events and 0.91 for resolved (past) events. Current temporality detection performance was the highest for hepatitis and pneumonitis ($F_1 = 0.96$) and the lowest for myocarditis ($F_1 = 0.86$). Condition-specific binary classification performance (precision, recall and F_1) is shown in Extended Data Fig. 1.

Multi-class Common Terminology Criteria for Adverse Events (CTCAE) grade assignment showed lower performance with an overall macro F_1 of 0.66 for current events (Fig. 3a). Performance varied by condition: dermatitis and thyroiditis achieved the highest performance ($F_1 = 0.73$ and 0.72, respectively), while colitis had the lowest performance ($F_1 = 0.51$). Intermediate grades (2–3) accounted for the majority of misclassifications. The system demonstrated stronger performance when severity was dichotomized at clinically relevant thresholds (grade <2 versus ≥2, and grade <3 versus ≥3), although performance varied according to irAE type (Fig. 3).

The system reliably captured clinician-documented attribution assessments, with a binary F_1 of 0.92 (current events) and 0.94 (past events). Certainty estimation remained high for binary classification (any stated certainty versus none; $F_1 = 0.93$–0.94).

The modest difference between temporality-agnostic and time point-specific detection ($F_1 = 0.93$ versus 0.92/0.91) indicates that residual errors were driven primarily by missed detections or incorrect grading rather than confusion between active and resolved events (Table 1).

Within the GPT-4.1-mini family, the single-inference variant (without self-consistency) reduced detection F_1 from 0.92 to 0.78, while approximately halving inference cost (US$0.011 versus US$0.021 per note), confirming the contribution of multi-run consensus to overall performance. Smaller models (4.1-nano and 4o-mini) and open-source alternatives consistently underperformed across detection and grading tasks (Supplementary Table 1). Notably, a rule-based baseline using customized keyword-matching regular expressions including 779 terms achieved reasonable performance for binary detection (that is,

https://doi.org/10.1038/s43018-026-01264-6

Table 1 | Retrospective performance of the irAE detection and grading pipeline across six irAE subtypes

	Grade						Attribution			Certainty		
	Binary F_1[a]			Macro F_1[b]			Binary F_1			Binary F_1		
Condition	Current	Past	Agnostic	Current	Past	Agnostic	Current	Past	Agnostic	Current	Past	Agnostic
Overall[c]	0.92	0.91	0.93	0.66	0.76	0.75	0.92	0.94	0.94	0.93	0.94	0.95
Thyroiditis	0.94	0.97	1.00	0.72	0.82	0.91	0.87	1.00	0.95	0.87	1.00	0.95
Hepatitis	0.96	0.91	0.97	0.66	0.82	0.74	0.98	0.95	0.97	0.98	0.98	0.98
Colitis	0.87	0.95	0.90	0.51	0.78	0.70	0.93	0.92	0.96	0.91	0.88	0.92
Pneumonitis	0.96	0.93	0.94	0.67	0.74	0.67	0.88	0.96	0.95	0.96	0.95	0.97
Myocarditis	0.86	0.81	0.85	0.67	0.52	0.64	0.93	0.91	0.91	0.90	0.88	0.90
Dermatitis	0.92	0.91	0.91	0.73	0.90	0.84	0.92	0.90	0.93	0.95	0.94	0.96

Temporality settings: current, current episode only; past, historical mentions only; agnostic, no temporal distinction. [a]Macro-averaged metrics for binary classification (frade 0 versus grade 1+) for the indicated temporality. These are the averages of metrics for both class 0 and class 1. [b]Multi-class macro F_1: macro-averaged F_1 score across all six grade classes (0–5) for the indicated temporality. For classes with no examples, F_1 is set to 1.0. [c]Overall (average across conditions): simple unweighted average of each metric across all six conditions (thyroiditis, hepatitis, colitis, pneumonitis, myocarditis and dermatitis). Note: results with GPT-4.1-mini.

flagging whether an irAE was present or absent), performing comparably to or better than even some LLM-based methods. This finding underscores the utility of simple pattern-matching approaches for case identification; however, regular expressions are inherently limited to surface-level keyword matching and cannot extract the granular clinical attributes (such as severity grade, attribution, certainty and supporting evidence spans) required for actionable irAE characterization.

Prospective silent validation (phase 2)

To assess real-world generalizability, we deployed the system in silent, prospective mode for 3 months (May–July 2025), processing 884 clinical notes in near-real time. The system generated 31,824 label adjudications across grade, attribution and certainty for both current and past time points. We found that detection performance remained robust, although attenuated relative to retrospective benchmarks.

Prospective performance was lower relative to retrospective benchmarks, as anticipated, owing to temporal drift in documentation patterns and case mix (Table 2, Fig. 3b and Supplementary Table 2). In addition to the results in Table 2, binary F_1 across grade, attribution and certainty labels was 0.72 for current events and 0.79 for past events. Attribution extraction remained strong ($F_1 = 0.77$), as did certainty classification ($F_1 = 0.80$). At the note level, model predictions matched human review in 62% of cases for grade (549/884), 87% for attribution (771/884) and 83% for certainty (732/884).

Randomized user effect study (phase 3)

To determine whether agentic assistance improves human curation of irAEs in practice, we conducted a randomized 2 × 2 crossover study comparing AI-assisted annotation to standard manual review among clinical research staff (Methods). We found that AI assistance substantially reduced annotation time, improved accuracy and enhanced interannotator agreement. In total, 21 participants enrolled in the study (20 clinical research coordinators (CRCs) and 1 clinical research nurse (CRN)), and 4 participants withdrew consent before starting annotations (all CRCs). All 17 remaining participants (16 CRCs and 1 CRN) completed the study (Supplementary Table 3), yielding 316 evaluable timing observations for the primary note-level analysis. Collectively, 24 note-level observations across participants were excluded owing to missing video recordings.

Table 3 presents results for the randomized user effect study. For our primary end point of annotation efficiency, AI assistance substantially reduced the time required to curate irAE labels from clinical notes. Median annotation time was 428 s (interquartile range (IQR) 253–671) with manual review versus 242 s (IQR 160–385) with AI assistance—a 43.5% reduction (Fig. 3c). In the primary generalized estimating equations (GEE) analysis of untransformed time, accounting for within-participant clustering, the estimated difference was −2.71 min (95% CI −3.89 to −1.53; $P = 6.98 \times 10^{-6}$). Neither note set nor method sequence significantly influenced efficiency in multivariable models, suggesting no substantial carryover effects.

For our secondary end point of performance, complete-match accuracy (requiring all six current grade labels to exactly match the gold standard) was modest overall, reflecting the stringency of this metric: 19.4% for manual annotation versus 25.9% with AI assistance. In GEE analysis, AI assistance increased the odds of complete accuracy by 45% (odds ratio 1.45; 95% CI 1.01 to 2.08; $P = 0.045$). This effect remained significant after adjustment for note set and sequence.

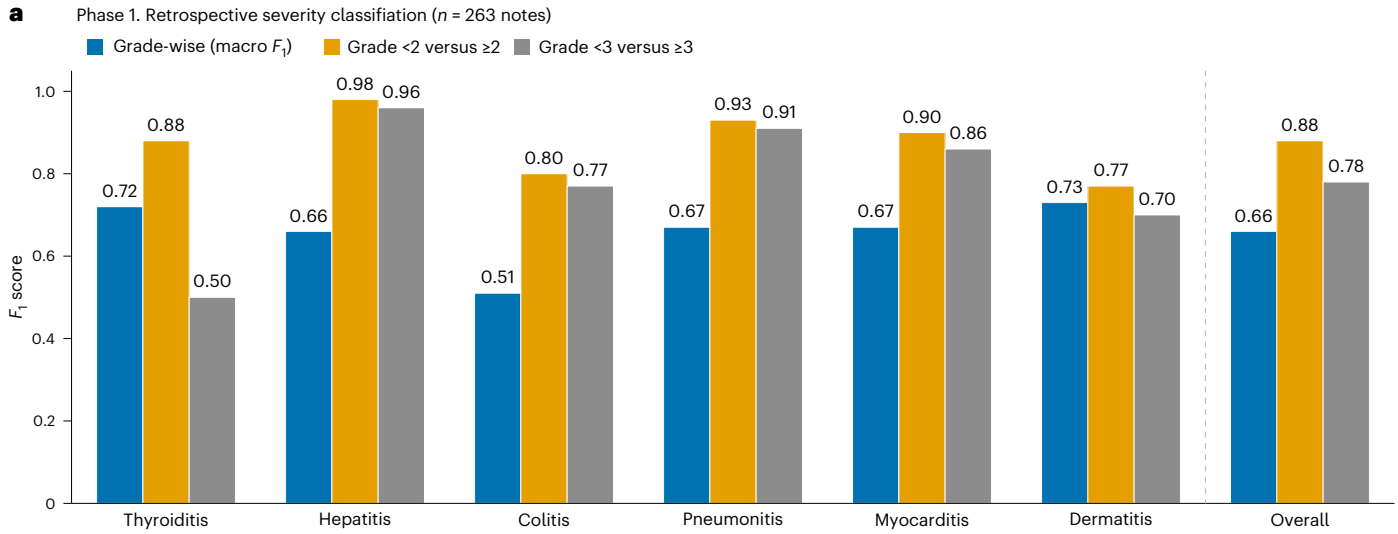
Further, for interannotator agreement, AI assistance markedly improved consistency across annotators (Fig. 3d). For current grade labels, Krippendorff's α increased from 0.22 (95% bootstrap CI 0.16 to 0.29) to 0.85 (95% CI 0.79 to 0.89) in note set A, and from 0.51 (95% CI 0.50 to 0.51) to 0.82 (95% CI 0.76 to 0.87) in note set B. The magnitude of improvement was substantial in both sets ($\Delta\alpha > 0.30$), indicating that AI assistance reduced between-annotator variability and promoted standardization. Exploratory analyses of attribution and certainty labels showed concordant improvements: for attribution, Krippendorff's α increased from 0.22 to 0.86 (note set A) and from 0.42 to 0.75 (note set B); for certainty, it increased from 0.18 to 0.84 and from 0.44 to 0.74, respectively.

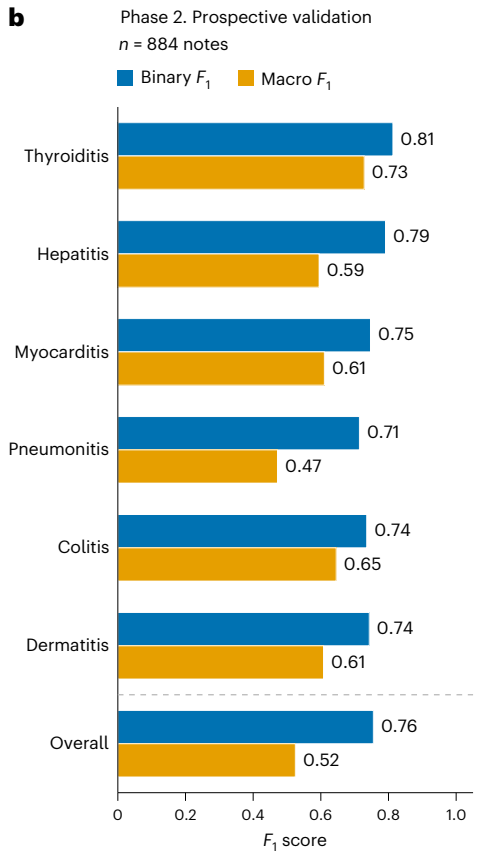
Among 17 complete poststudy respondents, mean System Usability Scale scores were 45.7 (manual) and 67.4 (AI-assisted; Supplementary Table 4). Poststudy ratings on five-point scales were higher for AI-assisted efficiency (3.9 versus 2.5) and confidence in assigned labels (3.8 versus 3.1). AI-capability confidence averaged 2.4 before and 3.0 after the study among these respondents; the differently worded questions are summarized descriptively Supplementary Table 4 (ref. 31). In total, 15 of 17 respondents (88%) preferred the AI-assisted workflow.

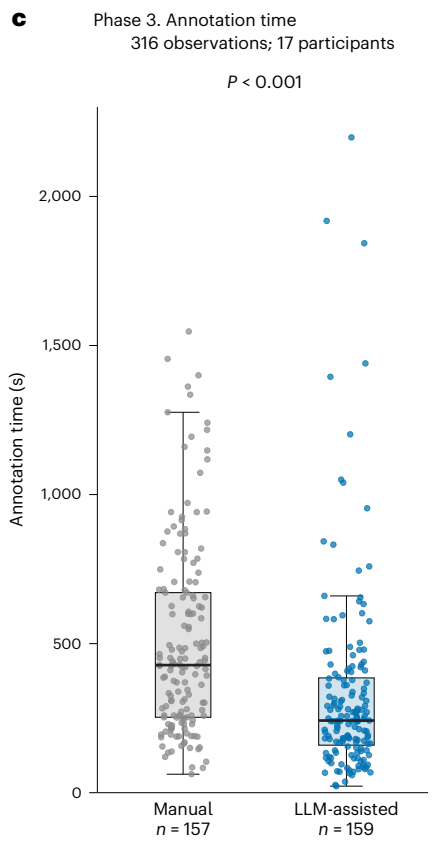
Discussion

The present study explores an agentic LLM system for the extraction of clinically salient attributes of irAEs from routine oncology documentation with performance that translates from retrospective benchmarking to prospective, real-world use and improved adverse event data quality. Across six high-priority irAEs, the system achieved strong note-level detection and clinician-stated attribution performance. Multi-class CTCAE grading was promising, with degradation at intermediate grades. Performance declined during a 3-month prospective silent deployment (binary $F_1 = 0.72$–0.79), underscoring the gap between retrospective evaluation and changing real-world conditions. In a randomized user effect study, agentic assistance reduced median annotation time by 43.5%, improved accuracy and improved interannotator

https://doi.org/10.1038/s43018-026-01264-6







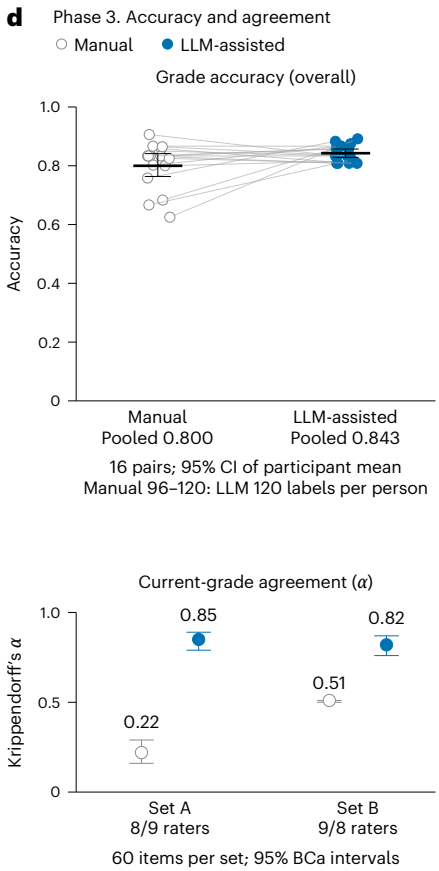


Fig. 3 | Performance and efficiency of AI-assisted irAE grading. a, Retrospective severity classification (phase 1, $n = 263$ notes) across irAE types. Points show F_1 scores for exact grade (blue), grade <2 versus ≥2 (orange), and grade <3 versus ≥3 (gray). **b**, Prospective validation (phase 2, $n = 884$ notes). Binary (blue) and macro (orange) F_1 scores combine grade, attribution and certainty labels across six conditions and two time points (36 assessments per note). **a** and **b** show aggregate performance estimates; no inferential tests are applied. **c**, Annotation time per note in the randomized user effect study (phase 3), comparing manual review (gray) with LLM-assisted review (blue). There are 316 participant-note observations (157 manual and 159 LLM assisted) from 17 participants reviewing 20 unique notes. Points show individual observations. Box plots display the median (center line) and the IQR (box bounds); whiskers extend to the most extreme observation within 1.5 × IQR of the box. Conditions were compared using a GEE with an exchangeable working correlation structure and robust sandwich variance estimators to account for within-participant clustering; the test was two-sided and was not adjusted for multiple comparisons. The primary raw-time GEE estimates a difference of −2.71 min (95% CI −3.89 to −1.53), $P = 6.98 \times 10^{-6}$. Median times were 428 s and 242 s (43.5% reduction). **d**, Overall grade accuracy and current-grade interannotator agreement (manual, gray; LLM-assisted, blue). Accuracy points and connecting lines show 16 paired participants; labels give pooled accuracy, and error bars show normal-approximation 95% intervals for participant means. Overall grade accuracy includes current and past labels. Agreement estimates use 60 note-condition items per set and 17 participants (manual/LLM-assisted raters: set A, 8/9; set B, 9/8). Reported 95% bias-corrected and accelerated bootstrap (BCa) intervals use 10,000 participant resamples. Accuracy is label-level agreement with the gold standard; interannotator agreement is Krippendorf's α for current grade labels, estimated separately for note set A and note set B. Overall grade accuracy: exact paired sign test, $P = 0.791$ (8 positive differences among 14 nonzero pairs; Table 3). Tests are two-sided and unadjusted for multiple comparisons. Interval overlap is not used as a significance test.

https://doi.org/10.1038/s43018-026-01264-6

Table 2 | Prospective deployment performance across 884 clinical notes

Category	Assessments	Notes	Binary F_1[a]	Macro F_1[b]
Overall	31,824	884	0.76	0.52
Colitis	5,304	884	0.74	0.65
Dermatitis	5,304	884	0.74	0.61
Hepatitis	5,304	884	0.79	0.59
Myocarditis	5,304	884	0.75	0.61
Pneumonitis	5,304	884	0.71	0.47
Thyroiditis	5,304	884	0.81	0.73

[a]Binary F_1 uses binarized grade, attribution and certainty labels across current and past time points. [b]Macro F_1 uses the full label categories for the same assessments. Each note contributes 36 assessments (six conditions, two time points and three label types).

reliability. Together, these results suggest that agentic LLM systems, coupled to human verification and evidence highlighting, could be further explored to shift irAE ascertainment from a labor-intensive retrospective review to scalable, efficient automated extraction with a potential to improve consistency and quality.

Prior approaches have largely focused on patient- or visit-level irAE classification, retrospectively and without the fine-grained attributes required for actionability and causal effects analysis (temporality, grade and attribution)[21,25–28]. One prior study evaluated simple prompting approaches on a subset of notes of 20 patients and reported macro F_1 scores of 0.50–0.57, lower than achieved by our agentic system[28]. However, as these evaluations were performed on entirely different datasets, such cross-study comparisons should be interpreted with caution and are not directly equivalent. Note-level classification is needed to support earlier diagnosis and management to improve outcomes and reporting of adverse events as they evolve over time, which is currently a major limitation for clinical trials and real-world evidence.

Further, the prior studies have focused on detecting the binary presence or absence of an irAE. Our work adds to prior literature by extracting not only presence but also temporality, grading, attribution and certainty with explicit extraction of supporting spans. These extracted attributes are critical for comprehensive phenotyping of irAEs for safety monitoring in both clinical and research settings, and therefore our system meaningfully adds to prior literature on automated adverse event classification and phenotyping[28,32].

The self-consistency mechanism, in which a judge agent reconciled three parallel model runs for a stage, contributed materially to detection performance (macro F_1 + 0.14 relative to the single-inference variant), consistent with the value of ensembling and cross-examination to reduce output variance[33,34]. Temporal misassignment, the system labeling an irAE as currently active when it was historical, or vice versa, was small ($|\Delta F_1| < 0.05$ across conditions). A cross-temporal analysis that credited a prediction if the irAE was detected in either timeframe yielded only modest gains (binary $F_1 = 0.93$), demonstrating that residual errors are primarily detection- or grading-related rather than due to timeframe confusion.

These results carry direct clinical and pharmacovigilance implications. Reliable, note-level extraction of irAE attributes addresses a persistent blind spot in structured EHR fields and claims data, where recall is limited even for hospitalization-associated events. The ability to surface candidate events with attributed causality and provisional grade shortly after documentation can shorten the time to recognition for life-threatening irAEs (for example, myocarditis and pneumonitis) and standardize ascertainment for research registries and safety reporting.

Multi-class CTCAE grading remains the most challenging subtask, which is unsurprising given the clinical context and nuance needed to grade irAEs. Grading requires a system to (1) identify which clinical findings documented in a note are actually sequelae of an irAE and (2) map those sequelae established grading criteria. In selected simple clinical presentations, this might be straightforward: thyroid replacement therapy is always an indicator of hypothyroidism (a subclass of thyroiditis) and is easily mapped to a grade 2 toxicity. However, take a simple extension of this case, where a patient has hypothyroidism treated with thyroid replacement therapy and is also admitted to the hospital. Here, the system must determine whether the hypothyroidism is causing or prolonging the admission to determine if it is a grade 2 or 3 event. Two additional considerations probably contribute to, and temper, these results. First, ambiguity at intermediate grades often reflects documentation incompleteness (for example, missing quantitative stool counts for grading of colitis and absent laboratory trend context for grading of hepatitis) rather than model error per se; such ambiguity similarly burdens human reviewers. Second, thresholds can be adapted to clinically meaningful bins (for example, none/low versus ≥ grade 3), which may improve effective performance for triage and reporting while preserving granular outputs for research. Incorporating structured signals (laboratories, medications such as corticosteroid initiation and imaging impressions) and temporal aggregation across notes are promising future strategies to disambiguate intermediate grades without sacrificing interpretability.

The transition from retrospective benchmarking to prospective deployment revealed expected but instructive performance attenuation. The system was optimized on a curated retrospective corpus of several hundred notes, then deployed against a continuous stream of real-world documentation generated years later on different EHR templates, encompassing a broader range of clinical presentations, irAE subtypes, symptom descriptions and documentation styles than were represented in the development set. This degradation, therefore, probably reflects the combined effects of distributional shift in case mix and clinical terminology, evolving EHR templates and note structures, and variation in documentation practices across clinician cohorts, rather than fundamental model limitations. The prospective phase thus underscores an operational consideration: deployed clinical AI systems require continuous monitoring and periodic recalibration[35], not one-time validation. Attribution and certainty extraction proved more robust to drift than severity grading, consistent with the relative stability of causal language conventions versus the context-dependent nature of quantitative symptom documentation.

The randomized user study provides controlled evidence that agentic AI assistance improves adverse event curation when deployed with human oversight. Three findings merit emphasis. First, efficiency gains were substantial, suggesting that AI-prepopulated labels shift annotator effort from generation to verification—a cognitively less demanding task. While efficiency gains from a mature LLM-assisted workflow are expected in principle, the contribution here is to quantify them under controlled, randomized conditions and to demonstrate concurrent improvements in accuracy and interannotator consistency—benefits that are not guaranteed a priori with AI assistance and efficiency gains. Second, there was a striking effect on interannotator consistency: manual annotation produced only fair agreement, whereas AI-assisted annotation achieved near-excellent reliability. This convergence effect probably arises because annotators anchor on model suggestions, reducing idiosyncratic interpretation of ambiguous documentation. For pharmacovigilance and registry curation, where consistency across reviewers is as important as individual accuracy, this standardization benefit may prove as valuable as time savings. However, the same anchoring mechanism raises the risk of automation complacency, whereby reviewers accept model outputs without sufficient independent scrutiny, potentially introducing systematic bias. Our study was not designed to disentangle consistency gains from complacency effects, and future work should explicitly test whether anchoring on AI suggestions compromises the detection

https://doi.org/10.1038/s43018-026-01264-6

Table 3 | HCI study comparing unassisted (control) and LLM-assisted irAE annotation

Section	Metric	Control	LLM-assisted	Improvement
Timing				
Overall	Median min/note (range)	7.13 (1.03 to 25.78)	4.03 (0.37 to 36.63)	43.5% reduction (P=6.98×10^{-6})
Accuracy (exact match versus gold)				
Grade (overall)	Proportion	0.800 (0.764 to 0.841)	0.843 (0.829 to 0.857)	0.043 (−0.006 to 0.086) (P=0.791)
Grade (current)	Proportion	0.763 (0.718 to 0.814)	0.809 (0.791 to 0.828)	0.046 (−0.014 to 0.101) (P=0.607)
Attribution (overall)	Proportion	0.864 (0.837 to 0.894)	0.928 (0.917 to 0.938)	0.064 (0.032 to 0.092) (P=3.05×10^{-5})
Attribution (current)	Proportion	0.841 (0.809 to 0.877)	0.933 (0.924 to 0.942)	0.092 (0.054 to 0.127) (P=6.10×10^{-5})
Certainty (overall)	Proportion	0.798 (0.766 to 0.835)	0.855 (0.847 to 0.862)	0.056 (0.021 to 0.087) (P=0.021)
Certainty (current)	Proportion	0.764 (0.718 to 0.818)	0.854 (0.842 to 0.866)	0.090 (0.042 to 0.131) (P=9.77×10^{-4})
Interannotator agreement (Krippendorff's α)				
Grade—note set A	α	0.22 (0.16 to 0.29)	0.85 (0.79 to 0.89)	+0.63
Grade—note set B	α	0.51 (0.50 to 0.51)	0.82 (0.76 to 0.87)	+0.31
Attribution—note set A	α	0.22 (0.07 to 0.35)	0.86 (0.82 to 0.90)	+0.64
Attribution—note set B	α	0.42 (0.33 to 0.49)	0.75 (0.66 to 0.82)	+0.33
Certainty—note set A	α	0.18 (0.13 to 0.22)	0.84 (0.77 to 0.88)	+0.66
Certainty—note set B	α	0.44 (0.25 to 0.52)	0.74 (0.65 to 0.81)	+0.30

Timing: medians (ranges) summarize 316 participant-note observations from 17 participants reviewing 20 unique notes (manual n=157; LLM-assisted n=159). The 43.5% reduction is calculated from the observed medians (428 s and 242 s). The P value is from the raw-time GEE with an exchangeable working correlation and participant-clustered robust variance; the model-estimated difference is −2.71 min (95% CI −3.89 to −1.53). Log-transformed time was analyzed separately as a sensitivity analysis. Accuracy analyses include 16 participants. Values are pooled label-level accuracy proportions; normal-approximation 95% intervals summarize participant means and paired differences. P values are from exact two-sided paired sign tests, excluding zero differences. Overall includes current and past labels; current includes current labels only. Agreement: Krippendorff's α and reported 95% bootstrap intervals are presented separately for note sets A and B. $\Delta\alpha$ is LLM-assisted minus manual. Interval overlap is not used as a significance test. All tests are two-sided and unadjusted for multiple comparisons.

of model errors, particularly for subtle or atypical presentations. Third, participants' qualitative feedback emphasized the value of highlighting the specific text spans extracted as evidence of an irAE in the clinical note (for example, the phrases describing respiratory symptoms that supported a pneumonitis determination) alongside each model output (Fig. 4). By linking each extracted label to its source evidence within the note, this feature enabled reviewers to more easily verify model conclusions against the original documentation directly, rather than re-reading entire notes, and served as a practical enabler of efficient human–AI collaboration.

These findings are consistent with a growing body of work suggesting that AI assistance can improve throughput in clinical documentation and annotation tasks, including ambient AI scribes that reduce documentation time by up to 30 min per clinician per day[36], and LLM-assisted radiology reporting that accelerates report generation while maintaining diagnostic accuracy[37]. The 88% preference for the AI-assisted workflow, coupled with improvements in perceived efficiency and confidence, suggests that such systems can achieve the often-elusive goal of user acceptance in clinical informatics.

Questions of generalizability, governance and cost warrant consideration. Interestingly, the architecture generalized across six irAEs with consistent detection performance, suggesting portability to a broader spectrum of adverse events. Nevertheless, this is a single-health-system study with local note templates and vernacular; external validation across institutions, EHRs and oncology subspecialties is necessary. Notably, all study phases—including the prospective silent deployment and the randomized user study—were conducted within this same healthcare system; consequently, no phase of this work establishes external validation, which remains the most important open question for real-world deployment. A practical barrier to adoption is cost and infrastructure. Agentic pipelines require multiple model calls per note, and our system currently relies on a proprietary model (GPT-4.1-mini) accessed via a commercial Application Programming Interface (API), which incurs per-token inference costs that may be prohibitive for small and medium-sized institutions lacking enterprise API agreements or dedicated compute infrastructure. In the present study, the best-performing agentic configuration cost US$0.021 per note (~US$18.50 for the 884-note prospective corpus), while open-source alternatives running on institutional graphics processing units (GPUs) eliminated API costs entirely, albeit with lower performance (Supplementary Table 1). Although our results with a mid-sized model suggest that performance–cost tradeoffs can be managed at scale, broader adoption will probably require validating open-source or locally deployable LLMs that can run behind institutional firewalls for maximum implementation across healthcare settings, reducing both cost and reliance on external vendors. Our preliminary benchmarking of open-source alternatives showed lower performance with current prompting strategies (Supplementary Table 1), but rapid improvements in open-weight models make this a tractable near-term goal. Robust monitoring for data drift will also be essential as indications and combination regimens expand[38].

Several limitations should be acknowledged. We focused on six irAEs representing a subset of the broader CTCAE v5.0 ontology; expansion to other toxicities will require targeted evaluation. These six subtypes were prioritized to represent a range of clinical severity and prevalence and do not capture the full heterogeneity of irAEs; evaluation on a broader range of other, more heterogeneous, or ambiguous presentations is an important direction for future work. While we compared retrospective performance on a range of LLMs, full validation using our primary system used a single proprietary model (GPT-4.1-mini). Performance may not transfer directly to other LLMs; systematic evaluation across open-source and locally deployable models is an important next step for broader accessibility. Ground-truth labels, although adjudicated, remain constrained by documentation quality and the inherent subjectivity of grading and attribution. The prospective deployment did not assess patient outcomes (for example, time to steroid initiation, hospitalization and treatment interruption). The preparatory field study was observational; causal effects on efficiency and agreement were isolated in the randomized study (phase 3). While we only studied irAEs, we believe that the results will be

https://doi.org/10.1038/s43018-026-01264-6

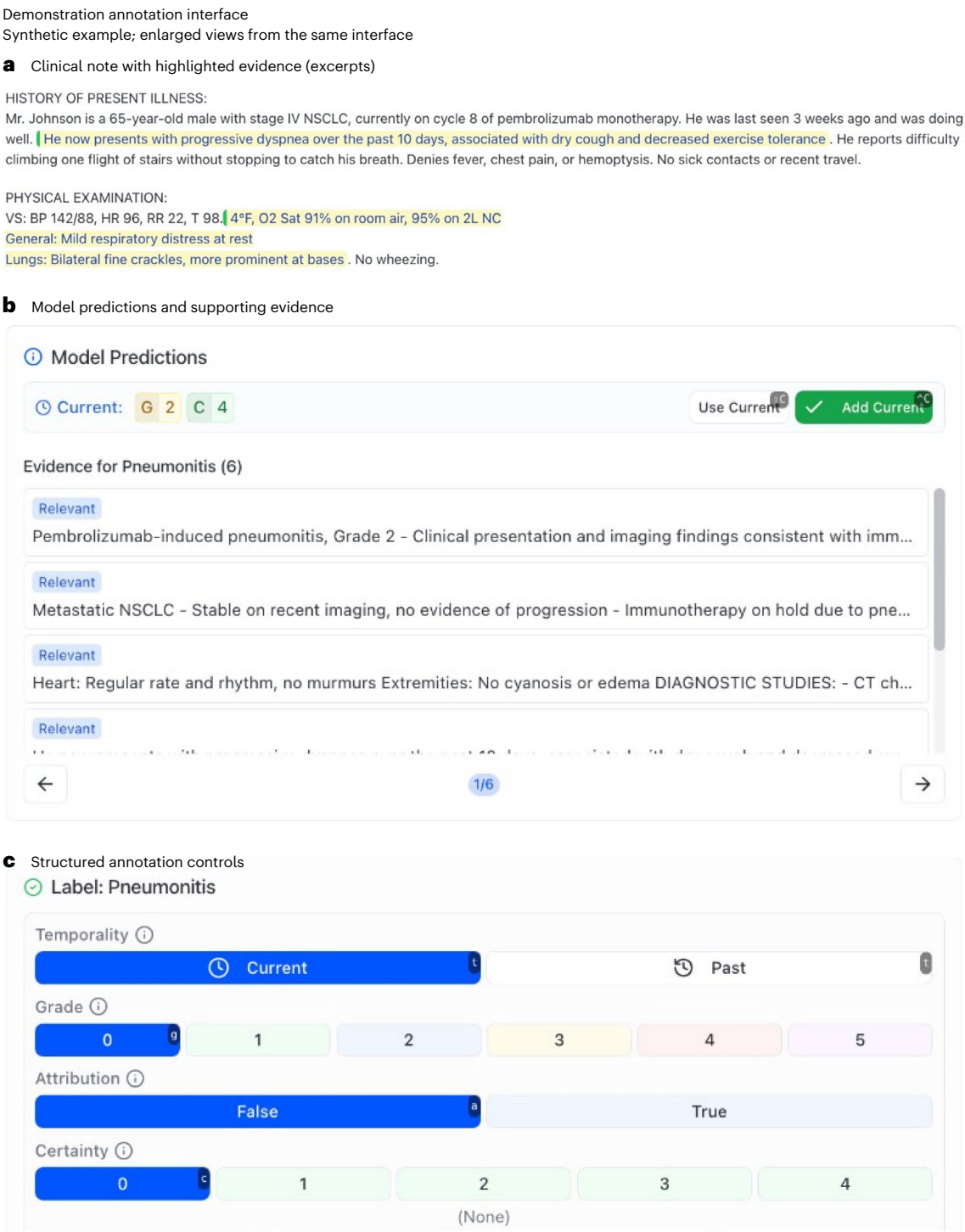


Fig. 4 | Annotation interface for irAE chart curation. a–c, Enlarged views from the same interface show a synthetic demonstration note with highlighted evidence (**a**), model predictions and supporting evidence (**b**) and structured annotation controls (**c**).

informative for other cancer adverse events and adverse events more broadly. In particular, the HCI results may be transferable for systems with similar performance and interface features.

Looking ahead, these findings carry both methodological and practical implications. Methodologically, agentic decomposition with self-consistency and explicit evidence extraction yielded accurate, auditable outputs suitable for clinical workflows. Practically, staging validation from retrospective testing to silent prospective monitoring and embedded human collaboration provides a template for closing the 'deployment gap' in clinical AI. Given the expanding survivor population exposed to ICIs and diffusion of care into nononcology settings, scalable, auditable assistance for irAE detection is poised to become a core informatics capability for oncology services and pharmacovigilance programs.

In summary, an agentic LLM system and associated interactive interface, validated across retrospective and prospective settings, can extract real-time, clinically detailed information about the presence, severity and attribution of irAEs. Agentic assistance improved the efficiency, performance and consistency of irAE curation when used with a human-in-the-loop. With external validation and outcome-focused

trials, such systems could enhance and accelerate real-time safety monitoring, trial curation and high-reliability oncology care.

Methods

Ethics statement

This research complies with all relevant ethical regulations. The study protocol received ethical approval from the Institutional Review Boards of Mass General Brigham (MGB) and Dana-Farber (DF)/Harvard Cancer Center (HCC; MGB 2022P002195, DF/HCC 24-735 and DF/HCC 20-328). A waiver of informed consent was granted by MGB and DF/HCC for the secondary use components, as the research involved minimal risk and did not alter any clinical care. Electronic informed consent was obtained for the HCI studies.

Note-level irAE system development and evaluation (phase 1)

For the data source and study population, we assembled a retrospective cohort of patients treated with ICIs at an academic medical center in Boston, Massachusetts, between 2015 and 2024. The inclusion criteria were: (1) age ≥18 years; (2) treatment with one or more of the following Food and Drug Administration-approved ICIs: tremelimumab, relatlimab, retifanlimab, pembrolizumab, nivolumab, atezolizumab, durvalumab, avelumab, cemiplimab, toripalimab and dostarlimab; and (3) the presence of at least one progress note or visit note in the clinical system. The number, age and sex of the patients whose notes were included in each phase are presented in Supplementary Table 5; sex was recorded as documented in the EHR.

To establish the gold-standard annotation procedure, annotation guidelines were developed to support manual ground-truth labeling of the presence of current and past irAEs and their key clinical attributes (see annotation guidelines in the study protocol)[31]. We focused on capturing six irAEs, representing the six most common and/or lethal irAEs and selected to capture 70–80% of clinically-relevant irAEs: pneumonitis, hepatitis, myocarditis, dermatitis, thyroiditis and colitis[30]. Each irAE was mapped to a set of National Cancer Institute CTCAE terms following the definitions used in the Alliance for Clinical Trial in Oncology irAE biobank (NCT04242095; see annotation guidelines in the study protocol)[31]. For each irAE type and temporality (current versus past) combination, we labeled its binary presence or absence, CTCAE grade, attribution and certainty as described in a note. The clinical attributes are defined as follows. (1) Severity: severity grading followed the National Cancer Institute CTCAE version 5.0, a standardized system commonly used for regulatory reporting. The values comprise grade 1, 2, 3, 4 and 5. (2) Attribution: attribution captured the clinician's explicitly documented assessment of causality between the adverse event and ICI therapy, recognizing that many symptoms in patients with cancer have multiple potential etiologies. The attributes comprise positive and negative. (3) Certainty: certainty reflected the degree of diagnostic confidence expressed in the documentation, guided in part by the Naranjo scale but adjusted for note-level, text-based determination[39]. The attributes comprise none, unlikely, possible, probable and definite.

Pilot rounds of 10–20 notes were dual-annotated by a physician and CRC with 4 years of oncology data abstraction experience, and annotation guidelines were iteratively updated for clarity and consistency until interannotator agreement exceeded 0.70 for each irAE type. All discrepancies were tracked and resolved through consensus discussion. Cases requiring adjudication were reviewed by an attending oncologist. The open-source Label Studio Community Platform v1.15 (HumanSignal), locally hosted behind our institution's firewall, was used as our annotation platform.

A sample of notes from 100 patient encounters, forming a total of 289 retrospective patient notes, was annotated for retrospective evaluation (Supplementary Table 5). These notes were divided into a development set of 26 notes, all from pneumonitis-positive patients, and a test set of 263 notes, used to evaluate the final developed system.

For the development of the agentic LLM architecture, OpenAI API-based models were accessed using our institutional HIPAA-compliant application programming inference access points. Open-source models were downloaded locally, behind our institutional firewall, to a dedicated server equipped with two NVIDIA RTX 6000 Ada GPUs.

The architecture used an agentic design pattern, decomposing the complex task of irAE extraction into specialized subtasks handled by distinct model instances (Fig. 2). A preprocessing agent first extracted and formatted relevant clinical text, handling the varied structures and formats present in EHR documentation. Subsequently, specialized extraction agents focused on specific attributes: one agent determined temporal status by analyzing verb tenses and temporal markers, another mapped clinical descriptions to CTCAE grades, a third assessed attribution by identifying causal language and clinical reasoning, and a fourth evaluated the certainty of diagnosis on the basis of hedging language and diagnostic workup discussions. We refer to this as the 'default' agentic architecture hereafter. Prompt engineering was optimized on the development set, and only for pneumonitis, to assess the generalizability of our approach to other irAE types.

Given the potential variability in LLM outputs and the risk of hallucination, we also evaluated adding a self-consistency mechanism[33,34,38,40]. Here, each note was processed through three parallel inference runs with controlled randomness. A final judge agent then synthesized these parallel outputs, implementing a majority-vote mechanism for discrete categories and selecting the most coherent extraction when unanimity was absent.

Finally, on the basis of preparatory field study work, the models were designed to extract 'string-matched' snippets directly from clinical notes, as well as a brief rationale for the determinations, to support human interaction in real-world downstream tasks. The snippets serve as supporting evidence for the determinations made at each stage of the agentic workflow. This functionality was then integrated into the user interface as features to aid human reviewers in their annotation tasks.

As comparative baseline methods, to contextualize the performance of our agentic system, we implemented several baseline approaches representing different points on the complexity–performance spectrum. These included rule-based extraction using regular expressions expanded from a set of terms previously used in our institution to identify patients admitted for irAEs[41] and zero-shot prompting. We compared open-source (Qwen-3-4B-Instruct, Qwen3-30B-A3B-Instruct-2507, Gemma-3-27B and GPT-OSS-20B) and proprietary OpenAI models (GPT-4.1-mini, GPT-4.1-nano and gpt-4o-mini) with medium reasoning effort on our institutional HIPAA-secure OpenAI API end point. Model selection was guided by the goal of representing a range of model sizes deployable on single- and dual-GPU institutional hardware for open-source candidates, alongside cost-constrained proprietary models available under Azure HIPAA-compliant agreements.

Precision, recall and macro-averaged F_1 scores were calculated overall and stratified by irAE type across all available classes for a given label. The inference cost per note was also calculated for all variants.

Prospective silent validation (phase 2)

Following retrospective analysis, we implemented the best setting of our system in a silent, prospective, real-time clinical environment to assess performance under operational conditions and to assess generalizability to evolving clinical language and documentation practices. The system was integrated with our institution's clinical data warehouse. Details of the integration infrastructure have been previously reported[38]. New physician, nursing, and physician assistant notes documented on patients treated with an ICI within the preceding year were automatically identified each morning, including progress notes, radiology reports, consultation reports, emergency department records and discharge summaries. Notes meeting criteria were immediately processed

through the agentic system. The first 102 notes were dual-annotated by a board-certified oncologist and a trained research assistant with experience as a medical scribe and clinical research, achieving an interannotator agreement across irAE type and temporality ($\kappa = 0.91$ for binary detection, $\kappa = 0.88$ for grade). The remaining notes were single-annotated by the research assistant. This prospective phase lasted for 3 months (May–July 2025), during which we processed 884 real-time clinical notes. Agreement between prospective model predictions and human determinations was quantified using agreement percentages and override rates for each label and across all labels.

Preparatory field study

Before the randomized user effect study (phase 3), we conducted a preparatory field study to refine the annotation interface (Fig. 4), establish baseline efficiency and agreement metrics for power calculations and assess the temporal stability of human–AI collaboration. We developed a web application (React 18.2 frontend and FastAPI backend) integrating the agentic system into the annotation workflow and iteratively refined it through think-aloud sessions with three CRCs; key features included evidence highlighting within the note viewer, keyboard shortcuts and prepopulation of structured fields with model suggestions, with audit trails capturing interactions, timings and modifications. Two CRCs then used the application over 3 months to assist chart curation for a separate retrospective study, with annotation time, interannotator agreement and human–model agreement tracked over time and reported descriptively.

Across 299 dual-annotated notes over six sequential (approximately weekly) periods, inter-rater agreement on grade was substantial overall (Cohen's $\kappa = 0.80$) and largely stable over time (a transient dip to $\kappa = 0.55$ for notes 101–150 rebounded to $\kappa = 0.71$–0.93), and agreement with the model on grade was high and stable (mean $\kappa = 0.89$). Time per note averaged 161 s versus an estimated 600-s unaided baseline (~3.7× faster), with efficiency improving over time and between-annotator variability narrowing, consistent with a learning curve. These findings informed the phase 3 design: annotation time was selected as the primary efficiency end point; the observed effect sizes and variance informed the power calculation; evidence highlighting and keyboard navigation were retained; and a 1-h training session was established as sufficient. Field study summary statistics are presented in Supplementary Table 6.

Randomized user effect study (phase 3)

To quantify the clinical impact of agentic assistance on irAE curation, we conducted a 2 × 2 randomized crossover study comparing AI-assisted annotation to standard manual review (Extended Data Fig. 2). This design was informed by a preparatory field study that informed the interface design, characterized learning curves, established baseline inter-rater agreement and identified annotation efficiency as the primary driver of real-world workflow burden. This study was approved by the institutional review board (DF/HCC 20-328).

We enrolled CRCs and CRNs from our department; there were no exclusion criteria. Following informed consent, participants completed a 1-h standardized training session with annotated examples[31] and an unobserved practice period using the annotation interface (Fig. 4). Participants were then randomized to one of four sequences in a crossover design that counterbalanced both intervention order (AI-assisted versus manual spreadsheet) and note set (set A versus set B). This structure was chosen to mitigate carryover effects from either the annotation method or note content, and enabling difference-in-differences estimation of treatment effects. A total of 17 participants completed the study. Participant characteristics, including professional role, years of oncology data-abstraction experience, attitudes toward AI and experience with AI are presented in Supplementary Table 3.

Clinical notes were drawn from the phase 1 retrospective corpus. Sets A and B were matched on note length and clinical complexity and number of irAE descriptions through manual review to ensure comparability across arms.

Sample-size planning was informed by the preparatory field study. After inflating the sample size to account for the within-participant correlation of 0.05, 15 participants completing 20 notes each (300 total notes) would provide 84.7% power to detect a standardized effect size of −0.43 (comparing AI assistance to manual) with two-sided $\alpha = 0.10$. We targeted the enrollment of 17–20 participants to account for potential dropouts and missing observations.

The primary end point was annotation efficiency, measured as time in seconds from note opening to final submission, captured via screen recording. The secondary end points were: (1) accuracy, defined as an exact match with the gold standard for all current irAE grades within a note—a stringent metric requiring perfect concordance across all six conditions; and (2) interannotator agreement, assessed using Krippendorff's α for ordinal grade labels.

For the primary efficiency end point, we used GEE with a Gaussian distribution, identity link, exchangeable working correlation and participant-clustered robust variance to estimate the difference in untransformed annotation time between conditions. Log-transformed time was analyzed as a sensitivity analysis; separate models adjusted for note set and sequence.

For complete-note accuracy (340 annotations from 17 participants; 170 per condition), we used a GEE with binomial outcomes to estimate the odds ratio of the exact gold-standard match comparing AI-assisted with manual annotation, with robust sandwich variance estimators. Exploratory label-level accuracy analyses used paired data from 16 participants: 312 of 320 note annotations could be linked to gold-standard records; 8 annotations without matching records were excluded. Point estimates pooled labels by condition. Normal-approximation 95% intervals summarized participant means and paired differences. Comparisons used exact two-sided paired sign tests, excluding zero differences. Further, interannotator reliability was assessed separately within each note set using Krippendorff's α for current grade labels. Bootstrap resampling (10,000 iterations, resampling participants with replacement) generated 95% confidence intervals (CIs) for each method. Questionnaires assessed usability, perceived efficiency and AI attitudes[31]. Survey analyses included 17 complete poststudy respondents and excluded five records with missing poststudy responses. Paired System Usability Scale[42] scores were compared with a two-sided paired t-test; efficiency and label-confidence ratings used two-sided Wilcoxon signed-rank tests with a normal approximation, average ranks for ties, zero differences excluded and no continuity correction. Differently worded pre-/post-AI-capability questions were summarized descriptively. Participants were compensated US$50 h^{-1}.

Statistics and reproducibility

No statistical method was used to predetermine sample size for the retrospective evaluation (phase 1) or the prospective silent deployment (phase 2). The phase 1 sample size was informed by sample sizes in previous studies and the capacity for manual double annotation. For the randomized user effect study (phase 3), sample size was determined a priori by power analysis using pilot estimates from the preparatory field study, as described above.

Robustness was instead assessed in three ways: through the self-consistency mechanism, in which the detection and grading stages were run three times independently per note and reconciled by a judge agent; through evaluation of the same fixed system in three nonoverlapping settings (retrospective, prospective and randomized user effect); and through bootstrap resampling (10,000 iterations, resampling participants with replacement) for interval estimation of agreement statistics as described above.

Notes were not randomized in phases 1 and 2. In phase 3, participants were randomized by computer-generated allocation to one of

four sequences in a 2 × 2 crossover design counterbalancing intervention order and note set. Data collection and analysis were not performed blinded to the conditions of the experiments. Collectively, 24 note-level observations across participants were excluded owing to missing video recordings.

Data distribution was assumed to be normal but this was not formally tested; the primary efficiency end point was additionally analyzed after log transformation as a sensitivity analysis, and nonparametric summaries (medians and IQRs) are reported alongside model-based estimates. All statistical tests were two-sided, and no adjustments were made for multiple comparisons. Statistical analyses used R 4.1.0, and the computational pipeline used Python 3.10. The annotation interface was built with React 18.2 and FastAPI. Final statistical reproduction used R 4.5.3 with geepack 1.3.13, irr 0.85 and boot 1.3-32. The accompanying analysis code records the full reproduction environment; Python dependency requirements are provided in the AEGIS repository (see the 'Code availability' section)[43].

The age and sex of the patients whose notes were analyzed in phases 1 and 2 are presented in Supplementary Table 5. These characteristics were recorded as documented in the EHR. Age and sex were not considered in the study design. Age and sex were not collected for phase 3 clinical research staff participants.

Further information on the research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Study protocols

The annotation guidelines developed and used for gold standard labeling of irAEs in clinical notes can be used to replicate labeling procedures used in this work. These guidelines, together with the training procedures and surveys used in the randomized user effects study, are available in the cited protocols.io collection[31].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The clinical notes and patient-level data that support the findings of this study were generated at Mass General Brigham and Dana-Farber Cancer Institute and are not publicly available. These data are free-text clinical documentation containing protected health information that cannot be de-identified without impacting the linguistic content that is the object of study, precluding their public release to protect patient privacy. Requests for access to additional data may be directed to the corresponding author (D.S.B., dbitterman@bwh.harvard.edu); requests will be acknowledged within 3 weeks. Good faith efforts will be made to explore data access, contingent on an executed institutional data-use agreement and evidence of appropriate institutional review board approval. Source data are provided with this paper.

Code availability

All custom code and prompts used to build and evaluate the agentic pipeline, the annotation interface and the statistical analyses are publicly available via GitHub at https://github.com/BittermanLab/AEGIS under the MIT license[43]. Proprietary models (GPT-4.1 and GPT-4.1-mini) were accessed through an institutional HIPAA-compliant API; open-source model weights are available from their original providers as cited in the Methods. No other custom code or mathematical algorithm was used in this study.

References

1. Martins, F. et al. Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance. *Nat. Rev. Clin. Oncol.* **16**, 563–580 (2019).
2. Reynolds, K. L. et al. Immune-related adverse events associated with immune checkpoint inhibitors: a call to action for collecting and sharing clinical trial and real-world data. *J. Immunother. Cancer* **9**, e002896 (2021).
3. Golder, S., Loke, Y. K., Wright, K. & Norman, G. Reporting of adverse events in published and unpublished studies of health care interventions: a systematic review. *PLoS Med* **13**, e1002127 (2016).
4. Hazell, L. & Shakir, S. A. W. Under-reporting of adverse drug reactions: a systematic review. *Drug Saf.* **29**, 385–396 (2006).
5. Derry, S., Loke, Y. K. & Aronson, J. K. Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. *BMC Med. Res. Methodol.* **1**, 7 (2001).
6. Ioannidis, J. P. A. & Lau, J. Improving safety reporting from randomised trials. *Drug Saf.* **25**, 77–84 (2002).
7. Ioannidis, J. P. & Lau, J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *J. Am. Med. Assoc.* **285**, 437–443 (2001).
8. Jha, A. K. et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *J. Am. Med. Inform. Assoc.* **5**, 305–314 (1998).
9. Center for Drug Evaluation and Research. Safety labeling update for capecitabine and fluorouracil (5-FU) on risks associated with dihydropyrimidine dehydrogenase (DPD) deficiency. *US Food and Drug Administration* https://www.fda.gov/drugs/resources-information-approved-drugs/safety-labeling-update-capecitabine-and-fluorouracil-5-fu-risks-associated-dihydropyrimidine (2026).
10. Mandelblatt, J. S. et al. Preliminary development and evaluation of an algorithm to identify breast cancer chemotherapy toxicities using electronic medical records and administrative data. *J. Oncol. Pract.* **11**, e1–e8 (2015).
11. Kalinich, M. et al. Prediction of severe immune-related adverse events requiring hospital admission in patients on immune checkpoint inhibitors: study of a population level insurance claims database from the USA. *J. Immunother. Cancer* **9**, e001935 (2021).
12. Postow, M. A., Sidlow, R. & Hellmann, M. D. Immune-related adverse events associated with immune checkpoint blockade. *N. Engl. J. Med.* **378**, 158–168 (2018).
13. Brahmer, J. R. et al. Management of immune-related adverse events in patients treated with immune checkpoint inhibitor therapy: American Society of Clinical Oncology clinical practice guideline. *J. Clin. Oncol.* **36**, 1714–1768 (2018).
14. Schneider, B. J. et al. Management of immune-related adverse events in patients treated with immune checkpoint inhibitor therapy: ASCO guideline update. *J. Clin. Oncol.* **39**, 4073–4126 (2021).
15. Yeoh, H.-L. et al. Immune-related adverse events secondary to immunotherapy in oncology: a guide for general practice. *Aust. J. Gen. Pract.* **52**, 378–385 (2023).
16. Jayathilaka, B., Mian, F., Franchini, F., Au-Yeung, G. & IJzerman, M. Cancer and treatment specific incidence rates of immune-related adverse events induced by immune checkpoint inhibitors: a systematic review. *Br. J. Cancer* **132**, 51–57 (2025).
17. Couey, M. A. et al. Delayed immune-related events (DIRE) after discontinuation of immunotherapy: diagnostic hazard of autoimmunity at a distance. *J. Immunother. Cancer* **7**, 165 (2019).
18. Tonorezos, E. et al. Prevalence of cancer survivors in the United States. *J. Natl Cancer Inst.* **116**, 1784–1790 (2024).
19. George, S. et al. The impact of adverse events on health care resource utilization, costs, and mortality among patients treated with immune checkpoint inhibitors. *Oncologist* **26**, e1205–e1215 (2021).

20. Gallifant, J., Celi, L. A., Sharon, E. & Bitterman, D. S. Navigating the complexities of artificial intelligence-enabled real-world data collection for oncology pharmacovigilance. *JCO Clin. Cancer Inform.* **8**, e2400051 (2024).
21. Sun, V. H. et al. Enhancing precision in detecting severe immune-related adverse events: comparative analysis of large language models and International Classification of Disease codes in patient records. *J. Clin. Oncol.* **42**, 4134–4144 (2024).
22. Hsiehchen, D., Watters, M. K., Lu, R., Xie, Y. & Gerber, D. E. Variation in the assessment of immune-related adverse event occurrence, grade, and timing in patients receiving immune checkpoint inhibitors. *JAMA Netw. Open* **2**, e1911519 (2019).
23. Davidson, D. C. et al. NCI MATCH (EAY131) arm Z1D: retrospective evaluation of attributions of adverse events experienced on nivolumab in patients with MSI-high tumors. *J. Clin. Oncol.* **42**, e14711 (2024).
24. Savova, G. K. et al. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer Res.* **79**, 5463–5470 (2019).
25. Gupta, S., Belouali, A., Shah, N. J., Atkins, M. B. & Madhavan, S. Automated identification of patients with immune-related adverse events from clinical notes using word embedding and machine learning. *JCO Clin. Cancer Inform.* **5**, 541–549 (2021).
26. Zitu, M. M. et al. Detection of patient-level immunotherapy-related adverse events (irAEs) from clinical narratives of electronic health records: a high-sensitivity artificial intelligence model. *Pragmat. Obs. Res.* **15**, 243–252 (2024).
27. Barman, H. et al. Identification and characterization of immune checkpoint inhibitor-induced toxicities from electronic health records using natural language processing. *JCO Clin. Cancer Inform.* **8**, e2300151 (2024).
28. Bejan, C. A. et al. irAE-GPT: leveraging large language models to identify immune-related adverse events in electronic health records and clinical trial datasets. *eBioMedicine* **0**, 106227 (2026).
29. Zhang, J. et al. AFlow: automating agentic workflow generation. In *International Conference on Learning Representations* Vol. 2025 (eds Yue, Y., Garg, A., Peng, N., Sha, F. & Yu, R.) 34040–34077 (2025).
30. Brahmer, J. R. et al. Society for Immunotherapy of Cancer (SITC) clinical practice guideline on immune checkpoint inhibitor-related adverse events. *J. Immunother. Cancer* **9**, e002435 (2021).
31. Procedures for 'An agentic AI system enhances clinical detection of immune-related adverse events in a multi-phase validation study.' *protocols.io* https://doi.org/10.17504/protocols.io.e6nvwxr3zgmk/v1 (2026).
32. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
33. Cohen, R., Hamri, M., Geva, M. & Globerson, A. LM vs LM: detecting factual errors via cross examination. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H., Pino, J. & Bali, K.) 12621–12640 (Association for Computational Linguistics, 2023).
34. Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. In *Proc. 11th International Conference on Learning Representations* https://openreview.net/pdf?id=1PL1NIMMrw (2023).
35. Davis, S. E., Greevy, R. A. Jr, Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).
36. Tierney, A. A. et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal. Innov. Care Deliv.* https://doi.org/10.1056/CAT.23.0404 (2024).
37. Huang, J. et al. Efficiency and quality of generative AI-assisted radiograph reporting. *JAMA Netw. Open* **8**, e2513921 (2025).
38. Gallifant, J. et al. Beyond the algorithm: a field guide to deploying AI agents in clinical practice. Preprint at *SSRN* https://doi.org/10.2139/ssrn.5732262 (2025).
39. Adverse drug reaction probability scale (Naranjo) in drug induced liver injury. in *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury* https://www.ncbi.nlm.nih.gov/books/NBK548069/ (National Institute of Diabetes and Digestive and Kidney Diseases, 2012).
40. Abbasian, M., Azimi, I., Rahmani, A. M. & Jain, R. Conversational health agents: a personalized large language model-powered agent framework. *JAMIA Open* **8**, ooaf067 (2025).
41. Abu-Shawer, O. et al. Novel platform leveraging electronic medical record (EMR) to triage patients admitted with high-grade immune-related adverse events (irAEs) to the immune-toxicity (ITOX) service. *J. Immunother. Cancer* **8**, e000992 (2020).
42. Sauro, J. Measuring usability with the System Usability Scale (SUS). *Measuring U* https://measuringu.com/sus (2011).
43. Gallifant, J. & Bitterman, D. AEGIS code. *Zenodo* https://doi.org/10.5281/zenodo.22981488 (2026).

Acknowledgements

We thank the clinical research coordinators and clinical research nurses who participated in the preparatory field study and the randomized user effect study, and the Mass General Brigham enterprise data and digital teams for supporting the prospective deployment.

Author contributions

J. Gallifant conceptualized the study, developed the methods and software, conducted validation and formal analysis, created visualization, performed the interviews, led the overall investigation and wrote the first draft of the manuscript. S.C. contributed to study design, methods and software development; creation of the visualizations; and formal analysis. K.-Y.S. contributed to the study design and led the statistical analysis. K.C.K. contributed to the study design, led the interviews, analyzed the results of the interviews and provided critical review and revision of the manuscript. P.F.D. and J. Guo contributed to the study design, data curation and project administration. B.Y. contributed to the study design, methods development, software development and validation. A.W. contributed to the data curation and project administration. B.K.Z. contributed to the data curation and project administration. M.J.H contributed to the study design, interpretation of findings and critical review and revision of the manuscript. A.G., B.R., D.C.C. and H.J.W.L.A. contributed to the study design, provided datasets, interpretation of findings and critical review and revision of the manuscript. B.H.K., R.H.M., T.L.N., P.N., J.D.S. and U.T. contributed to the study design, interpretation of findings and critical review and revision of the manuscript. P.C. oversaw the statistical analysis and contributed to the study design, interpretation of findings and critical review and revision of the manuscript. H.H. contributed to the study design, HCI methodology, interpretation of findings and critical review and revision of the manuscript. J.L.W., E.S., D.E.K. and G.K.S. contributed to the study conceptualization, funding acquisition, study design, interpretation of findings and critical review and revision of the manuscript. D.S.B. conceptualized the study, acquired funding, oversaw the overall conduct of the study from design through interpretation, provided supervision and provided critical review and revision of the manuscript. All authors have reviewed and approved the final manuscript.

Funding

The authors acknowledge financial support from the National Institutes of Health National Cancer Institute (U54CA274516-01A1 to J.

Gallifant, D.S.B, D.E.K. and G.K.S.; R01CA294033-01 to J. Gallifant, P.F.D., B.Y., M.H., J.L.W, M.J.H., E.S., D.E.K., G.K.S. and D.S.B.; 2U24CA248010 to D.S.B., G.K.S., H.H. and J.L.W.; and U01209414 to D.C.C. and R.H.M.); the American Cancer Society and American Society for Radiation Oncology, grant no. ASTRO-CSDG-24-1244514-01-CTPS (https://doi.org/10.53354/ACS.ASTRO-CSDG-24-1244514-01-CTPS.pc.gr.222210) to D.S.B.; a Patient-Centered Outcomes Research Institute (PCORI) Project Program Award (ME-2024C2-37484) to D.S.B.; and the Woods Foundation to D.S.B. All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

D.S.B.: associate editor for *JCO Clinical Cancer Informatics*, associate editor for *Annals of Oncology*, associate editor for Radiation Oncology of HemOnc.Org (not related to the submitted work); is on the Scientific Advisory Board of Mercurial AI and Blue Clay Health LLC; consulting fees from Inspire Exercise Medicine LLC. D.E.K.: consultant for AstraZeneca, Exact Sciences and Genentech/Roche. J.L.W.: research funding (AACR, Brown University, Brown University Health), editor-in-chief of *JCO Clinical Cancer Informatics*, consultant for Westat, The Lewin Group, Nemesis Health and UT Medical Branch; ownership of HemOnc.org LLC (not related to the submitted work). H.J.W.L.A.: consulting fees and/or stock from Onc.AI, Love Health, Sphera, Health-AI, Ambient and AstraZeneca (all not related to the submitted work). B.H.K.: consulting (GE Healthcare), advisory board (Day One Biopharmaceuticals), researching funding (National Institutes of Health (NIH) and St. Baldrick's Foundation), all unrelated to the submitted work. R.H.M.: advisory board (ViewRay and AstraZeneca), consulting (Varian Medical Systems and Pfizer), honorarium (Novartis, Springer Nature and American Society of Radiation Oncology), research funding (National Institutes of Health, ViewRay, AstraZeneca, Siemens Medical Solutions USA, Inc. and Varian Medical Systems). M.J.H.: consulting (Replimmune, Deloitte, Guidepoint and GLG). J.D.S.: research funding (Merck, BMS, Regeneron, Debiopharm, Siemens and EMD Serono), consulting (IntraGel, Merck KGA, EMD Serono, Lantheus and GSK), scientific advisory board (Aveta Biomics and Immunitas); expert witness fees (Burns and White, Morgan and Morgan, Sacchetta and Baldino, and Offutt Simmons Simonton), all not related to the submitted work but are included for comprehensiveness and transparency.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-026-01264-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-026-01264-6.

Correspondence and requests for materials should be addressed to Danielle S. Bitterman.

Peer review information *Nature Cancer* thanks Jakob Kather and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

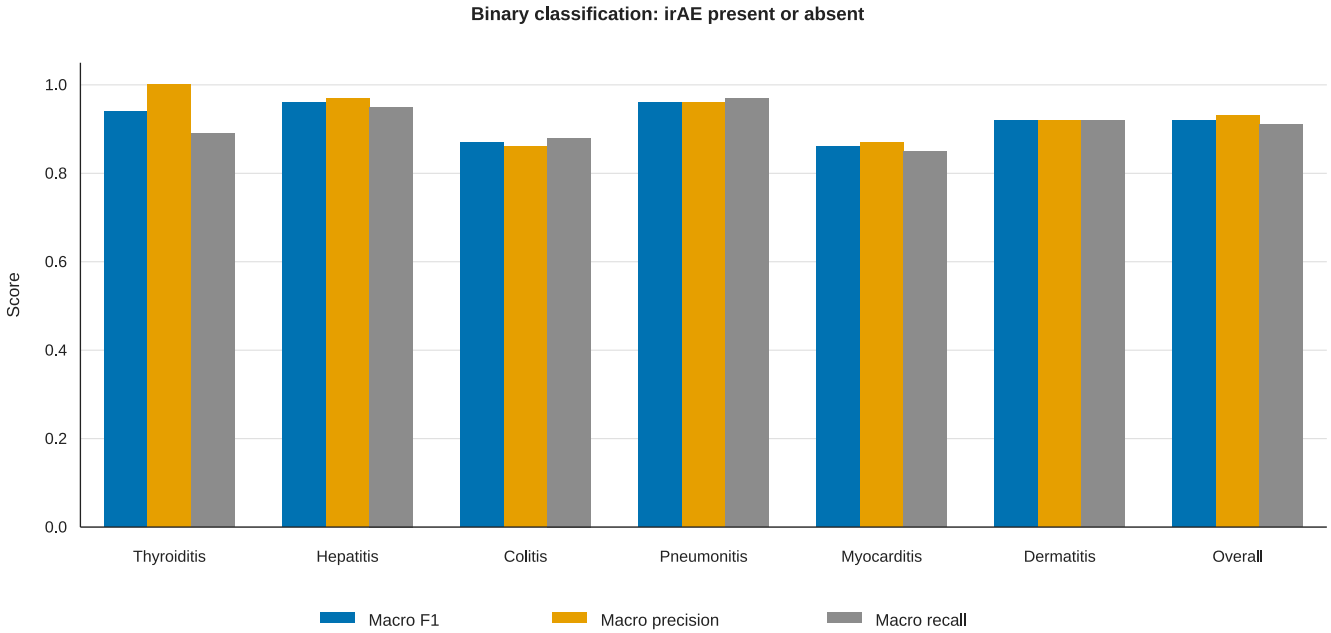
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

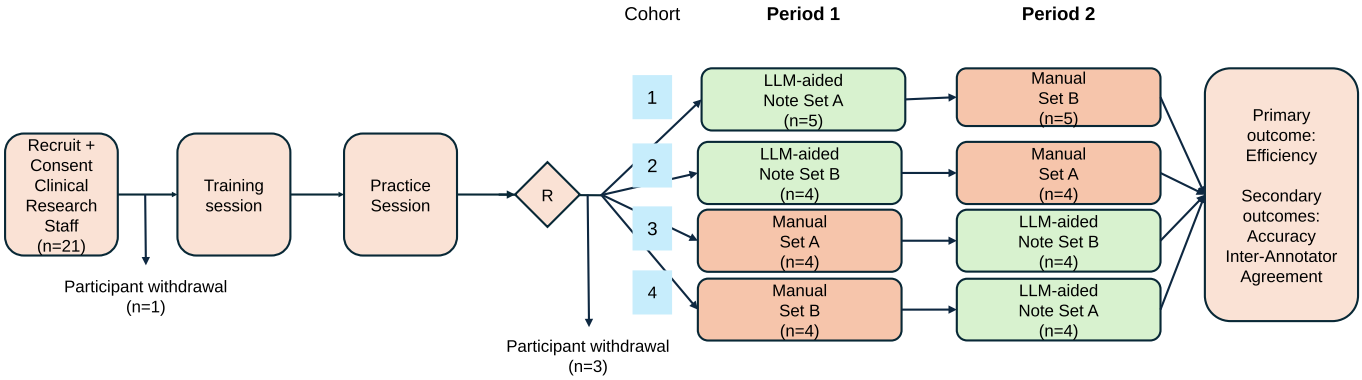
[1]Mass General Brigham, Boston, MA, USA. [2]Harvard Medical School, Boston, MA, USA. [3]Dana-Farber Cancer Institute, Boston, MA, USA. [4]MIT Sloan School of Management, Boston, MA, USA. [5]Center for Clinical Cancer Informatics and Data Science, Legorreta Cancer Center, Brown University, Providence, RI, USA. [6]Brown University Health Cancer Institute, Rhode Island Hospital, Providence, RI, USA. [7]Maastricht University, Maastricht, the Netherlands. [8]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD, USA. [9]University of Pittsburgh, Pittsburgh, PA, USA. [10]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. ✉e-mail: dbitterman@bwh.harvard.edu

https://doi.org/10.1038/s43018-026-01264-6



Extended Data Fig. 1 | Condition-specific binary irAE detection performance in the retrospective evaluation. F1 score (blue), precision (orange) and recall (grey) for current binary irAE detection (Grade 0 versus Grade ≥ 1), using GPT-4.1-mini with self-consistency. Metrics are macro-averaged across the two classes for each of six irAE subtypes; Overall is the unweighted mean across conditions. n = 263 clinical notes in the held-out retrospective test set. Values are point estimates computed against the expert-derived gold standard; no inferential tests are shown.

https://doi.org/10.1038/s43018-026-01264-6



Extended Data Fig. 2 | Design of the randomized user effect study. Schematic of the 2 × 2 randomized crossover design (Phase 3). Seventeen clinical research staff were randomized to one of four sequences counterbalancing intervention order (LLM-assisted versus manual annotation) and note set (Set A versus Set B), so that every participant completed both conditions and both note sets. Each participant reviewed 20 unique notes across the two note sets. The timing analysis included 316 participant-note observations (157 manual, 159 LLM-assisted); 24 observations lacked video recordings. No inferential tests are shown in this schematic.

nature portfolio

Corresponding author(s): Danielle S. Bitterman

Last updated by author(s): Aug 30, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect the clinical data used in this study. The randomized user effect study collected irAE labels via the custom agentic system that is the focus of this manuscript, and for which code will be released upon acceptance.
Data analysis	R v4.1.0 and Python v3.13 were used for statistical analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data Availability

The clinical notes and patient-level data that support the findings of this study were generated at Mass General Brigham and Dana-Farber Cancer Institute and are not publicly available. These data are free-text clinical documentation containing protected health information that cannot be de-identified without impacting the

linguistic content that is the object of study, precluding their public release in order to protect patient privacy. The numerical data underlying Figures 3a–d, Extended Data Figure 1 and Tables 1–3 are provided as Source Data with this paper. Requests for access to additional data may be directed to the corresponding author (D.S.B., dbitterman@bwh.harvard.edu); requests will be acknowledged within 3 weeks. Good faith efforts will be made to explore data access, contingent on an executed institutional data-use agreement and evidence of appropriate Institutional Review Board approval. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Age and sex of the patients whose notes were analysed in Phases 1 and 2 are reported in Supplementary Table 4. Sex and gender were recorded as documented in the electronic health record for patients.
Reporting on race, ethnicity, or other socially relevant groupings	Phase 1 and 2: We report sex, race and ethnic group in Supplementary Table 4. These data are collected as documented in the electronic health records during routine clinical care. While not a primary focus of this manuscript. we include them to provide line demographic information for transparency into the potential generalizability of methods. Race and sex are not reported for Phase 3 participants. Information on role, experience, and attitidudes toward AI were collected on these participants as these are more likely to be relevant for the endpoints.
Population characteristics	We report age, sex, race group, ethnic group, and primary language in our patient cohorts. While not a primary focus of this manuscript. we include them to provide baseline demographic information for transparency into the potential generalizability of methods. We do not report age and sex for the clinical research staff participants in Phase 3.
Recruitment	Clinical research coordinators and research nurses were recruited for the randomized user effect study via emails to institutional listservs.
Ethics oversight	This research complies with all relevant ethical regulations. The study protocol received ethical approval from the Institutional Review Boards of Mass General Brigham and Dana-Farber/Harvard Cancer Center (MGB 2022P002195, DF/HCC 24-735, DF/HCC 20-328). A waiver of informed consent was granted by MGB and DF/HCC for the secondary use components, as the research involved minimal risk and did not alter any clinical care. Electronic informed consent was obtained for the human-computer interaction studies.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable for retrospective components (descriptive performance). For randomized user effect study: Based on pilot data from the preparatory field study, we estimated a clinically meaningful efficiency gain of 450 seconds per note (SD = 192 seconds). After inflating the sample size to account for the within-participant correlation of 0.05, 15 participants completing 20 notes each (300 total notes) would provide 84.7% power to detect a standardized effect size of -0.43 (comparing AI assistance to manual) with two-sided $\alpha = 0.10$. We targeted enrollment of 17–20 participants to account for potential dropouts and missing observations.
Data exclusions	Twenty-four note-level observations across participants were excluded due to missing video recordings.
Replication	Replication in the sense of repeated independent experiments was not applicable because the study analysed pre-existing clinical documentation rather than experimental samples. Robustness was instead assessed through the self-consistency mechanism (three independent inference runs per note for the detection and grading stages, reconciled by a judge agent), through evaluation of the same fixed system in three non-overlapping settings (retrospective, prospective and randomized user effect), and through bootstrap resampling (10,000 iterations) for interval estimation.
Randomization	Participants were randomized using randomization software.
Blinding	Blinding of participants and researchers was not possible because there is no way to blind them to the use (or not) of agentic AI assistance.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	N/A, this was not a clinical trial
Study protocol	N/A, this was not a clinical trial
Data collection	For the data source and study population used for system development (Phase 1), we assembled a retrospective cohort of patients treated with immune checkpoint inhibitors at an academic medical center in Boston, MA between 2015 and 2024. Inclusion criteria were: (1) age ≥18 years, (2) treatment with one or more of the following FDA-approved ICIs: tremelimumab, relatlimab, retifanlimab, pembrolizumab, nivolumab, atezolizumab, durvalumab, avelumab, cemiplimab, toripalimab, dostarlimab, and (3) presence of at least one progress note or visit note in the clinical system. For the prospective silent pilot (phase 2), we implemented the best setting of our system in a silent, prospective, real-time clinical environment to assess performance under operational conditions and to assess generalizability to evolving clinical language and documentation practices at Dana-Farber/Brigham Cancer Center, from May-June 2025. The randomized user effect study (phase 3) took place via virtual participation from August-September 2025 at Dana-Farber/ Brigham Cancer Center. For Phase 1 and 2, we collected demographic information including age, sex, race, ethnicicty, and primary language. These were collected from the electronic health records. For Phase 3, we collected information on role, abstraction experience, and attitudes toward AI as these are most likely to be relevant for the endpoints of this study, via self-report.
Outcomes	The outcomes for retrospective and prospective system performance were precision, recall, and macro-averaged F1. We also collected agreement between prospective model predictions and human determinations using agreement percentages and override rates for each label and across all labels for Phase 2. The primary endpoint pf the randomized user effect study was annotation efficiency, measured as time in seconds from note opening to final submission, captured via screen recording. Secondary endpoints were: (1) accuracy, defined as exact match with the gold standard for all current irAE grades within a note—a stringent metric requiring perfect concordance across all six conditions; and (2) inter-annotator agreement, assessed using Krippendorff's alpha for ordinal grade labels. For the primary efficiency endpoint, we used generalized estimating equations (GEE) with an exchangeable correlation structure to account for within-participant clustering and estimated the reduction in annotation time attributable to AI assistance. Sensitivity analyses examined log-transformed times, adjusting for note set and sequence effects. For the accuracy endpoint, we used GEE with binomial outcomes to estimate the odds ratio of exact gold-standard match comparing AI-assisted to manual annotation, with robust sandwich variance estimators. Additional exploratory analyses relaxed the stringent complete-match criterion to examine accuracy for individual label types (grade, attribution, certainty) and by temporality. Further, inter-annotator reliability was assessed separately within each note set using Krippendorff's alpha for current grade labels. Bootstrap resampling (10,000 iterations, resampling participants with replacement) generated 95% confidence intervals for each method. Participants were clinical research coordinators and clinical research nurses recruited from a single academic department, and participation was voluntary; self-selection toward staff more interested in or comfortable with AI tools is therefore possible, which could bias the observed efficiency and usability gains upward. The crossover design mitigates this for the within-participant comparison, since each participant serves as their own control, but it does not remove the potential effect on generalizability to a broader curation workforce.

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A