Evaluating Medical LLMs by Levels of Autonomy: A Survey Moving from Benchmarks to Applications

Anonymous ACL submission

Abstract

Medical large language models achieve strong scores on standard benchmarks; however, transferring those results to safe and reliable performance in clinical workflows remains a challenge. This survey reframes evaluation through a levels-of-autonomy lens (L0–L3), spanning informational tools, information transformation and aggregation, decision support, and supervised agents. We align existing benchmarks and metrics with the actions permitted at each level and their associated risks, making the evaluation targets explicit. This motivates a level-conditioned blueprint for selecting metrics, assembling evidence, and reporting claims, alongside directions that link evaluation to oversight. By centering autonomy, the survey moves the field beyond score-based claims toward credible, risk-aware evidence for real clinical use.

1 Introduction

Large language models (LLMs) have advanced rapidly on medical benchmarks (Singhal et al., 2023; Hendrycks et al., 2020; Nazi and Peng, 2024). Both general-purpose (Qwen Team, 2025; DeepSeek-AI, 2025; OpenAI, 2025) and domain-specialized (Singhal et al., 2025; Bolton et al., 2024) LLMs now achieve high scores on licensing-style examinations and medical Q&A benchmarks (Oliveira et al., 2025; Tang et al., 2023). These headline results suggest that LLMs could provide meaningful assistance to clinicians and patients across a range of information-centric workflows.

However, benchmark correctness alone is not sufficient for clinical use (Hager et al., 2024; Ma et al., 2025). Clinical deployment requires consistency, fairness, auditability, calibrated uncertainty, and demonstrably safe clinical reasoning (Omiye et al., 2023; Fehr et al., 2024). Most benchmarks are Q&A-centric, which rarely probe these aspects, allowing unsafe reasoning and missing context to go undetected (Soroush et al., 2024; Asgari et al., 2025). A recent systematic review of 519 evaluations of healthcare applications of LLMs shows that 44.5% evaluated medical knowledge (like exam questions) and 19.5% focused on diagnosis, but only 5% used real patient data. (Bedi et al., 2025b). A rigorous evaluation that spans factual grounding, reasoning quality, uncertainty calibration, safety, and human preferences is required (Tam et al., 2024; Shool et al., 2025).

This survey reviews the state of LLM evaluation in the medical field and identifies its limitations. We first summarize how LLMs are being applied in the medical field and what current benchmarks measure. While these scores are informative, they only provide quick capability snapshots and overlook integration into real workflows, calibration, and traceable provenance. Therefore, we move from scores to applications, treating evaluation as a means of showing that a system is sufficient for a defined purpose and scope at a specific autonomy level. For each level, we define the scope, typical applications, evaluation focus, boundaries, and challenges: **L0** Inform (no personalized advice); **L1** Information Transformation & Aggregation (structure, summarize); **L2** Decision Support (recommendations and personalized advice); and **L3** Agents Under Human Supervision (plan + invoke tools/APIs to enqueue actions under explicit review). This organization makes evaluation targets explicit: factual grounding at L0/L1; calibrated reasoning and coverage at L2; tool-use safety and auditability at L3. Regarding the challenges in each level, they are cumulative, not isolated: each higher autonomy level inherits unresolved issues from lower levels. As autonomy and permitted actions expand, new risks emerge that are specific to that level's capabilities. Finally, we outline future work and recommendations for developing more robust evaluation frameworks to ensure that LLM-based systems can be trusted in clinical practice.

	L0-Inform	L1-Transformation & Aggregation	L2-Decision Support	L3-Agents under Supervision
Scope	Explain / Inform; NO advice	Structured extraction and aggregation	Recommendation; NO action	Plan, act under supervision
Application	HealthSearchQA, Patient Summary	EHR Transformation, Retrieve & Aggregate	Diagnostic Reasoning, Patient Education	EHR co-pilot, Clinical Simulation
Example	What is MRI?	From this discharge note, extract the medication list.	Given the summary, is MRI appropriate?	Assemble prior results, draft an order, queue for sign-off.
Challenge	Hallucination, Bias	Attribution, Privacy	Reasoning Consistency, Fairness	Tool-use, Human-AI Interaction

Table 1: Overview of autonomy levels (L0–L3) for medical LLMs, showing for each level the scope, typical applications, an example question/task, and key challenges. The rightward arrow indicates an increase in autonomy and risk. A detailed worked example/case study of an EHR copilot instantiated at different autonomy levels is provided in Appendix A.5. Example autonomy escalation criteria that specify when evidence at one level is sufficient to justify a move to the next are provided in Appendix A.3.

2 Related Work

Contemporary medical-LLM surveys mostly list datasets (e.g., USMLE/OKAP), task scenarios, and evaluation modes, but they rarely organize evaluation targets by autonomy level or permitted clinical actions. Representative examples include a review contrasting closed and open-ended tasks and discussing agentic settings (Chen et al., 2025c), and a systematic review highlighting reliance on general-purpose models and accuracy metrics with limited calibration and safety assessment (Shool et al., 2025). Parallel strands propose conversation-quality and safety evaluations (Abbasian et al., 2024), human-rater rubrics such as QUEST (Tam et al., 2024), multi-dimensional criteria in SCORE (Tan et al., 2024), and reporting guidance in TRIPOD-LLM (Gallifant et al., 2025). We address this gap by mapping evaluation objectives and metrics to autonomy levels (L0–L3). This mapping clarifies what evidence is sufficient for a system's intended role, which risks must be tested and where human oversight is required. A detailed comparison with existing survey papers is summarized in Tables 5 and 6 in the Appendix.

Autonomy scales outside medical-LLM evaluation exist, but they target system design or human factors, rather than evaluation blueprints. In clinical decision research, levels of autonomy delineate who acts and who bears responsibility (Festor et al., 2021). In the broader agent literature, five-level design frameworks define autonomy via user roles (operator → observer) (Feng et al., 2025a), and industry taxonomies describe L0–L5 agentic behavior (Kirkovska et al., 2025). None of these prescribe autonomy-conditioned metrics or align evaluation with healthcare oversight. Our survey fills that gap by integrating autonomy scales with concrete measurement choices for medical LLMs.

3 Evaluation Methodologies

We start by listing current benchmarks and metrics for medical LLMs. This sets clear measurement boundaries before we move to the applications. For LLMs in the medical field, benchmarks provide low-cost, repeatable evidence about specific capabilities within a bounded scope; they surface failure modes early, track progress over time, and help align claims with a system's intended role (L0–L3). First, we group benchmarks by task and summarize the usefulness of each, as well as the L0–L3 level(s) they inform (§3.1). We then summarize both automated and human metrics, along with their limitations (§3.2). Table 2 provides a compact illustration of our taxonomy with a few representative rows; the full expanded table appears in Appendix A.1.

3.1 Benchmarks

Exam Q&A. Answer multiple-choice and short-answer questions from professional medical examinations across specialties. Typical benchmarks include MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). This task provides a low-cost

Task	Usefulness	Level	Representative benchmarks	Example
Knowledge recall / exam Q&A	Low-cost unit tests of basic clinical knowledge.	L0	MedMCQA; MultiMedQA	Is it safe to take ibuprofen in early pregnancy?
Summarization / transformation	Checks whether models produce faithful summaries.	L1	MIMIC-IV-BHC; MTS-Dialog; ACI-Bench	Condense a multi-paragraph ICU discharge note into a 2–3 sentence brief hospital course
Decision support	Probes decision-making in simulated cases.	L2	HealthBench; MedHELM; DexBench	Given 7 days of CGM traces and meal logs, generate some advice.

Table 2: Condensed illustration of benchmark classes; the full expanded version appears in Appendix A.1

way to test factual breadth and specialty coverage. It is useful as a unit test for L0–L1 informational tools and for identifying coarse knowledge gaps.

Summarization. Generate faithful summaries, patient-friendly explanations, or structural rewrites of clinical text. Typical benchmarks include MS^2 (Wang et al., 2022) and PubMed long-document summarization (Cohan et al., 2018). This task evaluates whether outputs are faithful to the source (no important omissions) and structurally complete for explanation-oriented use cases at L0–L1.

Retrieval-augmented generation (RAG). Answer questions grounded in retrieved documents with explicit attribution to evidence passages. Typical benchmarks include BEIR (Thakur et al., 2021) and TREC-COVID (Roberts et al., 2021). This task demonstrates whether responses are supported by cited sources and avoid contradicting them, enabling checks on evolving topics at L0–L1.

Information extraction. Extract and normalize clinical entities, relations, and codes from notes and reports. Typical benchmarks include CBLUE (Zhang et al., 2021). This task establishes reliable structuring and normalization to support downstream aggregation in L1 systems.

Decision support. Make thresholded recommendations or triage decisions in vignettes or multi-turn clinical scenarios scored by clinician rubrics. Typical benchmarks include HealthBench (Arora et al., 2025) and MedHELM (Bedi et al., 2025a). This task targets selective reliability for L2 systems by quantifying calibrated behavior at thresholds and surfacing harm-proximal errors in simulation. Despite strong exam-style performance, overall success rates on HealthBench remain modest: GPT-3.5-Turbo satisfies only about 16% of physician-authored behavioral criteria, GPT-4o around 32%, and OpenAI's o3 roughly 60%, with more challenging "HealthBench Hard" subset capped at about 32% even for the best model.

Clinical dialogue. Conduct conversations with patients or clinicians to achieve goals while communicating safely and clearly. Typical benchmarks include MedDialog (Zeng et al., 2020) and MedDG (Liu et al., 2022). This task assesses communication effectiveness and safety across L0–L2.

3.2 Automated vs. Human Metrics

Automated Metrics. Automated metrics quantify answer correctness, calibration, faithfulness, and retrieval quality. For Q&A, Exact Match and F1 summarize accuracy, while ECE, Brier, and NLL capture probability calibration (Guo et al., 2017; Brier, 1950; Manning et al., 2008). In selective prediction, risk-coverage curves relate error to the answered fraction and justify abstention on uncertain cases (Geifman and El-Yaniv, 2017; Geifman and El-Yaniv, 2019; Traub et al., 2024). For summarization, lexical overlap (ROUGE/chrF) and embedding-based similarity (BERTScore/BLEURT) are commonly complemented with source-grounded checks for omissions and contradictions (Lin, 2004; Popović, 2015; Zhang et al., 2020; Sellam et al., 2020). In RAG settings, retrieval ranking is assessed with Recall@k, MRR, and nDCG (Manning et al., 2008).

Human Evaluation. Clinician raters typically apply behaviorally anchored rubrics (BARS-style) to score clinical correctness, coverage, contextualization, reasoning transparency, uncertainty handling, readability, actionability/safety, and empathy; inter-rater agreement is summarized with Cohen's κ or ICC (Holland et al., 2022; McHugh, 2012). For patient-facing text, readability is assessed using the Flesch-Kincaid or SMOG indices (Singh et al., 2024; Badarudeen and Sabharwal, 2010).

LLM-as-Judge (LAJ). To scale open-ended assessments, many studies adopt LLM-as-judge:

strong models render pairwise preferences or rubric-based scores, often with prompts that request quoted evidence. To control bias, studies randomize order or blind positions and, when feasible, ensemble multiple judges. (Zheng et al., 2023; Li et al., 2024c; Gu et al., 2024; Zhu et al., 2025; Tan et al., 2025). LAJ can track human preferences well in aggregate but is vulnerable to position and verbosity biases and self-preference effects (Chen et al., 2024; Ye et al., 2025a; Wataoka et al., 2024).

4 Applications and Autonomy Levels

From Scores to Applications. Static benchmarks provide an overview of the model capabilities, but they do not capture real workflow context, calibration, provenance, and audit trails, or role-specific needs. We therefore treat evaluation as showing that a system is **sufficient for a defined purpose and scope** at a given autonomy level. Concretely, each autonomy level is organized into five parts: (1) **Definition & scope**; (2) **Typical applications**; (3) **Evaluation focus**; (4) **Scope boundaries**; and (5) **Challenges**. Benchmarks remain useful as ingredients in this assessment, not the destination.

4.1 L0 - Inform

Definition and Scope. At autonomy level L0, the system functions purely as an informational tool: it explains medical concepts and provides a general background in plain language. It neither tailors advice to an individual patient nor initiates clinical decisions. Outputs are educational in nature and include an explicit non-advice disclaimer.

Typical Applications. Representative L0 tasks include answering common health questions (e.g., "What is MRI?"), producing lay summaries of technical passages, and simplifying lab reports. Public datasets used as proxies include consumer Q&A corpora and patient-facing summarization sets: HealthSearchQA within the MultiMedQA (Singhal et al., 2023) suite, the TREC 2017 LiveQA Medical task (Abacha et al., 2017), PubMedQA for abstract-level research comprehension, and consumer-question summarization datasets such as MeQSum and MEDIQA'21 (Ben Abacha and Demner-Fushman, 2019; Ben Abacha et al., 2021).

Evaluation Focus. L0 evaluation should primarily focus on accuracy, completeness, and readability (Srinivasan et al., 2025). Some benchmarks also include structured rubrics or provide reference contexts that enable additional checks: HealthBench (Arora et al., 2025) grades free-text answers with physician-written rubrics, MultiMedQA reports human ratings for long consumer answers, and PubMedQA uses the abstract as the reference context. We reference these only to motivate the axes here, not to require source citation as a core L0 practice.

Scope Boundaries. L0 outputs are intentionally non-patient-specific and involve minimal reasoning: they recall and lightly synthesize facts but do not conduct case workups, triage, or make recommendations (those belong to L2 Decision Support). As a result, L0 content may omit person-specific contraindications or time-sensitive context and should avoid language that could be interpreted as advice. Readability should target plain-language norms to reduce misunderstanding, and common safety disclaimers may be shown without tailoring. Some L0 benchmarks include rubrics or reference contexts for scoring, but we do not treat source citation as a core requirement for L0.

Challenges. Hallucination: Even at L0, models often produce fluent but unfaithful text: Broad surveys separate hallucinations into intrinsic and extrinsic, link them to training/decoding choices and weak grounding, and recommend source- and task-aware evaluations instead of generic overlap scores (Ji et al., 2023). In summarization specifically, human studies show that systems invent unsupported details and that standard n-gram metrics miss these errors, motivating faithfulness-oriented checks instead (Maynez et al., 2020). For patient-facing summaries, clinical audits advise using medical rubrics that check whether each claim is supported by the underlying notes and whether important facts are missing-rather than relying on readability alone. (Asgari et al., 2025). However, because these tools audit rather than eliminate hallucinations, the problem persists. Recent clinical summarization audits quantify how often hallucinations occur in practice. In psychiatric discharge summaries, human-written summaries were rated higher than GPT-4 summaries (3.78 vs 3.12 on a 5-point scale), and hallucinations appeared in 40% of AI summaries (Schmidgall et al., 2024). In an emergency-department study of 100 encounters, only 33% of GPT-4 summaries were error-free, while 42% contained hallucinations and 47% omitted clinically relevant information (Williams et al., 2025). **Bias:** We use bias to mean systematic tendencies that push outputs away from truth or intended scope (e.g., sensitivity to superficial cues), such that the same input intent

can yield different explanations. These tendencies plausibly arise from the pretraining objective and are further shaped by alignment procedures that optimize against human preferences (Ouyang et al., 2022; Sharma et al., 2023; Xu et al., 2025b; Shen et al., 2025). In L0 applications, such biases surface as assistants mirroring user beliefs rather than correcting them (Sharma et al., 2023) and shifting outputs under small, meaning-preserving prompt changes such as formatting tweaks, option order in multiple-choice settings, or early anchoring hints (Sclar et al., 2023; Pezeshkpour and Hruschka, 2023; Lou et al., 2024; Zhou et al., 2024; Li et al., 2024b; Ye et al., 2025b; Li et al., 2024a; RRV et al., 2025). They persist in practice because preference-optimized objectives and prompt conventions are integral to how LLMs are used (Kadavath et al., 2022; Steyvers et al., 2025). Consequently, mitigations such as reporting performance ranges across prompt formats, anchor-aware templates, and a conservative tone can reduce bias at L0, but they do not reliably remove it. (Sclar et al., 2023; Pezeshkpour and Hruschka, 2023; Lou et al., 2024; Steyvers et al., 2025).

4.2 L1 - Information Transformation & Aggregation

Definition and Scope. This stage turns raw, heterogeneous clinical data into standardized, computable representations and then combines them with external evidence to produce grounded outputs. In practice, health systems map local EHR fields to an interoperability standard or a research schema, attach machine-readable provenance for auditability, and operate over de-identified corpora such as MIMIC-IV that illustrate the target tables (encounters, labs, medications) and privacy constraints (Abacha et al., 2021; Alsentzer et al., 2023; Zhang et al., 2024a; Liu et al., 2021).

Typical Applications. EHR data transformation: Typical pipelines mix schema harmonization (FHIR/OMOP ETL), clinical NLP to extract entities/attributes from notes, concept normalization to standard vocabularies, and de-identification. i2b2/n2c2 shared tasks supply widely used de-identified note sets for de-identification, concept extraction, relation labeling, and medication-change context, enabling objective measurement of span-level and mapping accuracy (Li et al., 2016; Abacha et al., 2021; Nowak et al., 2023; Mahajan et al., 2023; Henry et al., 2020). **Retrieve & Aggregate:** On top of the transformed corpus, systems index structured facts (problem lists, meds, labs) and unstructured notes, then pair them with external sources (guidelines, reviews) via retrieval-augmented generation (RAG). Retrieval metrics (Recall@k, nDCG, MRR) assess whether the right evidence is fetched; generation metrics (faithfulness/attribution, grounded-answer rate) check that outputs rely on retrieved passages rather than model priors (Cohan et al., 2020; Zhang et al., 2024a; Tang et al., 2023). Representative resources include BEIR for generalizable retrieval evaluation, TREC-COVID for high-stakes, rapidly evolving topics, and domain-specific retrievers such as MedCPT for biomedical search (Thakur et al., 2021; Roberts et al., 2021; Jin et al., 2023).

Evaluation Focus. Transformation: report extraction/normalization scores (e.g., span-level F1, concept mapping accuracy), coverage/completeness of key fields, and lineage completeness. **Retrieval & Aggregation**: report Recall@k/nDCG/MRR; grounded-answer and attribution rates to retrieved passages; contradiction-to-source; and selective prediction/abstention rates under uncertainty (Alsentzer et al., 2023; Abacha et al., 2021; Zhang et al., 2024a).

Scope Boundaries. L1 improves structure, traceability, and access to evidence, but it does not perform the patient-specific reasoning required for diagnosis, test selection, or treatment trade-offs. Real clinical workups must integrate temporality, comorbidities, contraindications, and uncertainty-capabilities not captured by schema or Recall@k alone. Studies also show that (i) retrieval can surface conflicting or outdated sources and (ii) LLMs may misattribute or over-trust citations; recent medical RAG evaluations highlight these gaps even when retrieval quality is strong (Roberts et al., 2021; Peng et al., 2023; Hueber and Kleyer, 2023; Golan et al., 2023; Dhanvijay et al., 2023; Sezgin et al., 2023; Team et al., 2024). Thus, L0+L1 should be paired with higher-level (L2+) decision-focused assessments before deployment.

Challenges. Attribution: Attribution concerns whether aggregated outputs actually rely on and are traceable to the retrieved sources. Automated audits in the medical domain report that models often cite papers that are only loosely relevant or that do not fully support the generated claims (Wu et al., 2025). Even with explicit attribution metrics and fine-grained factuality scoring, models can pass retrieval tests while still incorporating priors or blending multiple sources in ways that obscure

provenance (Yang et al., 2025). Techniques like Self-RAG improve on-demand retrieval and self-critique, yet do not eliminate misattribution when sources disagree, are low quality, or when prompts nudge the model toward fluent synthesis over faithful quotation (Asai et al., 2023; Jung et al., 2024). Thus, attribution still remains a challenge. **Completeness:** Completeness addresses whether transformed corpora and their aggregations cover all clinical facts (problems, meds, labs, temporality, context) without omissions. Comparative studies find that concept-recognition tools are inconsistent and often miss negations, misread abbreviations, and struggle with ambiguity or misspellings, which leads to missing or distorted facts downstream (Lossio-Ventura et al., 2023). Beyond extraction, mapping text spans to standard concepts is fragile because benchmark datasets contain many ambiguous terms and do not align well with UMLS coverage, so the 'correct' code is often unclear from the beginning (Newman-Griffis et al., 2021). In practice, long EHR narratives, events spread across notes, and uneven local coding leave gaps that retrieval can't fill when the structured substrate is incomplete. These issues persist because gold standards underrepresent edge cases, annotation guidelines vary, and many L1 evaluations emphasize span-level F1 or Recall@k over end-to-end coverage of clinically critical fields (Lossio-Ventura et al., 2023). **Privacy & lineage:** It evaluates whether L1 transformations are governable and safe to share. De-identification of clinical notes reduces direct identifiers but does not preclude membership inference against downstream models (Sarkar et al., 2024). At the model layer, training data extraction and related attacks demonstrate that LLMs can memorize and regurgitate snippets of their training corpora (Carlini et al., 2021). These realities make rigorous lineage essential: machine-readable provenance (i.e., what data, which ETL/normalizers) should be recorded using established schemas, allowing organizations to audit and reproduce outputs (Lebo et al., 2013; Mitchell et al., 2019; Gebru et al., 2021). Yet provenance remains challenging in practice because pipelines are multi-hop and frequently updated; components are swapped or fine-tuned; and evidence bases evolve.

4.3 L2 - Decision Support

Definition and Scope. At autonomy level L2, the system provides patient-specific recommendations that can assist clinical decision-making. By design, L2 depends on upstream EHR transformation and retrieval/aggregation (L1) but adds reasoning over the individual's data and clinical context.

Typical Applications. Diagnostic reasoning: it reads a patient's history, exam, labs, and imaging to propose differentials with brief rationales and possible next steps; evaluations commonly use vignette-based case sets and prompting schemes that elicit stepwise clinical reasoning (Goh and colleagues, 2024; Savage et al., 2024a). **Medication decision support:** it integrates a patient's active medications, allergies, problems, and recent labs to surface potential adverse drug events, drug–drug interactions, and dosing/contraindication checks; widely used resources include the 2018 n2c2 ADE shared task and the SemEval-2013 DDIExtraction task for literature-based DDI detection (Henry et al., 2020; Segura-Bedmar et al., 2013). **Patient education:** Patient education tools support patient–clinician communication by turning medical information into clear, usable messages and scaffolding two-way conversations. Typical functions include generating plain-language explanations of conditions, tests, and procedures; prompting patients to ask key questions; and using teach-back so patients restate key points to confirm understanding (Shoemaker et al., 2013; Centers for Disease Control and Prevention, 2020; Agency for Healthcare Research and Quality, 2023; Institute for Healthcare Improvement, n.d.; IPDAS Collaboration, 2024; Stacey et al., 2021). Taken together, these tasks are L2 because they require reasoning over an individual patient's context to generate recommendations or tailored explanations.

Evaluation Focus. For L2 decision support, evaluation should foreground three axes beyond L0's accuracy/completeness/readability: reasoning consistency, calibration and abstention, and safety. For reasoning, score the process, not just final answers: stepwise diagnostic logic and evidence use should be judged against clinician-authored rubrics or case rationales. For calibration, report reliability at the case level, plus selective prediction curves (risk–coverage) with a tunable "don't know / escalate" option. For safety, it is essential to track rates of contraindications and guideline violations.

Scope Boundaries. At L2, systems provide decision support by combining a patient's EHR context with external evidence to offer guidance. However, they do not plan tasks, call tools or APIs, or change the record, so a clinician must review and decide.

This preserves clinical authority but requires clinicians to convert text into orders and messages, which adds to their workload and increases the risk of transcription errors. These limits motivate a shift to agentic L3 configurations that keep a human in the loop while reducing cognitive burden.

Challenges. Reasoning consistency & faithfulness: It concerns whether patient-specific answers are stable across prompt phrasings and whether the rationales actually support the recommendation. Comparative guideline evaluations demonstrate that modifying formats or instructions can significantly impact model outputs and agreement, underscoring the sensitivity of prompts in medical settings (Wang et al., 2024). Beyond prompt sensitivity, in a study of seven LLMs on 2,400 abdominal-pain cases, clinicians achieved around 90% diagnostic accuracy, whereas model performance was on average 16–25 percentage points lower and more frequently violated guideline-concordant workup recommendations (Hager et al., 2024). Even when stepwise prompts are used to elicit reasoning, studies find that the generated "explanations" can be unfaithful (Turpin et al., 2023; Madsen et al., 2024; Kuang et al., 2025; RRV et al., 2025). In diagnostic contexts, structured prompting can improve transparency but does not guarantee faithful causal grounding of the final answer (Savage et al., 2024b; Dineen et al., 2025). The result in practice is that two seemingly careful L2 prompts may yield different plans with rationales that read well but do not reliably reflect the model's decision process, which keeps consistency and faithfulness an open problem for deployment (Turpin et al., 2023). **Confidence calibration:** Here the question is whether stated or implicit confidence tracks correctness so that uncertain cases can be flagged or deferred. Recent medical evaluations show that simple proxies correlate only weakly with error (Bentegeac et al., 2025). Semantic-entropy approaches detect confabulations and better prioritize abstention, but their reliability varies by task and domain (Farquhar et al., 2024; Kossen et al., 2024; Penny-Dimri et al., 2025). In real L2 use, these trade-offs mean systems may sound certain on incorrect recommendations or abstain too rarely on edge cases (Bentegeac et al., 2025; Farquhar et al., 2024; Feng et al., 2025b). **Fairness:** The core issue is whether recommendations generalize equitably across patient subgroups and clinical contexts. Specialized benchmarks and audit tools reveal that biases are present in long-form answers and clinical recommendations, which plain accuracy scores fail to capture (Prakash et al., 2024). Purpose-built bias benchmarks for clinical LLMs report subgroup-linked shifts in outputs, while triage studies using counterfactual tests reveal intersectional differences across sex and race (Zhang et al., 2024b; Lee et al., 2025). A recent evaluation in healthcare showed that changing a patient's race or gender caused GPT-4 to miss the top correct diagnosis in 37% of cases. (Zack et al., 2024) A recent scoping review highlights uneven coverage across medical fields and limited clinician-in-the-loop evaluation-leaving blind spots in where and how biases manifest (Cheng et al., 2025). Consequently, subgroup reliability remains a persistent challenge (Prakash et al., 2024; Cheng et al., 2025).

4.4 L3 - Agents under human supervision

Definition and Scope. We define L3 agents as systems that plan and invoke tools to initiate actions in clinical workflows while keeping a clinician explicitly "in the loop" for review, modification, and sign-off. Core capabilities are task planning, retrieval, and safe tool use, with human oversight enforced at key checkpoints (e.g., order "draft" states).

Typical Applications. Clinical Copilot: Clinical copilots plan tasks, fetch chart context, and then draft orders, messages, referrals, or care plans via tool calls, pausing for human sign-off (e.g., Polaris, Rx Strategist, Almanac Copilot) (Mukherjee et al., 2024; Van et al., 2024; Zakka et al., 2024). **Sandboxed Simulation:** It embeds agents in controlled clinics to probe plan–act–check loops with audit trails-benchmarks like AgentClinic, AI Hospital, ClinicalLab, and 3MDBench (Schmidgall et al., 2024; Fan et al., 2024; Yan et al., 2024; Sviridov et al., 2025; Yue et al., 2025; Chen et al., 2025a; Xu et al., 2025a; Jiang et al., 2025). **Operation & EHR Automation:** It coordinates non-diagnostic workflows (e.g., prior authorization) and return proposed actions for approval (RxLens; multi-agent prior-auth pipelines) (Jagatap et al., 2025). In contrast to L2 (textual recommendations), these L3 systems invoke tools to initiate actions but keep a clinician in the loop for review and sign-off.

Evaluation Focus. At L3, evaluation shifts from answer quality to supervised action quality: studies should first demonstrate end-to-end task success under human oversight. Second, they must verify tool-use correctness: every API call, order, and parameter matches clinical intent. Finally, they should require auditability via machine-readable

provenance of data, prompts, models, tools, parameters, and approvals, leveraging standards such as FHIR `Provenance` and `AuditEvent` so that actions can be traced for post-hoc review and governance.

Scope Boundaries. L3 agents are limited to draft-and-queue actions under explicit human oversight: they may plan tasks and call approved tools/APIs, but execution requires a clinician's independent review. Looking ahead, closed-loop system deployments can enable agents to execute pre-approved, low-risk steps when explicit policies and fine-grained access controls are met, rather than requiring individual approval for each action.

Challenges. Tool-use failures: These arise when an agent plans correctly but issues the wrong API call or constructs malformed parameters, so the action fails even if the reasoning was sound. In one benchmark of 34 complex tasks, only 50% of tasks were successfully completed, with failures attributed to improper planning, non-functional code generation, or inadequate self-correction loops (Lu et al., 2025). **Human–AI interaction dynamics:** These refer to how clinicians review and sign off on queued actions. Studies of decision support reveal an overreliance on automated suggestions and alert fatigue, where users either ignore or overaccept system outputs, resulting in oversight gaps even when accuracy is reasonable (Khera et al., 2023; Abdelwanis et al., 2024). The problem persists because busy workflows, lengthy sessions, and inconsistent interface cues make it challenging to calibrate attention and sustain critical review at every step. **Auditability and accountability:** These require a traceable record of what data were accessed, which tools were called with what parameters, and who approved the final action. Standards such as FHIR `AuditEvent` and `Provenance` define the necessary primitives, and governance frameworks emphasize logging and traceability for post-hoc review (HL7 International, 2025a,b; National Institute of Standards and Technology, 2023; World Health Organization, 2025). Yet multi-tool, multi-service agent stacks often lack end-to-end lineage across steps, so reconstructing a failure or near-miss remains difficult in practice. **Scalability of training:** Medical agents cannot typically be trained directly in live care settings because unconstrained exploration is ethically and logistically infeasible, pushing work toward simulated environments (Schmidgall et al., 2024; Fan et al., 2024). However, strong performance in such synthetic setups does not guarantee reliable transfer to real-world workflows: benchmarks often simplify states and constraints, leaving significant gaps between simulation and reality (Jayaraman et al., 2024; Hargrave et al., 2024).

5 Future Work

Closed-looped System We expect clinical deployments to shift from single-model helpers to closed-loop, hospital-scale systems composed of cooperating, role-specialized agents that escalate to clinicians at predefined gates (Schmidgall et al., 2024; Borkowski and Ben-Ari, 2025). We call for simple, auditable protocols for handoff, disagreement resolution, and safety gating of tool use, preceded by simulation and shadow deployments. We recommend reporting operational outcomes (e.g., deferral rates) so the community can compare architectures and converge on safe patterns (Liu et al., 2020).

Guarantees. We call for reframing "good" performance around risk-controlled selectivity: in clinical settings, systems should act only when a target risk can be met and otherwise defer, evaluated by risk–coverage rather than accuracy. In this agenda, reliable high accuracy on a subset with explicit deferral (e.g., 99% accuracy on 20% task) is preferable to broad moderate accuracy (e.g., 80% on 80%), because the former enables out-of-scope detection and smooth human routing while the latter obscures which cases are incorrect. We recommend that future evaluations report calibrated confidence on the acted-on subset and coverage at target risk, along with slice breakdowns (Guo et al., 2017).

6 Conclusion

This survey reframes the evaluation of medical LLMs around levels of autonomy (L0–L3). We emphasize risk coverage rather than average accuracy: safe systems act only when they can meet a target risk and defer otherwise. At L0–L1, the focus is factual fidelity, bias, and structural correctness with clear grounding to sources. At L2–L3, the bar rises to uncertainty, subgroup robustness, tool-use correctness, and verified human oversight.

Taken together, this level-conditioned lens turns benchmark scores into clinically relevant claims. Evaluations are most persuasive when they make the level explicit, pair target risk with achieved coverage, report performance across levels, and verify both confidence calibration and oversight checkpoints. We hope this provides a clear language for building medical LLMs that are not only capable but also reliably useful and safe in practice.

Limitations

This survey is necessarily selective and time-bounded; model releases, datasets, and guidance evolve rapidly, so some details may become outdated. Evidence in the literature remains uneven: many studies emphasize lab benchmarks over prospective or randomized evaluations, and reporting quality is inconsistent.

References

Asad Aali, Dave Van Veen, Yamin Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash Tehrani, Jangwon Kim, and Akshay Chaudhari. 2025. Mimic-iv-ext-bhc: Labeled clinical notes dataset for hospital course summarization. PhysioNet Repository.

A. B. Abacha, E. Agichtein, Y. Pinter, and D. Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. *TREC*, pages 1–12.

A. B. Abacha, Y. M'rabet, Y. Zhang, C. Shivade, C. Langlotz, and D. Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.

Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li-Jia Li, Ramesh Jain, and Amir M. Rahmani. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *npj Digital Medicine*, 7(1):82.

Moustafa Abdelwanis, Hamdan Khalaf Alarafati, Maram M. S. Tammam, and Mecit Can Emre Simsekler. 2024. Exploring the risks of automation bias in healthcare artificial intelligence applications: A bowtie analysis. *Journal of Safety Science and Resilience*, 5:460–469.

Agency for Healthcare Research and Quality. 2023. Tool: Teach-back.

Ahmed M. Alkalbani, Ahmed S. Alrawahi, Ahmed Salah, Vafa Haghighi, Yong Zhang, Saif Alkindi, and Quan Z. Sheng. 2025. A systematic review of large language models in medical specialties: Applications, challenges and future directions. *Information*, 16(6):489.

E. Alsentzer, M. J. Rasmussen, R. Fontoura, A. L. Cull, B. Beaulieu-Jones, K. J. Gray, D. W. Bates, and V. P. Kovacheva. 2023. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *NPJ Digital Medicine*, 6(1).

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Sameer Badarudeen and Sanjeev Sabharwal. 2010. Assessing readability of patient education materials: Current role in orthopaedics. *Clinical Orthopaedics and Related Research*, 468(10):2572–2580.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, and 1 others. 2025a. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, and 1 others. 2025b. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 333(4):319–328.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of ACL*, pages 2228–2234.

Asma Ben Abacha, Dina Demner-Fushman, Shweta Yadav, and Deepak Gupta. 2021. MEDIQA 2021: Consumer health question summarization, multi-answer summarization, and radiology report summarization. `https://sites.google.com/view/mediqa2021`. NAACL-BioNLP shared task site.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor–patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Raphaël Bentegeac, Bastien Le Guellec, Grégory Kuchcinski, Philippe Amouyel, and Aghiles Hamroun. 2025. Token probabilities to mitigate large language models overconfidence in answering medical questions: Quantitative study. *Journal of Medical Internet Research*, 27:e64348.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text. *Preprint*, arXiv:2403.18421.

Andrew A. Borkowski and Alon Ben-Ari. 2025. Multiagent ai systems in health care: Envisioning next-generation intelligence. *Federal Practitioner*, 42(5):188–193.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Nathan Brown, Marco Fiscato, Marwin Segler, and Alain C. Vaucher. 2019. Guacamol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108.

Maria Ana Cardei, Josephine Lamp, Mark Derdzinski, and Karan Bhatia. 2025. Dexbench: Benchmarking llms for personalized decision making in diabetes management. *Preprint*, arXiv:2510.00038.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, pages 2633–2650.

Centers for Disease Control and Prevention. 2020. Clear communication index user guide.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327. Association for Computational Linguistics.

Shan Chen, Pedro Moreira, Yuxin Xiao, Sam Schmidgall, Jeremy Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle S. Bitterman. 2025a. Medbrowsecomp: Benchmarking medical deep research and computer use. *Preprint*, arXiv:2505.14963.

Tong Chen, Zimu Wang, Yiyi Miao, Haoran Luo, Yuanfei Sun, Wei Wang, Zhengyong Jiang, Procheta Sen, and Jionglong Su. 2025b. Medfact: A large-scale chinese dataset for evidence-based medical fact-checking of llm responses. *Preprint*, arXiv:2509.17436.

Xinyu Chen, Jiayi Xiang, Shudong Lu, Yifan Liu, Meng He, and Dan Shi. 2025c. Evaluating large language models and agents in healthcare: Key challenges in clinical applications. *Intelligent Medicine*, 5(2):151–163.

Lionel Tim-Ee Cheng, Jasmine Chiat Ling Ong, Zhen Ling Teo, Ting Fang Tan, Narrendar RaviChandran, Fei Wang, Leo Anthony Celi, Marcus Eng Hock Ong, and Nan Liu. 2025. A scoping review and evidence gap analysis of clinical AI fairness. *npj Digital Medicine*. PMCID: PMC12167363.

Jan Clusmann, Franziska R. Kolbinger, Hannah S. Muti, Zoe I. Carrero, Jens-Niklas Eckardt, Nima Ghaffari Laleh, Carsten M. L. Löffler, Stefan-Christian Schwarzkopf, Matthias Unger, Gabriëlle P. Veldhuizen, Sabine J. Wagner, and Jakob N. Kather. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3:141.

A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *Preprint*, arXiv:2104.06486.

A. K. D. Dhanvijay, M. J. Pinjar, N. Dhokane, S. R. Sorte, A. Kumari, H. Mondal, and A. K. Dhanvijay. 2023. Performance of large language models (chatgpt, bing search, and google bard) in solving case vignettes in physiology. *Cureus*, 15(8).

Jacob Dineen, Aswin RRV, Qin Liu, Zhikun Xu, Xiao Ye, Ming Shen, Zhaonan Li, Shijie Lu, Chitta Baral, Muhao Chen, and Ben Zhou. 2025. Qa-lign: Aligning llms through constitutionally decomposed qa. *Preprint*, arXiv:2506.08123.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Preprint*, arXiv:2402.09742.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. A dataset of simulated patient–physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9:313.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Jana Fehr, Brian Citro, Rohit Malpani, Christoph Lippert, and Vince I. Madai. 2024. A trustworthy ai reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, 6:1267290.

Kevin J. Feng, David W. McDonald, and Amy X. Zhang. 2025a. Levels of autonomy for ai agents. arXiv:2506.12469.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. 2025b. Bird: A trustworthy bayesian inference framework for large language models. *Preprint*, arXiv:2404.12494.

Paul Festor, Ibrahim Habli, Yan Jia, Anthony Gordon, A. Aldo Faisal, and Matthieu Komorowski. 2021. Levels of autonomy and safety assurance for ai-based clinical decision systems. In *SAFECOMP 2021 Workshops*, pages 291–296. Springer.

Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Sherri Chen, Giovanni E. Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Christopher Fernandes, and 1 others. 2025. The tripod-llm reporting guideline for studies using large language models. *Nature Medicine*, 31(1):60–69.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR.

Elizabeth Goh and colleagues. 2024. Large language model influence on diagnostic reasoning. *JAMA Network Open*.

R. Golan, S. J. Ripps, R. Reddy, J. Loloi, A. P. Bernstein, Z. M. Connelly, N. S. Golan, R. Ramasamy, S. Ripps, and R. V. Reddy. 2023. Chatgpt's ability to assess quality and readability of online medical information: evidence from a cross-sectional study. *Cureus*, 15(7).

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622.

Mason Hargrave and 1 others. 2024. Epicare: A reinforcement learning benchmark for dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, volume 37, pages 130536–130568.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Özlem Üzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

HL7 International. 2025a. Hl7 fhir auditevent resource. `https://build.fhir.org/auditevent.html`. Accessed Oct 5, 2025.

HL7 International. 2025b. Hl7 fhir provenance resource. `https://build.fhir.org/provenance.html`. Accessed Oct 5, 2025.

Jaycelyn R. Holland, Donald H. Arnold, Holly R. Hanson, Barbara J. Solomon, Nicholas E. Jones, Tucker W. Anderson, Wu Gong, Christopher J. Lindsell, Travis W. Crook, and Daisy A. Ciener. 2022. Reliability of the behaviorally anchored rating scale (bars) for assessing non-technical skills of medical students in simulated scenarios. *Medical Education Online*, 27(1).

Yuchen Hua, Hyeok Na, Zeng Li, Fang Liu, Xia Fang, David Clifton, and John Torous. 2025. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8:1611.

A. J. Hueber and A. Kleyer. 2023. Quality of citation data using the natural language processing tool chatgpt in rheumatology: creation of false references. *RMD open*, 9(2).

Institute for Healthcare Improvement. n.d. Ask me 3: Good questions for your good health. Accessed 2025-10-05.

IPDAS Collaboration. 2024. International patient decision aid standards (ipdas) collaboration. Accessed 2025-10-05.

Akshay Jagatap, Srujana Merugu, and Prakash Mandayam Comar. 2025. Rxlens: Multi-agent llm-powered scan and order for pharmacy. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 822–832, Albuquerque, New Mexico. Association for Computational Linguistics.

Prithvi Jayaraman, Jack Desman, Majid Sabounchi, Girish N. Nadkarni, and Ankit Sakhuja. 2024. A primer on reinforcement learning in medicine for clinicians. *npj Digital Medicine*, 7(1):337.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*.

D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11).

A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1).

Dongwon Jung, Qin Liu, Tenghao Huang, Ben Zhou, and Muhao Chen. 2024. Familiarity-aware evidence compression for retrieval-augmented generation. *Preprint*, arXiv:2409.12468.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S. Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W. Safranek, Abid A. Anwar, Andrew Zhang, Aidan Gilson, Maxwell B. Singer, Amisha Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, and Zhiyong Lu. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Preprint*, arXiv:2406.12036.

Rohan Khera, Melissa A. Simon, and Joseph S. Ross. 2023. Automation bias and assistive ai: Risk of harm from ai-driven clinical decision support. *JAMA*, 330(23):2255–2257.

Anita Kirkovska, Nico Finelli, and David Vargas. 2025. The six levels of agentic behavior. Vellum AI Blog.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Jiayi Kuang, Haojing Huang, Yinghui Li, Xinnian Liang, Zhikun Xu, Yangning Li, Xiaoyu Tan, Chao Qu, Meishan Zhang, Ying Shen, and Philip S. Yu. 2025. Atomic thinking of llms: Decoupling and exploring mathematical reasoning abilities. *Preprint*, arXiv:2509.25725.

Timothy Lebo, Satya S. Sahoo, Deborah L. McGuinness, and 1 others. 2013. PROV-O: The PROV ontology. W3C Recommendation.

Joseph Lee, Tianqi Shang, Jae Young Baik, Duy Duong-Tran, Shu Yang, Lingyao Li, and Li Shen. 2025. From promising capability to pervasive bias: Assessing large language models for emergency department triage. *arXiv preprint arXiv:2504.16273*.

Bangzheng Li, Ben Zhou, Xingyu Fu, Fei Wang, Dan Roth, and Muhao Chen. 2024a. Famicom: Further demystifying prompts for language models with task-agnostic performance estimation. *Preprint*, arXiv:2406.11243.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2024b. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? *Preprint*, arXiv:2311.09702.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024c. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical–disease relation extraction. *Database*, 2016:baw068.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

G. Liu, Y. Liao, F. Wang, B. Zhang, L. Zhang, X. Liang, X. Wan, S. Li, Z. Li, and S. Zhang. 2021. Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):3786–3797.

W. Liu, J. Tang, Y. Cheng, W. Li, Y. Zheng, and X. Liang. 2022. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459.

Xiaoxuan Liu, S. C. Rivera, D. Moher, and et al. 2020. Consort-ai extension: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence. *Nature Medicine*, 26:1364–1374.

Juan Antonio Lossio-Ventura, Ran Sun, Sebastien Boussard, and Tina Hernandez-Boussard. 2023. Clinical concept recognition: Evaluation of existing systems on ehrs. *Frontiers in Artificial Intelligence*, 5:1051724.

Zhong Lou, Chen Zhao, Wenhao Li, Chenyu Zhang, Qi Zhu, Xuehai Pan, Mohit Bansal, and Kai-Wei Chang. 2024. Anchoring effects in large language models. *arXiv preprint arXiv:2412.06593*.

Ruofan Lu, Yichen Li, and Yintong Huo. 2025. Exploring autonomous agents: A closer look at why they fail when completing tasks. *arXiv preprint arXiv:2508.13143*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022. Biored: A rich biomedical relation extraction dataset. *Briefings in Bioinformatics*.

Zizhan Ma, Wenxuan Wang, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Wenting Chen, and Linlin Shen. 2025. Beyond the leaderboard: Rethinking medical benchmarks for large language models. *Preprint*, arXiv:2508.04325.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? *arXiv preprint arXiv:2401.07927*.

Diwakar Mahajan, Jennifer J. Liang, Ching-Huei Tsou, and Özlem Uzuner. 2023. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of Biomedical Informatics*, 144:104432.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F. Molina, and Ahmed Alaa. 2025. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *Preprint*, arXiv:2505.22919.

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, Jun Gao, Ya-Geng-Shun Wang, Jun Ji, Zhen Qiu, Min Li, Changhong Qian, Tengda Guo, Shu Ma, Zhi Wang, and Yue-Dong Tang. 2024. The application of large language models in medicine: A scoping review. *iScience*, 27(5):109713.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229.

Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, Cezanne Bianco, Swapnil Sharma, Rae Lasko, Michelle Voisard, Sanchay Harneja, Darya Filippova, Gerry Meixiong, Kevin Cha, Amir Youssefi, and 7 others. 2024. Polaris: A safety-focused llm constellation architecture for healthcare. *Preprint*, arXiv:2403.13313.

National Institute of Standards and Technology. 2023. Artificial intelligence risk management framework (ai rmf 1.0). `https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf`. Accessed Oct 5, 2025.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. *Informatics*, 11(3):57.

Denis Newman-Griffis, Guy Divita, Bart Desmet, Ayah Zirikly, Carolyn P Rosé, and Eric Fosler-Lussier. 2021. Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. *Journal of the American Medical Informatics Association*, 28(3):516–532.

S. Nowak, D. Biesner, Y. Layer, M. Theis, H. Schneider, W. Block, B. Wulff, U. Attenberger, R. Sifa, and A. Sprinkart. 2023. Transformer-based structuring of free-text radiology report databases. *European Radiology*, 33(6):4228–4236.

Juliana Damasio Oliveira, Henrique D. P. Santos, Ana Helena D. P. S. Ulbrich, Julia Colleoni Couto, Marcelo Arocha, Joaquim Santos, Manuela Martins Costa, Daniela Faccio, Fabio O. Tabalipa, and Rodrigo F. Nogueira. 2025. Development and evaluation of a clinical note summarization system using large language models. *Communications Medicine*, 5(1):376.

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *npj Digital Medicine*, 6:195.

OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-10-02.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, and A. Ray. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35.

A. Pal, L. K. Umapathi, and M. Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. *Conference on health, inference, and learning*, pages 248–260.

C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, and T. Magoc. 2023. A study of generative large language model for medical research and healthcare. *NPJ Digital Medicine*, 6(1).

Jahan C. Penny-Dimri, Magdalena Bachmann, William R. Cooke, Sam Mathewlynn, Samuel Dockree, John Tolladay, Jannik Kossen, Lin Li, Yarin Gal, and Gabriel Davis Jones. 2025. Reducing large language model safety risks in women's health using semantic entropy. *Preprint*, arXiv:2503.00269.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. 2018. Molecular sets (moses): A benchmarking platform for molecular generation models. *Preprint*, arXiv:1811.12823.

Maja Popović. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Sushant Prakash, Katherine Heller, Alan Karthikesalingam, Christopher Semturs, Joelle Barral, Greg Corrado, Yossi Matias, Jamila Smith-Loud, Ivor Horn, and Karan Singhal. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30:3590–3600.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R. Hersh. 2021. Searching for scientific evidence in a pandemic: An overview of trec-covid. *Journal of Biomedical Informatics*, 121:103865.

Aswin RRV, Jacob Dineen, Divij Handa, Md Nayem Uddin, Mihir Parmar, Chitta Baral, and Ben Zhou. 2025. Thinktuning: Instilling cognitive reflections without distillation. *Preprint*, arXiv:2508.07616.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, et al., and Xiaoqian Jiang. 2024. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific Reports*, 14(29669).

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024a. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine*, 7(20).

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024b. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine*, 7(1):20.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*. ICLR 2024 camera-ready version.

Isabel Segura-Bedmar, Paloma Martínez, and David Herrero-Zazo. 2013. Semeval-2013 task 9: Extraction of drug–drug interactions from biomedical texts (ddiextraction 2013). In *Proceedings of SemEval 2013*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

E. Sezgin, F. Chekeni, J. Lee, and S. Keim. 2023. Clinical accuracy of large language models and google search responses to postpartum depression questions: cross-sectional study. *Journal of Medical Internet Research*, 25.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell,

Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Ming Shen, Zhikun Xu, Jacob Dineen, Xiao Ye, and Ben Zhou. 2025. Bow: Reinforcement learning for bottlenecked next word prediction. *Preprint*, arXiv:2506.13502.

Sarah J. Shoemaker, Michael S. Wolf, and Cindy Brach. 2013. *The Patient Education Materials Assessment Tool (PEMAT) and User's Guide (Version 1.0)*. Agency for Healthcare Research and Quality, Rockville, MD. AHRQ Publication No. 14-0002-EF. Updated August 2014.

Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(117).

Som Singh, Aleena Jamal, and Fawad Qureshi. 2024. Readability metrics in patient education: Where do we innovate? *Clinics and Practice*, 14(6):2341–2349.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, and S. Pfohl. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Ali Soroush, Benjamin S. Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W. Charney, Girish N. Nadkarni, and Eyal Klang. 2024. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI*, 1(5).

Adarsh Srinivasan, Jacob Dineen, Muhammad Umar Afzal, Muhammad Uzair Sarfraz, Irbaz B. Riaz, and Ben Zhou. 2025. Recap: Transparent inference-time emotion alignment for medical dialogue systems. *Preprint*, arXiv:2509.10746.

Dawn Stacey, Robert J. Volk, and IPDAS Evidence Update Leads. 2021. The international patient decision aid standards (ipdas) collaboration: Evidence update 2.0. *Medical Decision Making*, 41(7):729–733.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7:221–231.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 2025. 3mdbench: Medical multimodal multi-agent dialogue benchmark. *arXiv preprint arXiv:2504.13861*.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V. Stolyar, Katelyn Polanska, Karleigh R. McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1):258.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR 2025 (published version); arXiv:2410.12784.

Ting Fang Tan, Kabilan Elangovan, Jasmine Ong, Nigam Shah, Joseph Sung, Tien Yin Wong, Lan Xue, Nan Liu, Haibo Wang, Chang Fu Kuo, Simon Chesterman, Zee Kin Yeong, and Daniel S. W. Ting. 2024. A proposed s.c.o.r.e. evaluation framework for large language models: Safety, consensus, objectivity, reproducibility and explainability. *arXiv preprint arXiv:2407.07666*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivi'ere, M. S. Kale, and J. Love. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

N. Thakur, N. Reimers, A. R""uckl'e, A. Srivastava, and I. Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Aravinthan J. Thirunavukarasu, Daniel Shu Wei Ting, Karthik Elangovan, Luis Gutierrez, Thelma F. Tan, and Daniel Shu Jin Ting. 2023. Large language models in medicine. *Nature Medicine*, 29:1930–1940.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jaeger. 2024. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*. NeurIPS 2024 (spotlight).

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations

in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Phuc Phan Van, Dat Nguyen Minh, An Dinh Ngoc, and Huy Phan Thanh. 2024. Rx strategist: Prescription verification using llm agents system. *Preprint*, arXiv:2409.03440.

Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 7(1):41.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Xiaosong Wang, Yifan Peng, Le Lu, and 1 others. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Christopher Y. K. Williams, Jaskaran Bains, Tianyu Tang, Kishan Patel, Alexa N. Lucas, Fiona Chen, Brenda Y. Miao, Atul J. Butte, and Aaron E. Kornblith. 2025. Evaluating large language models for drafting emergency department encounter summaries. *PLOS Digital Health*, 4(6):e0000899.

World Health Organization. 2025. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. `https://www.who.int/publications/i/item/9789240084759`. Accessed Oct 5, 2025.

K. Wu, R. Wang, H. Quan, C. Lin, E. Lo, H. Tu, F. Yu, and H. T. Lin. 2025. An automated framework for assessing how well large language models cite relevant medical references. *Nature Communications*, 16(1):427.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2025. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *Preprint*, arXiv:2402.13178.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May D. Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Xin Liu, Carl Yang, Yang Xie, and Wenqi Shi. 2025a. Medagentgym: A scalable agentic training environment for code-centric reasoning in biomedical data science. *Preprint*, arXiv:2506.04405.

Zhikun Xu, Ming Shen, Jacob Dineen, Zhaonan Li, Xiao Ye, Shijie Lu, Aswin RRV, Chitta Baral, and Ben Zhou. 2025b. Tow: Thoughts of words improve reasoning in large language models. *Preprint*, arXiv:2410.16235.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, Li Zhu, and Xuandong Zhao. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *Preprint*, arXiv:2406.13890.

Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S. Bitterman, Jasmine C. L. Ong, Daniel S. W. Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(2):2.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2025a. Justice or prejudice? quantifying biases in llm-as-a-judge. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR 2025 Poster; arXiv:2410.02736.

Xiao Ye, Shaswat Shrivastava, Zhaonan Li, Jacob Dineen, Shijie Lu, Avneet Ahuja, Ming Shen, Zhikun Xu, and Ben Zhou. 2025b. Cc-learn: Cohort-based consistency learning. *Preprint*, arXiv:2506.15662.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10.

Erjian Yu, Xiaofeng Chu, Wei Zhang, Xiangbin Meng, Yaodong Yang, Xue Ji, and Chao Wu. 2025. Large language models in medicine: Applications, challenges, and future directions. *International Journal of Medical Sciences*, 22(11):2792–2801.

Matthew Yue, Zhikun Xu, Vivek Gupta, Thao Ha, Liesal Sharabi, and Ben Zhou. 2025. Relate-sim: Leveraging turning point theory and llm agents to predict and understand long-term relationship dynamics through interactive narrative simulations. *Preprint*, arXiv:2510.00414.

T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D. W. Bates, and R.-E. E. Abdulnour. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1).

Cyril Zakka, Joseph Cho, Gracia Fahed, Rohan Shad, Michael Moor, Robyn Fong, Dhamanpreet Kaur, Vishnu Ravi, Oliver Aalami, Roxana Daneshjou, Akshay Chaudhari, and William Hiesinger. 2024. Almanac copilot: Towards autonomous electronic health record navigation. *Preprint*, arXiv:2405.07896.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

N. Zhang, M. Chen, Z. Bi, X. Liang, L. Li, X. Shang, K. Yin, C. Tan, J. Xu, and F. Huang. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR 2020)*.

Y. Zhang, M. Lang, J. Jiang, Z. Gao, F. Xu, T. Litfin, K. Chen, J. Singh, X. Huang, and G. Song. 2024a. Multiple sequence alignment-based rna language model and its application to structural inference. *Nucleic Acids Research*, 52(1).

Yubo Zhang, Shudi Hou, Mingyu Derek Ma, Wei Wang, Muhao Chen, and Jieyu Zhao. 2024b. CLIMB: A benchmark of clinical bias in large language models. *Preprint*, arXiv:2407.05250.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(1):909.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS 2023 Datasets and Benchmarks Track*.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024. Conceptual and unbiased reasoning in language models. *Preprint*, arXiv:2404.00205.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR 2025 Spotlight; arXiv:2310.17631.

A Appendix

A.1 Full Table

A.2 LLM Usage

We used LLMs for language polishing and organization only; all technical claims and citations were authored and verified by the authors

Table 3: Expanded benchmarks by task class. Each row lists what the task can establish for clinical readiness, common evaluation metrics, the target level (L0–L3), and representative benchmarks.

Task	Usefulness	Metric	Level	Benchmark
Knowledge recall / exam Q&A	Breadth of factual recall and clinical reasoning on multiple-choice or open-ended exam questions across specialties and difficulty levels. Evaluates recall under constrained formats and coverage gates, and whether models hallucinate when unsure (Arora et al., 2025).	Exact match/F1; accuracy; subject/difficulty slices; contamination checks; calibration on unanswerable or uncertain questions.	L0	*MedMCQA* (Pal et al., 2022); *MultiMedQA* (MedQA, MedMCQA, PubMedQA, MMLU) (Singhal et al., 2023); *Mirage RAG suite (MMLU-Med, MedQA-US, PubMedQA, BioASQ Y/N)* (Xiong et al., 2024)
Summarization / transformation	Fidelity and completeness of summaries or transformations of clinical notes, conversations, or research literature. Tests whether models can produce coherent, structurally complete summaries without omissions or hallucinations (DeYoung et al., 2021; Aali et al., 2025).	Hallucination/omission rate; ROUGE/BERTScore/chrF; section completeness; clinical correctness; expert ratings.	L1	MS^2 (DeYoung et al., 2021); *MIMIC-IV-BHC* (Aali et al., 2025); *MTS-Dialog* (Ben Abacha et al., 2023); *ACI-Bench* (Yim et al., 2023)
Retrieval-augmented QA	Attribution and faithfulness of answers to retrieved sources and freshness/recency of information. Evaluates how well models retrieve and ground answers in relevant documents (Xiong et al., 2024).	Faithfulness/attribution; source contradiction rate; Recall@k, nDCG, MRR; answer correctness; freshness.	L0,L1	*Mirage RAG benchmark (MMLU-Med, MedMCQA, PubMedQA, BioASQ)* (Xiong et al., 2024); *HealthSearchQA (part of MultiMedQA)* (Singhal et al., 2023)
Evidence based fact checking	Reliability of claims and ability to verify or refute medical statements using evidence. Useful for ensuring LLM outputs do not propagate misinformation.	Claim classification accuracy; evidence recall/precision; F1 for true/false/unfounded labels; citation quality.	L0,L1	*MedFact* (Chen et al., 2025b)
Information extraction / coding	Structured accuracy on entity recognition, relation extraction, coding, and normalization tasks. Establishes ability to extract structured data from unstructured texts (Luo et al., 2022; Li et al., 2016; Uzuner et al., 2011; Zhang et al., 2021).	Mention/cluster F1; relation F1; coding/normalization accuracy; entity linking accuracy.	L1	*BioRED* (Luo et al., 2022); *BC5CDR* (Li et al., 2016); *n2c2 2010 (i2b2)* (Uzuner et al., 2011); *CBLUE* (Zhang et al., 2021)

Table 3: Expanded benchmarks by task class (continued).

Task	Usefulness	Metric	Level	Benchmark
Decision support / triage (simulation)	Selective reliability for clinical decision making: calibration at deployable thresholds, risk coverage trade-offs, harm proxies, and quantitative reasoning. Includes simulation of triage, diagnosis, personalized diabetes management, and medical calculations (Arora et al., 2025; Bedi et al., 2025a; Cardei et al., 2025; Mehandru et al., 2025; Khandekar et al., 2024).	ECE; Brier; NLL; risk coverage curves; contraindication/near miss rates; accuracy, groundedness, safety, clarity, actionability; MAE for calculations.	L2	*HealthBench* (Arora et al., 2025); *MedHELM* (Bedi et al., 2025a); *DexBench* (Cardei et al., 2025); *ER-Reason* (Mehandru et al., 2025); *MedCalc-Bench* (Khandekar et al., 2024)
Clinical dialogue	Communication quality and human factors in multi-turn doctor patient conversations or simulated OSCE interviews. Measures goal completion, uncertainty marking, empathy, and adherence to safety rails; also covers note generation from visit dialogues (Zeng et al., 2020; Yim et al., 2023; Fareez et al., 2022).	Goal completion; uncertainty/hedging tags; rubric-based ratings; guideline-contradiction flags; empathy/communication scores; note-generation quality; ROUGE/BERTScore.	L2,L3	*MedDialog* (Chinese/English) (Zeng et al., 2020); *MTS-Dialog* (Ben Abacha et al., 2023); *ACI-Bench* (Yim et al., 2023); *OSCE simulated interview dataset* (Fareez et al., 2022)
Multimodal (imaging + text)	Linkage between radiology images and free text reports or classification labels; evaluates image understanding, report generation, and cross-modal retrieval (Johnson et al., 2019; Wang et al., 2017).	Report correctness; finding detection/linking; classification accuracy; precision/recall/F1; bounding box/segmentation metrics.	L2,L3	*MIMIC-CXR*, *MIMIC-CXR-JPG* (Johnson et al., 2019); *NIH ChestX-ray* (Wang et al., 2017)
Patient retrieval	Ability to retrieve relevant literature or similar patient summaries to support clinicians. Tests retrieval quality and ranking of semantically similar patients or articles (Zhao et al., 2023).	Recall@k; nDCG; MRR; patient-similarity accuracy; retrieval precision.	L1	*PMC-Patients* (Zhao et al., 2023)
Molecular / drug discovery	Validity and diversity of generated molecules and optimization of proxy properties for research settings.	Validity/diversity/novelty proxies; property-optimization success; synthetic accessibility; logP and QED scores.	L3	*MOSES* (Polykovskiy et al., 2018); *GuacaMol* (Brown et al., 2019)

A.3 Autonomy Escalation Criteria

Current level	Next level	Example evidence required before escalation
L0 – Inform	L1 – Information Transformation & Aggregation	Stable performance on medical Q&A and patient-facing explanation benchmarks in the intended specialties; hallucination and unsafe-content rates below predefined thresholds on representative L0 queries; readability and lay appropriateness confirmed by human raters; and risk-controlled selectivity (ability to abstain or defer) for out-of-scope informational queries.
L1 – Information Transformation & Aggregation	L2 – Decision Support	High and stable extraction / normalization performance for clinically critical fields (e.g., problems, medications, laboratory results, temporality) on internal and public test sets; coverage of mandatory fields with low rates of clinically important omissions; RAG evaluations showing that responses are grounded in retrieved evidence with low contradiction to source rates; privacy and provenance checks demonstrating acceptable de-identification risk and machine readable lineage; and, for target L2 tasks, pilot studies showing that clinicians can reliably base case workups on the transformed EHR substrate.
L2 – Decision Support	L3 – Agents under Human Supervision	All L2 requirements satisfied for the tasks in scope, including calibrated risk coverage with a tuned abstention policy; simulation or shadow deployments showing high end-to-end task success under human oversight for representative workflows; tool-use correctness audits (API choice, parameters, sequencing) with low failure and silent-error rates; clinician-in-the-loop studies indicating acceptable edit / override rates and no strong evidence of harmful automation bias; and complete, queryable audit trails (e.g., FHIR Provenance / AuditEvent) for data access, prompts, model invocations, tool calls, and approvals.

Table 4: Example autonomy escalation criteria: illustrative evidence requirements for moving a medical LLM system from one autonomy level (L0–L2) to the next (L1–L3). We intentionally do not prescribe specific numerical thresholds here, because acceptable error rates, coverage targets, and abstention rates depend on deployment context, task criticality, baseline performance of existing workflows, and local regulatory or institutional standards. Concrete numerical criteria should therefore be instantiated by deployers in collaboration with clinical, safety, and governance stakeholders and updated over time as evidence, norms, and models evolve.

A.4 Survey Comparison

Table 5: Comparison between our survey and prior medical LLM evaluation surveys focusing on specific aspects of evaluation.

Survey	Focus of prior survey	Difference from our autonomy-aware survey
Evaluating large language models and agents in healthcare: Key challenges in clinical applications (Chen et al., 2025c)	Examines the evaluation of large language models and agents in healthcare, emphasizing agentic, tool-using scenarios and challenges in clinical applications.	Provides a detailed discussion of agentic and workflow-oriented use cases, but does not organize evaluations by levels of autonomy or explicitly link metrics to escalation decisions across levels. Our survey structures agentic evaluations along an autonomy ladder and specifies the required evidence for each level.
A systematic review of large language model (LLM) evaluations in clinical medicine (Shool et al., 2025)	Systematically reviews evaluations of LLMs in clinical medicine, with a strong emphasis on benchmark and task performance, often centered on accuracy-heavy metrics.	Covers a broad range of clinical evaluations but primarily treats metrics as task-level accuracy and performance measures, without structuring them by autonomy level or mapping them to deployment decisions. We instead embed such metrics within an autonomy-aware evaluation framework.
A framework for human evaluation of large language models in healthcare derived from literature review (QUEST) (Tam et al., 2024)	Proposes the QUEST framework for human evaluation of LLMs in healthcare, focusing on dimensions such as quality, utility, and safety of model outputs as judged by human raters.	Provides rich guidance on human evaluation metrics but does not systematically connect these metrics to levels of autonomy or to the escalation of model responsibilities. Our survey incorporates human evaluation as one component of the evidence required at each autonomy level.
A Proposed S.C.O.R.E. Evaluation Framework for Large Language Models: Safety, Consensus, Objectivity, Reproducibility and Explainability (Tan et al., 2024)	Introduces a multidimensional framework (S.C.O.R.E.) for evaluating LLMs along safety, consensus, objectivity, reproducibility, and explainability.	Offers a multidimensional view of evaluation but does not explicitly map these dimensions to an autonomy axis or specify what level of S.C.O.R.E. evidence is required before a model can be escalated to higher-impact roles. Our survey directly ties such multidimensional metrics to autonomy levels and escalation criteria.
Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI (Abbasian et al., 2024)	Defines foundation metrics for evaluating the effectiveness of healthcare conversations powered by generative AI, focusing on conversation quality and interaction outcomes.	Concentrates on conversational metrics rather than autonomy-aware evaluation, and does not map conversational evidence to what the system is allowed to do in clinical workflows. Our framework relates conversational evaluation to specific autonomy levels and associated permissions.

Table 6: Comparison between our survey and additional related surveys identified in our extended search.

Survey	Focus of prior survey	Difference from our autonomy-aware survey
The future landscape of large language models in medicine (Clusmann et al., 2023)	Discusses the broader landscape of how large language models could fit into medicine, including opportunities, risks, and potential future directions for clinical integration.	Provides a broad, forward-looking view of LLMs in medicine but does not map evaluations to specific levels of autonomy, making it difficult to decide when to escalate an LLM from a chatbot to higher-impact roles in real-world applications. Our survey explicitly aligns evaluations and metrics with autonomy levels.
A scoping review of large language models for generative tasks in mental health care (Hua et al., 2025)	Reviews the use of LLMs for generative tasks in mental health care, highlighting applications, challenges, and issues such as non-standard and ad-hoc evaluation scales.	Identifies important evaluation problems (e.g., heterogeneous and ad-hoc scales) but does not specify what concrete evidence is sufficient to deploy a system in real mental health settings. We instead specify the necessary evaluations for each task within each autonomy level.
The application of large language models in medicine: A scoping review (Meng et al., 2024)	Synthesizes 550 studies on LLMs in medicine, characterizing the gap between benchmark-focused evaluations and real-world deployment across many clinical and research settings.	Clearly documents the benchmark–deployment gap but does not provide a concrete plan to close it. In contrast, we pair evaluation metrics with corresponding checks for each autonomy level, turning this gap into a deployable test plan.
Large language models in medicine (Thirunavukarasu et al., 2023)	Surveys opportunities, risks, and potential applications of LLMs in medicine, with emphasis on clinical use cases and safety concerns.	Discusses opportunities and risks across medical research and practice but does not specify evaluations tied to what the model is allowed to do in clinical workflows. Our framework links required evidence directly to permitted actions at each autonomy level.
A Systematic Review of Large Language Models in Medical Specialties: Applications, Challenges and Future Directions (Alkalbani et al., 2025)	Systematically reviews LLM applications across medical specialties, reporting heterogeneous results and emphasizing the need for standards in evaluation and deployment.	Catalogs applications and argues that standards are necessary, but does not describe which additional evidence is required at each level of autonomy. We identify the evidence and checks required before escalation between levels.
A comprehensive survey of large language models and multimodal large language models in medicine (Xiao et al., 2025)	Offers general guidance on medical LLMs and MLLMs, including architectures, datasets, applications, and evaluation, with emphasis on multimodal capabilities.	Provides broad evaluation guidance for LLMs and MLLMs but is not autonomy-aware and does not indicate what evidence must be demonstrated before escalation. Our survey is explicitly structured around autonomy levels and escalation criteria.
Large Language Models in Medicine: Applications, Challenges, and Future Directions (Yu et al., 2025)	Reviews applications, challenges, and future directions of LLMs in medicine, and calls for standardized benchmarks and multimodal evaluations.	Calls for standardized benchmarks and multimodal evaluations, but does not provide a schema that specifies what additional evidence is needed for each level of autonomy. In contrast, we specify which checks are necessary at each level and how they relate to real-world deployment decisions.

A.5 Case Study & Worked example: Evaluating an EHR copilot across autonomy levels

In this worked example, we show how the same high-level application—an EHR copilot for outpatient anticoagulant management—can be instantiated and evaluated at different autonomy levels (L0–L3). The goal is to make concrete how our autonomy lens guides (i) which actions the system is allowed to take, and (ii) which evidence and metrics are required at each level, building directly on the definitions in Section 4.

Clinical scenario. A primary care clinician is starting a new oral anticoagulant for a patient with non-valvular atrial fibrillation. The overall workflow includes (1) educating the patient about the medication, (2) reviewing the chart for contraindications and drug to drug interactions, (3) proposing a dosing and monitoring plan, and (4) drafting orders and patient-facing instructions in the EHR for clinician sign-off. We consider four configurations of an EHR copilot for this scenario, one per autonomy level.

L0: Inform-only patient education

At autonomy level L0, the copilot acts purely as an informational tool. It does *not* access patient-specific EHR data and does *not* give individualized recommendations. Instead, it:

- answers general questions such as "What is an anticoagulant?", "Why is this medication prescribed?", or "What are common side effects?";
- rewrites clinician-authored discharge instructions into plain language;
- explains standard lab tests (e.g., "What does a kidney function test measure?").

The outputs are generic and include explicit disclaimers (e.g., "This is general information and not personal medical advice.").

Evaluation at L0. For this configuration, evaluation targets match our L0 focus in Section 4.1:

- **Factual accuracy and completeness**: question answering accuracy on consumer-health datasets (e.g., HealthSearchQA-style corpora), coverage of key points in patient education materials.
- **Readability and organization**: readability indices and human ratings of clarity, structure, and appropriateness of language for lay readers.
- **Safety and tone**: rubric-based assessment of whether answers avoid giving individualized advice, avoid discouraging guideline-concordant care, and encourage seeking professional help when appropriate.

No calibration, personalization, or tool-use metrics are required, because the system does not see patient data and cannot act on the record.

L1: Information transformation and aggregation over the EHR

At autonomy level L1, the copilot is allowed to read the patient's EHR but remains descriptive and non-prescriptive. It can:

- extract and normalize key facts from the chart (problems, active medications, allergies, recent labs, relevant imaging reports) into a standardized schema (e.g., FHIR resources or an OMOP-style common data model);
- generate a concise clinical summary highlighting bleeding and thrombotic risk factors (e.g., prior gastrointestinal bleeding, renal function, age, concomitant antiplatelets);
- retrieve and surface relevant guideline snippets or drug monographs via retrieval-augmented generation, with inline citations.

The copilot presents *what is known* and *what the literature says* but does not propose a concrete plan for this patient.

Evaluation at L1. Relative to L0, evaluation now needs to cover transformation quality, aggregation, and provenance, as outlined in Section 4.2:

- **Extraction and normalization**: span-level precision/recall/F_1 for entities (problems, medications, allergies, labs) on de-identified note benchmarks; concept mapping accuracy to standard vocabularies.
- **Retrieval and aggregation**: retrieval metrics (e.g., Recall@k, MRR, nDCG) for guideline and evidence search, plus grounded-answer and contradiction-to-source rates for generated summaries.

- **Completeness**: coverage of clinically critical fields in the structured snapshot (e.g., proportion of cases where all guideline-relevant risk factors were surfaced at least once).

- **Privacy and lineage**: documentation that transformed views and summaries are backed by traceable data sources (e.g., pointers to underlying FHIR resources) and that de-identification or access control is appropriate for the use case.

The system is still not making recommendations, so decision level safety and calibration remain out of scope. However, poor completeness or misattribution at L1 would propagate to higher levels, motivating careful measurement here.

L2: Patient-specific anticoagulant decision support

At L2, the copilot is allowed to combine the structured EHR view and external evidence to generate *patient-specific recommendations*, but it still does not call ordering APIs. For the anticoagulant scenario, it may:

- propose one or more anticoagulant options and doses, explicitly citing renal function, weight, comorbidities, and concomitant medications;

- recommend baseline and follow-up labs (e.g., creatinine at specified intervals) and clinic follow-up timing;

- flag potential contraindications or drug–drug interactions and suggest alternative options or escalation to a specialist;

- generate tailored patient-facing explanations (e.g., "Because of your kidney function and other medications, drug X at dose Y is recommended over drug Z.").

The clinician remains the final decision-maker: the copilot's output is an advisory note the clinician can accept, modify, or ignore.

Evaluation at L2. Here, evaluation must extend beyond content quality to cover reasoning, calibration, and harm-proximal safety, as emphasized in Section 4.3:

- **Diagnostic and therapeutic correctness**: accuracy of recommended drug choice, dose, and monitoring plan on vignette-style cases scored against guideline-consistent reference answers or expert panels.

- **Reasoning consistency and faithfulness**: rubric-based scoring of stepwise rationales (e.g., does the explanation correctly reference the factors that actually justify the recommendation?), and stability of recommendations under meaning-preserving prompt variants.

- **Calibration and abstention**: risk–coverage curves over cases when the copilot is allowed to abstain (e.g., "insufficient information, escalate to specialist"), reporting accuracy on answered cases versus coverage, and evaluating whether uncertain or borderline cases are correctly flagged.

- **Safety metrics**: rates of contraindicated recommendations, guideline violations, and other harm-proximal errors (e.g., recommending a standard dose despite severely reduced renal function).

- **Fairness slices**: performance and safety broken down by clinically relevant subgroups (e.g., age, sex, comorbidity patterns) to detect systematic gaps that simple average accuracy would miss.

Importantly, even very strong vignette performance does not by itself justify L3 deployment: as long as the system only produces text and does not invoke tools, it remains at L2.

L3: Tool-using EHR copilot under explicit human supervision

At L3, the copilot augments its L1 and L2 capabilities with controlled tool use. In our scenario, this may include:

- querying EHR APIs to confirm the most recent creatinine, weight, and medication list, rather than relying solely on the clinician's current view;

- drafting medication and lab orders (e.g., drug, dose, frequency, start date, lab panel and schedule) in a "pending" state in the order entry system;

- preparing structured follow-up tasks (e.g., reminders, referral orders) and patient messages or visit summaries;

- logging provenance (which data, which model, which tools) for each drafted action.

Crucially, all drafted actions are *queued for explicit clinician review and sign-off*. The agent is not allowed to execute orders or send messages autonomously; it operates "under supervision" as in our L3 definition.

Evaluation at L3. For L3, the primary target shifts to supervised action quality and human–AI interaction, as discussed in Section 4.4:

- **End-to-end task success**: in simulation or shadow mode, the proportion of encounters where the agent-produced plan and drafted actions are judged clinically acceptable, and the degree to which they reduce manual workload (e.g., reduction in time spent on order entry and documentation).
- **Tool/API correctness**: validity of API calls (correct endpoints, parameters, and patient identifiers), syntactic correctness (e.g., FHIR-conformant requests), and semantic correctness (the data retrieved or orders drafted match the intended patient and plan).
- **Oversight effectiveness**: rates at which clinicians accept, modify, or reject drafted actions; qualitative or quantitative signals of over-reliance (rubber-stamping) or alert fatigue; changes in error rates with versus without the copilot.
- **Auditability and governance**: completeness of machine-readable provenance (e.g., FHIR AuditEvent and Provenance resources) that tie each action to its inputs, prompts, model versions, tools, and human approvals, enabling post-hoc investigation of failures or near-misses.
- **Operational safety**: monitoring for rare but high-impact failure modes in pilot deployments (e.g., mis-addressed messages, wrong-patient orders) and documenting mitigation strategies (access controls, guardrails, kill-switches).

These measurements go beyond traditional benchmark scores. They are required to justify even supervised L3 use, because errors now manifest as mis-drafted actions rather than just incorrect text.

Summary of the worked example

This worked example illustrates how autonomy levels structure both system design and evaluation:

- L0 constrains the copilot to generic education, evaluated on accuracy, readability, and safe communication.
- L1 adds EHR-based transformation and aggregation, evaluated on extraction quality, completeness, retrieval, and provenance.
- L2 introduces patient-specific recommendations, evaluated on correctness, reasoning, calibration, safety, and fairness.
- L3 adds tool use under supervision, evaluated on supervised action quality, tool correctness, oversight, auditability, and operational safety.

Framing the same anticoagulant-management copilot across L0–L3 shows how our autonomy taxonomy operates as a practical blueprint: it makes the system's allowed actions explicit, and aligns those actions with level-appropriate evidence and metrics, turning benchmark performance into credible, risk-aware claims about clinical utility.