

LLMs and Agents in Pharmacovigilance: Division of Labour and Governance

Qiang Wang¹, Great Unicorn Intelligence Research Team^{2*}

Abstract Pharmacovigilance (PV) aims to detect new drug–adverse event associations early and have qualified people assess them. Reviewing recent studies, preprints and regulatory documents, we map where LLMs and agents fit across the workflow, from case processing to causality assessment. Extraction precision is high on well-constrained tasks; MedDRA errors mainly involve non-existent lowest level terms; and expert agreement on causality remains limited, especially for the intermediate "possible" category. Using disproportionality formulas with reproducible illustrative data, we argue that signal statistics belong in deterministic code, and propose a gating framework tiered by model influence and decision consequence.

Keywords: pharmacovigilance; individual case safety report; MedDRA coding; signal detection; large language model; human review; regulatory science

Introduction

Pre-marketing clinical trials are limited in size and duration and enrol selected populations, so rare adverse reactions, risks of long-term exposure and risks in special populations often emerge only after approval. Pharmacovigilance is the body of science and practice built to close that gap. Module IX of the EU Good Pharmacovigilance Practices (GVP) defines a signal as information arising from one or multiple sources, including observations and experiments, which suggests a new potentially causal association, or a new aspect of a known association, between an intervention and an event, and which is judged sufficiently likely to justify verifactory action¹. China's Good Pharmacovigilance Practice took effect on 1 December 2021 and requires marketing authorisation holders to establish and continuously improve a pharmacovigilance system².

Day to day, this system runs on a great deal of manual reading, judging, data entry and review. A spontaneous report must first be checked against the minimum criteria for a valid ICSR; its patient, product, event, timing and outcome elements must be extracted; the event description must be coded in MedDRA; seriousness and expectedness must be determined; a case narrative must be written; and the case must be submitted in ICH E2B(R3) format within the regulatory timeline^{3,4}. New sources such as literature monitoring, social media and patient support programmes keep enlarging the inflow. When ICH E2D(R1) was finalised in September 2025, it added specific rules for digital platforms, including social media, websites, forums and apps³.

Author information: ¹ Qiang Wang, Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd.; research interests: artificial general intelligence, computer vision; Shenzhen, Guangdong, China.

² Great Unicorn Intelligence Research Team, Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd.; research interests: AI agents, AGI training; Shenzhen, Guangdong, China. Corresponding author. E-mail: research@hxallai.com.

LLMs have turned machine reading of free text from a research topic into a deployable capability. But PV is a heavily regulated field whose purpose is patient safety. Extract the wrong drug name and a case may vanish from the statistics; choose the wrong term and a signal may be diluted; write "probable" where the evidence supports only "possible" and a risk-management decision may shift. The CIOMS Working Group XIV report, published in December 2025, warns that automation pursued solely for efficiency and financial targets may remove the value added by humans in the loop and ultimately compromise the PV system's distinctive strength, which was designed to capture data for individual-case causality assessment⁵.

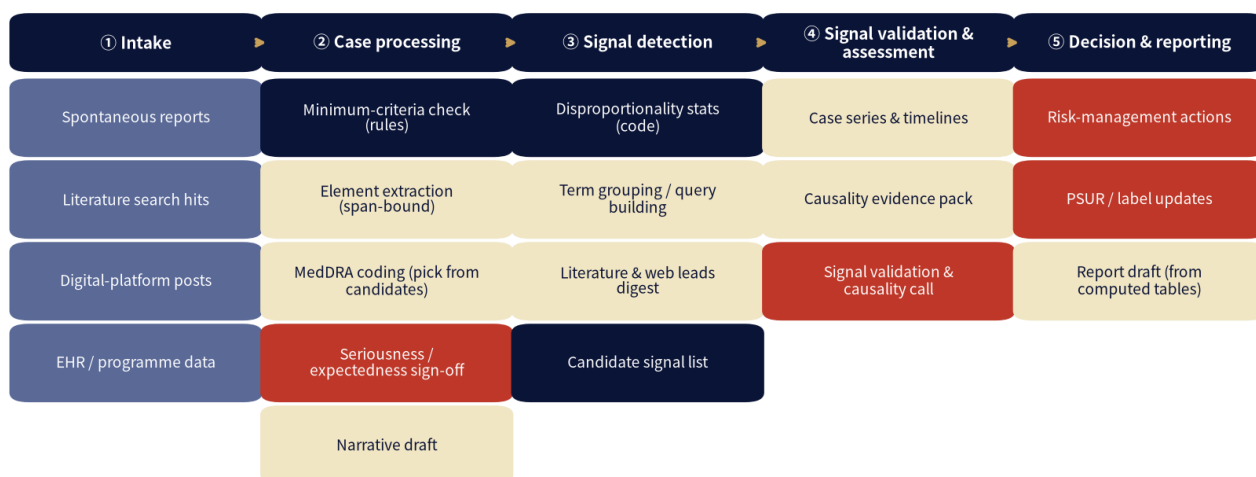
The question this paper addresses is therefore not whether LLMs can be used in pharmacovigilance, but: at which stage, in what way and under what controls can LLMs and agents raise efficiency without lowering the quality of safety judgements?

We make four contributions:

1. We organise the 2025–2026 evidence by the five stages of PV and place the metrics self-reported by each study in one figure, showing how strongly task type determines performance;
2. We give the formulas for four disproportionality measures (PRR, ROR, IC and EBGM) with a reproducible illustrative dataset, explain what shrinkage does, and set out what LLMs should and should not do in signal detection;
3. We translate the requirements of CIOMS, ICH, FDA and EMA documents into system design constraints and propose a gating framework tiered by model influence and decision consequence;
4. We state Great Unicorn Intelligence's viewpoint and give predictions for the next two to three years, each with its rationale and a method for testing it.

All model performance figures quoted here are self-reported by the original studies under their own datasets and settings. Because metrics, data and evaluation protocols differ, they cannot be compared directly across studies. All worked examples use clearly labelled illustrative data and involve no real company or patient data.

Where LLMs and agents enter the pharmacovigilance workflow, and where humans decide



Data source
LLM drafts + human review
Named-person decision
Deterministic code / rules

Colours show the division of labour proposed in this paper (opinion); based on ICH E2D(R1), GVP Module IX and the CIOMS WG XIV report

Fig. 1 | Where LLMs and agents enter the PV workflow, and where human gates sit. Compiled by the authors from ICH E2D(R1), GVP Module IX and the CIOMS report^{1,3,5}; colours indicate the division of labour proposed in this paper and represent our opinion.

Background and Related Work

PV data standards: ICSR, E2B(R3), MedDRA

The individual case safety report (ICSR) is the basic data unit of PV. ICH E2D(R1) defines it as a description of an adverse event, adverse drug reaction or other observation in an individual patient at a specific point in time, and states that a reportable ICSR must meet four minimum criteria: at least one adverse event or other observation, at least one suspect or interacting medicinal product, an identifiable patient and an identifiable reporter³. These four criteria decide whether a document constitutes a case, and therefore the day from which the regulatory clock starts.

ICH E2B(R3) specifies the data elements and message format for electronic ICSR transmission, covering transmission identification, primary source, literature references, study information, patient characteristics, reactions and events, test results and drug information⁴. For LLMs, E2B(R3) matters because it provides a ready-made, regulator-recognised output structure. Extraction is not "summarise the case freely"; it is "populate a structure whose fields and value ranges are defined".

MedDRA (the Medical Dictionary for Regulatory Activities) is the standard medical terminology maintained under ICH. According to the Introductory Guide for MedDRA Version 29.1 (September 2026), it has a five-level hierarchy (System Organ Class, SOC; High Level Group Term, HLGT; High Level Term, HLT; Preferred Term, PT; and Lowest Level Term, LLT) with 27 SOCs; each LLT links to exactly one PT, each PT has a primary SOC, and LLTs can serve as the basis for auto-encoding⁶. Coding is usually performed at LLT level, while analysis and retrieval are usually performed at PT level or above. New MedDRA versions are released periodically, coded data must be traceable to the version used, and version management is itself a compliance requirement.

ICH E2E sets out the structure of pharmacovigilance planning, including the safety specification (identified risks, potential risks and missing information) and the pharmacovigilance plan⁷. It is a reminder that signal detection and assessment are not free-standing data mining but planned activities organised around a product's known risks and knowledge gaps.

From rules and classical NLP to LLMs

Text processing in PV is not new. Rules, dictionaries and statistical models have long been used to identify and code adverse reactions; they are transparent and efficient but struggle with the linguistic variability of clinical narratives. Pre-trained language models and LLMs understand context better and adapt more readily, yet their regulatory acceptance is limited by hallucination, reduced transparency and reproducibility, sensitivity to parameter settings and the continuing need for human-in-the-loop validation⁸.

Three recent reviews provide a broad picture. Kim et al., following PRISMA 2020, searched studies published from January 2022 to March 2026 and included 83 empirical studies. LLM applications were

concentrated in constrained extraction and classification tasks, including signal evaluation, clinical-note extraction, social media surveillance and literature screening. Most studies were retrospective, benchmark-based or proof-of-concept; hallucination, low specificity, prompt sensitivity, narrow datasets and weak external validation were common limitations. The authors conclude that current evidence supports supervised, task-specific use of LLMs rather than autonomous PV decision-making⁹. El Sayed et al. included 67 studies and found that safety-signal detection was the most common application (50.5%), that external validation was the weakest APPRAISE-AI domain (mean score 52%) and that prospective implementation studies were rare¹⁰. Wu et al., writing from the US FDA, propose a framework of "computable PV": routine, well-defined tasks such as completeness checks, duplicate detection and structured extraction are increasingly amenable to automation, whereas case-level causality assessment remains hard to formalise and still relies on expert judgement. LLMs bring cross-task capability but also challenges of reliability, auditability and governance, so hybrid architectures combining large models, small models and rule-based components are becoming necessary¹¹.

Regulatory guidance and requirements

The CIOMS Working Group XIV report (December 2025) is currently the international guidance document devoted specifically to AI in PV. It sets out seven guiding principles (a risk-based approach, human oversight, validity and robustness, transparency, data privacy, fairness and equity, and governance and accountability) and distinguishes human-in-the-loop (HITL, where the decision is the result of human-machine interaction), human-on-the-loop (HOTL, where the machine produces a result that a human then checks) and human-in-command (HIC). The extent and nature of oversight should be proportionate to risk⁵.

The EMA reflection paper on AI in the medicinal product lifecycle (adopted September 2024) states that AI can support adverse event report management and signal detection and that PV may allow a more flexible approach to modelling and deployment, including incremental learning; but it remains the marketing authorisation holder's responsibility to validate, monitor and document model performance and to include AI operations in the PV system¹².

The FDA draft guidance of January 2025, *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products*, proposes a seven-step risk-based credibility assessment framework: define the question of interest; define the context of use (COU); assess model risk; develop a credibility assessment plan; execute the plan; document results and deviations; and determine whether the model is adequate for the COU. Model risk combines two factors: model influence (the contribution of model-derived evidence relative to other evidence) and decision consequence (the significance of an adverse outcome resulting from an incorrect decision)¹³. The draft explicitly lists "identifying, evaluating, and processing for reporting postmarketing adverse drug experience information" among AI uses. In postmarketing PV, process documentation is generally not submitted in advance but maintained under the sponsor's standard operating procedures and made available on request, for example during inspection; sponsors remain responsible for postmarketing safety surveillance and reporting regardless of the technology used¹³. Where a COU involves a human in the loop, evaluation should consider the performance of the human-AI team rather than the model in isolation¹³. At the time of writing, the FDA website still lists the document as a draft.

The ten guiding principles of good AI practice in drug development, issued jointly by FDA and EMA in January 2026, cover human-centric design, a risk-based approach, adherence to standards (including GxP), a clear context of use, multidisciplinary expertise, data governance and documentation, model design and development, risk-based performance assessment, life-cycle management and clear essential information¹⁴. The requirement that data provenance, processing steps and analytical decisions be documented in a detailed, traceable and verifiable manner bears directly on PV audit expectations.

At the level of electronic records, 21 CFR 11.10 requires closed systems to use secure, computer-generated, time-stamped audit trails that independently record the date and time of operator entries and actions that create, modify or delete records, without obscuring previously recorded information, and to use operational system checks to enforce permitted sequencing of steps¹⁵. These two requirements translate almost word for word into design constraints for agent systems: every tool call and every human sign-off must be logged, and the model must not be able to skip steps.

Table 1 summarises what these documents imply for system design.

Table 1 | Key points of PV-related regulatory and professional documents and their implications for LLM/agent system design.

Document	Points relevant to this paper	Design implication
ICH E2D(R1), 2025 ³	Four ICSR minimum criteria; screening of digital platforms and "day zero"	Check minimum criteria with deterministic rules; record when information first met them
ICH E2B(R3) ⁴	ICSR data elements and message format	Constrain extraction output to E2B fields and value ranges, not free text
MedDRA 29.1, 2026 ⁶	Five levels, 27 SOCs, each LLT linked to one PT	Codes must exist in the locked dictionary version; record the version with the case
ICH E2E ⁷	Safety specification and PV plan	Organise signal detection around identified/potential risks and missing information
GVP Module IX ¹	Definition of a signal; signal management process	A statistical "candidate" is not a signal; human validation and assessment required
CIOMS WG XIV, 2025 ⁵	Seven principles; HITL/HOTL/HIC	Choose oversight mode by risk; document the rationale
EMA reflection paper, 2024 ¹²	MAH validates, monitors and documents model performance	Bring model monitoring into the PV system master documentation
FDA draft guidance, 2025 ¹³	Seven-step credibility framework; model influence × decision consequence	Tier each COU separately; evaluate the human–AI team
FDA–EMA principles, 2026 ¹⁴	Traceable data governance; life-cycle management	Data snapshots, model versions, parameters and decisions all traceable
21 CFR 11.10 ¹⁵	Time-stamped audit trails; sequencing checks	Tamper-evident logs; process order enforced by the system
China GVP, 2021 ²	MAH establishes and improves the PV system	Use of AI tools must sit inside the MAH's system and SOPs

Core Methods and Technical Analysis

This section works through the five stages of PV. For each, we describe the nature of the task, review what published studies did and found, and draw out engineering design points.

Automated ICSR extraction

The task. The input is free text (clinical notes, call records, emails, articles, posts); the output is a set of structured case elements: patient information, suspect and concomitant drugs, event description, onset and outcome dates, dechallenge and rechallenge, seriousness criteria and reporter details. The difficulty is that indications, medical history, allergies and genuine adverse events often appear side by side, and the model must separate "why the drug was given" from "what happened after it was given".

Evidence. Ludwig et al., in *Drug Safety*, used a two-pass workflow: the first pass screens an entire electronic health record (EHR) note for adverse drug events; the second adds 20 structured fields to each event¹⁶. Across 372 de-identified notes the model identified 191 events, of which 180 were true positives: 94.2% precision and 84.1% recall (F1 = 88.9%). Seriousness was correct in every case and label status in 93.9%; MedDRA LLT coding was correct for 92.5% of valid events, with errors mostly being LLTs that do not exist. Notably, 84.9% of notes contained no adverse event at the first pass, which kept inference costs to USD 0.18 per note¹⁶.

Plasek et al. compared models for identifying adverse event–drug relationships in inpatient and outpatient notes; GPT-4o achieved an F1 of 0.524, against 0.381 for GPT-4o-mini. Common errors included classifying pre-existing allergies as adverse events and conflating therapeutic indications with adverse effects¹⁷. Both studies address relation extraction, yet their numbers differ widely because task definitions, data sources and evaluation rules (strict versus relaxed matching) differ. "Accuracy" means little without a task definition.

For Chinese text, Ma et al. built a Chinese adverse-event corpus from clinical notes at Xiangya Hospital (a 2,510-record gold standard and a 5,144-entry knowledge base) and compared non-augmented, statically augmented and retrieval-augmented generation (RAG) prompting. DeepSeek-V3 with RAG reached an F1 of 0.9638 at the overlap-matching level; RAG raised GPT-4o's recall from 0.6419 to 0.9241; and on progress notes reflecting real-world prevalence the RAG prompt maintained a specificity of 0.9821¹⁸. For Chinese clinical text, using a curated domain knowledge base as the retrieval source is an effective way to suppress hallucination and lift recall.

In industrial signal assessment, Warner et al. at Eli Lilly used GPT-4o to extract case elements, such as risk factors and responses to dechallenge and rechallenge, from series of ICSR narratives in a machine-first, human-verified workflow. Across five historical signal assessments, F1 ranged from 0.444 to 1.000 for risk factors and from 0.429 to 0.909 for dechallenge and rechallenge; the authors saw potential time savings even after accounting for verification¹⁹. The spread itself is the lesson: the same model performs very differently on different elements, so evaluation before deployment must be element by element, not an overall average.

Design points. Four engineering principles follow:

1. **Constrain the output.** Use E2B(R3) fields as the schema; the model may populate only defined

fields, each with a value range or controlled vocabulary; fields that cannot be confirmed from the source stay empty rather than being inferred.

2. **Bind evidence.** Every populated field carries a pointer to the source span (document ID and character range). Reviewers see not the model's conclusion but the conclusion plus its source.
3. **Decide minimum criteria by rule.** Whether a valid ICSR exists is determined by deterministic rules applied to the extracted elements, with the time of determination recorded; the model is never asked "is this a case?".
4. **Two passes plus stratified review.** A high-recall first pass, a detailed second pass, and review sampled by risk tier; serious cases, fatal cases and special-population cases are all reviewed by a person.

MedDRA coding: retrieve–select–validate

The task. Coding maps the reporter's verbatim description to an LLT, from which the PT and higher levels follow through the dictionary structure. It looks like translation but is really retrieval and selection within a controlled vocabulary: the right answer must be an entry that actually exists in the dictionary.

Evidence. In the Ludwig study, LLT errors were mostly non-existent LLTs¹⁶, a typical failure of generative models, which produce a plausible-sounding medical phrase instead of selecting from the dictionary. Yan et al. at Hengrui Pharmaceutical built a RAG-plus-agent coding framework for clinical data management and validated it on 2,712 unique adverse event terms. Automated coding cut mean coding time from 8.0 to 3.9 minutes per record (48.7%), covered 95.8% of terms and achieved 70% precision against the manual gold standard; spelling-error detection reached 85.1% precision on real-world data, reducing manual review workload by about 70%; compliance relied on local deployment and audit trails aligned with 21 CFR Part 11²⁰. A precision of 70% means every code still needs human confirmation. The value of such systems lies in shortening confirmation time, not in removing confirmation.

For term grouping, the SafeTerm system embeds query terms and MedDRA PTs in the same vector space and uses cosine similarity with extreme-value clustering to produce a ranked PT list. Against 110 tier-1 Standardised MedDRA Queries (SMQs), recall reached 94% at moderate thresholds and precision up to 89% at higher thresholds; at the optimal threshold of 0.70, overall recall was 48% and precision 45%²¹. A similar recall–precision trade-off was observed against 104 FDA Office of New Drugs Custom Medical Queries (OCMQs)²². Such tools work well as candidate generators when building queries; medical staff still decide which PTs to include.

On normalising lay expressions, Zhou et al. evaluated six social media medical concept normalisation datasets and found widespread incorrect mappings, invalid concepts and long-tail distributions; after data-quality enhancement and multi-source fusion, metrics on rare concepts improved by up to 38%²³. The "gold standard" used to train and evaluate coding models needs quality review too.

Design points. A three-step design is advisable:

1. **Retrieve candidates** from the locked MedDRA version with lexical and vector search;
2. **Let the model select** among candidates and state its reasons, with no terms allowed outside the candidate set;

3. **Validate against the dictionary and record the version:** the output must exist in the locked version with "current" status or it is blocked; the version number is written to the case record.

This removes "invented term" errors by construction. What remains is "a real term, wrongly chosen", which can be managed through sampled review and consistency statistics.

Literature and social media monitoring

The task. Literature monitoring requires MAHs to search the scientific literature regularly for individual cases and safety information concerning their products; digital-platform monitoring faces larger volumes of text that is more informal and less complete. ICH E2D(R1) states that MAHs should regularly screen digital platforms under their responsibility at a frequency that allows adverse events to be identified and reported within the required timelines; the regulatory clock (day zero) starts as soon as information meeting the ICSR minimum criteria is posted. MAHs are not expected to screen platforms outside their responsibility, but if they collect data from such a platform in a planned manner, the activity should be treated as an organised data collection system³.

The implication for system design is direct: once an automated monitoring tool is deployed, the moment it "sees" information may become the start of the compliance clock. Monitoring scope, screening frequency and triage deadlines must therefore be fixed in SOPs in advance; the tool must not be allowed to widen its own scope.

Evidence. AskGLP1 shows the scale at which LLMs can process online patient narratives. A MedDRA-grounded, LLM-assisted pipeline screened nearly 8.5 million Reddit posts, distilled 163,135 medication-experience posts and, with 2,506 WebMD reviews, built a knowledge graph of 385,790 drug-side-effect and drug-purpose relations. Online narratives shared the most common gastrointestinal symptoms with FAERS but foregrounded everyday burdens such as fatigue and food aversion, whereas FAERS more often included injection-site reactions and serious diagnoses such as pancreatitis; the agent answered with verifiable evidence and declined requests for medical advice²⁴. Such data help in understanding patient experience and generating hypotheses, but sources cannot be verified and duplication and exaggeration are common, so they should not feed directly into disproportionality statistics.

A second risk comes from the model itself. Using FAERS structured data and a persona-based framework, Liu et al. found that adding clinically irrelevant socio-demographic information (education, housing stability, insurance and so on) systematically shifted adverse-event predictions, with disadvantaged groups often assigned higher likelihoods. Bias appeared both explicitly, with persona attributes cited in reasoning traces, and implicitly, as inconsistent predictions without mention of the attributes²⁵. For PV, the context given to a model should be minimised to what the task needs, and personal information irrelevant to the judgement should not enter the prompt.

Design points. Literature and web-monitoring agents should: fix the list of sources and screening frequency in SOPs; record, for each potential case, when it first met the minimum criteria; label web sources by credibility and store and analyse them separately from formal case data; and strip task-irrelevant personal information from prompts.

Statistics in code, explanation by LLM

The task. Spontaneous reporting databases lack a denominator (the number of exposed patients), so

incidence cannot be computed directly. Disproportionality analysis instead compares within the database: is the proportion of reports for a given event among reports for a given drug clearly higher than the proportion of that event among reports for all other drugs? For each drug–event combination we build a 2×2 table: a is the number of reports with the drug and the event, b with the drug and other events, c with other drugs and the event, and d with other drugs and other events; $N = a + b + c + d$.

Four common measures. Table 2 gives the formulas. PRR and ROR are frequentist measures^{26,27}; the information component (IC) of the Bayesian Confidence Propagation Neural Network (BCPNN) and the EBGM of the Gamma Poisson Shrinker (GPS) are Bayesian measures. What the Bayesian measures share is shrinkage of small counts, which stops one or two chance reports from producing a very large ratio^{28–30}.

Table 2 | Formulas for four disproportionality measures. E is the expected count.

Measure	Formula	Common threshold
PRR ²⁶	$PRR = [a/(a+b)] / [c/(c+d)]$	Evans et al.: $PRR \geq 2$, $\chi^2 \geq 4$ and $a \geq 3$
ROR ²⁷	$ROR = (a \cdot d)/(b \cdot c)$; 95% CI = $\exp[\ln ROR \pm 1.96 \cdot \sqrt{(1/a+1/b+1/c+1/d)}]$	Lower 95% confidence limit > 1
IC (BCPNN) ^{28,30}	$E = (a+b)(a+c)/N$; shrunk IC = $\log_2[(a+0.5)/(E+0.5)]$; IC025 = $\log_2$ of the 2.5% quantile of $\text{Gamma}(a+0.5, E+0.5)$	IC025 > 0
EBGM (GPS) ²⁹	λ = true ratio; prior is a mixture of two gamma distributions; posterior geometric mean EBGM = $\exp(E[\ln \lambda \mid a, E])$; EB05 = posterior 5% quantile	EB05 above a preset threshold (e.g. 2)

Worked example. So that readers can reproduce the figures, we constructed illustrative data for four drug–event combinations (each combination sits in its own illustrative database of $N = 100,000$ reports; the data are fictitious and correspond to no real medicine). In practice the EBGM prior parameters should be fitted to the whole database by maximum likelihood; for demonstration we use $\alpha_1 = 0.2$, $\beta_1 = 0.1$, $\alpha_2 = 2$, $\beta_2 = 4$ and mixing proportion $p = 1/3$. The computation code is archived with this paper.

Table 3 | Worked disproportionality calculations. Illustrative data, no real medicines.

Combination	a	b	c	d	E	PRR	χ^2 (Yates)	ROR (95% CI)	Shrunk IC / IC025	EBGM / EB05
S1 Drug A – elevated liver enzymes	20	980	200	98,800	2.20	9.90	137.7	10.08 (6.34–16.03)	2.93 / 2.22	8.57 / 5.83
S2 Drug B – interstitial lung disease	3	297	30	99,670	0.10	33.23	58.4	33.56 (10.19–110.56)	2.55 / 0.50	13.26 / 4.21
S3 Drug C – headache	60	1,940	2,400	95,600	49.20	1.22	2.3	1.23 (0.95–1.60)	0.28 / –0.11	1.17 / 0.94
S4 Drug D – rhabdomyolysis	8	4,992	40	94,960	2.40	3.80	11.4	3.80 (1.78–8.13)	1.55 / 0.38	2.76 / 1.23

Reproducing S1: $PRR = (20/1,000)/(200/99,000) = 0.020/0.00202 = 9.90$; $ROR = (20 \times 98,800)/(980 \times 200) =$

10.08; $E = 1,000 \times 220 / 100,000 = 2.2$; shrunk IC = $\log_2(20.5/2.7) = 2.93$.

What the table shows. First, S2 rests on only three reports. Its PRR is 33.2, but the confidence interval is very wide (ROR 10.19 to 110.56); shrinkage lowers the IC from an unshrunk 4.92 to 2.55 and the EBGM from about 33 to 13.3. This is exactly what shrinkage is for: combinations with little evidence should not rise to the top by chance. Second, S3 has the most reports (60) but a ratio close to 1, and no method flags it; many reports of a common event do not imply an association. Third, S4 meets the Evans criteria, its IC025 is positive and its EB05 exceeds 1, yet it would not be flagged under an EB05 ≥ 2 threshold. Different methods and thresholds give different answers. Fig. 2 shows how, with the observed-to-expected ratio fixed at 5, the shrunk measures approach the raw ratio as the number of reports grows.

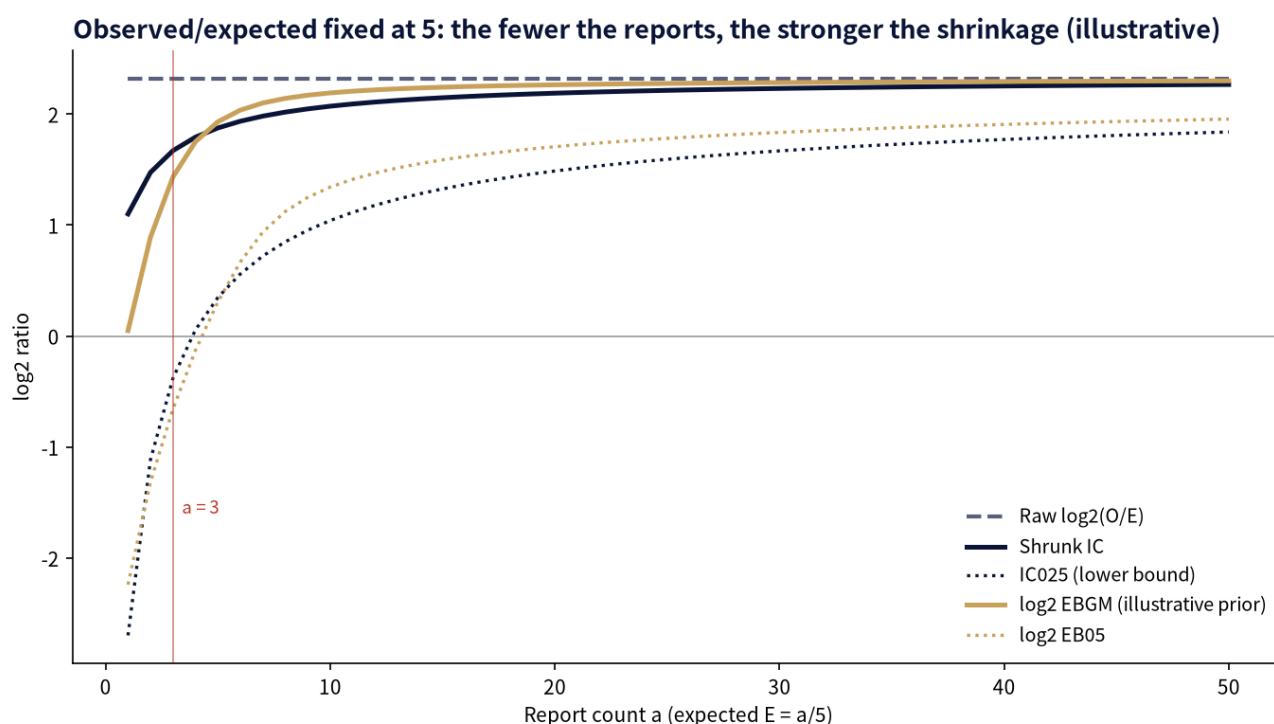


Fig. 2 | Raw ratio versus shrunk estimates at different report counts with the observed-to-expected ratio fixed at 5. Illustrative data; IC computed following Norén et al. 2013³⁰; EBGM uses an illustrative prior²⁹.

Recent statistical developments. Disproportionality analysis is itself evolving. The MCSG method of Bright et al. estimates background rates with a hierarchical Bayesian model and iteratively removes counts with low probability to mitigate masking (where many reports for some drugs or events depress the ratios of other combinations); it outperformed conventional methods on synthetic data and a FAERS subset³¹. Balogh et al. used contrastive learning to learn adverse event representations from spontaneous reports, reaching an AUC of 0.88 on a drug–event association benchmark, compared with 0.76–0.83 for ROR and for representations generated by large language models³². Kefala et al. used DeepSeek V3 to extract adverse reactions from 17,763 label versions of 1,513 centrally authorised EU products, building a time-indexed reference set (110,823 drug–adverse event associations) for evaluating whether signal detection methods can surface signals before regulatory confirmation³³.

Where LLMs belong in signal detection. These developments suggest a clear division of labour: statistics are computed by deterministic code, so results can be reproduced digit for digit and replayed during inspection. LLMs are suited to four supporting jobs: grouping terms, to help build or review the set of PTs used in an analysis²¹; assembling case narratives related to a candidate signal into case-series summaries; building reference sets from labels and the literature³³; and explaining in natural language what statistical results mean and where their limits lie. An LLM should not directly "decide whether a combination is a signal", nor compute ratios "in its head". The pharmacovigilance assistant of Calle et al. offers a template: the system queries FAERS through OpenFDA, computes predefined summary quantities and then generates constrained narratives and charts from the tables already computed; deterministic execution, versioned queries and cached responses allow exact reruns and audits³⁴.

Causality assessment: limited agreement

The task. Causality assessment weighs temporal relationship, dechallenge and rechallenge, alternative causes, prior evidence and objective findings, typically using the WHO-UMC system or the Naranjo scale. It is the stage of PV that depends most heavily on professional judgement.

Evidence. Heckmann et al. compared several biomedical LLMs and prompting strategies on 150 ICSRs; the top-performing combination agreed with human assessors on the final causality category in 64% of cases, with lower agreement on key items such as labelled status, temporal plausibility, alternative causes and objective evidence³⁵. The same group then evaluated Naranjo assessments with GPT-5.2 on 723 FAERS cases: at baseline, agreement was 74.1% on question 5 (alternative causes) and 65.4% on question 10 (objective evidence), and Bayesian optimisation of case-specific temperature raised agreement on causality category from 45.0% to 72.0%³⁶. Ha et al. tested WHO-UMC assessment on 55 FAERS cases (337 drug-level assessments). Cohen's κ between model and experts ranged from 0.368 to 0.641 and accuracy from 0.583 to 0.804; F1 for the "certain", "probable" and "unlikely" categories was about 0.79, 0.79 and 0.90, but only 0.293 for "possible". The authors conclude that LLMs are not suitable for independent decision-making but can assist preliminary triage³⁷.

The MARCA study of October 2026 divided the work among four agents (literature retrieval, label assessment, objective-evidence classification and clinical causal reasoning), with agents generating rationales and humans assisting with scoring. On 215 ICSRs and 1,562 drug–event pairs, agreement on final causality category was 79.4% ($\kappa = 0.587$), but κ for the alternative-causes question was only 0.269³⁸.

Shi et al. revealed another kind of risk. In drug-induced liver injury risk assessment, models answering closed-book seldom abstained when direct evidence was absent (abstention rates of 5.3% to 33.3%) and gave high-confidence "low risk" answers for drugs that do carry liver-injury risk; under an evidence-gated protocol that required answers based only on supplied evidence, every model abstained when direct evidence was absent³⁹. The result shows that "no source, no answer" should be the default setting for drug-safety applications, not an option.

Design points. In causality assessment, the LLM is better cast as the *preparer of assessment materials*: it organises supporting and opposing evidence item by item along the scale, marks the timeline, lists alternative causes and cites the source text for every item. The final causality judgement is made by a named pharmacovigilance physician. The system should record differences between the model's

suggestion and the human conclusion as data for continuous evaluation.

Comparison and Evaluation

Fig. 3 brings together the main metrics self-reported by the studies cited here. To stress again: these figures come from different tasks, data and evaluation rules, and can be used to observe patterns, not to rank models.

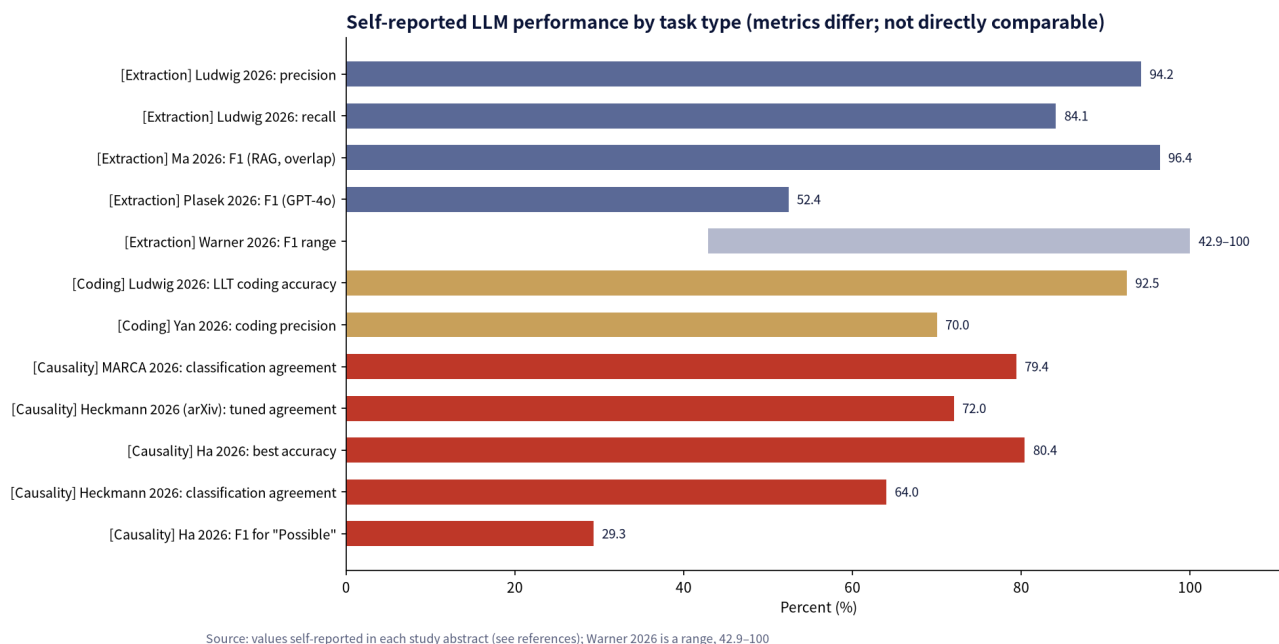


Fig. 3 | Self-reported LLM performance by task type. Data from^{16–20,35–38}; metrics and evaluation rules differ and cannot be compared directly.

Table 4 | Overview of representative studies.

Stage	Study	Data	Main results	Key takeaway
Extraction	Ludwig et al., 2026 ¹⁶	372 EHR notes	Precision 94.2%, recall 84.1%; LLT coding 92.5%	Two-pass; errors mostly non-existent LLTs
Extraction	Ma et al., 2026 ¹⁸	2,510-record Chinese gold standard	F1 0.9638 with RAG	Retrieval augmentation lifts recall markedly
Extraction	Plasek et al., 2026 ¹⁷	505 inpatient, 2,555 outpatient notes	GPT-4o F1 0.524	Indications and allergy history easily confused
Extraction	Warner et al., 2026 ¹⁹	5 historical signal assessments	F1 0.429–1.000	Machine first, human verified
Coding	Yan et al., 2026 ²⁰	2,712 AE terms	Time down 48.7%; precision 70%	Local deployment and audit trails
Grouping	SafeTerm, 2025 ²¹	110 SMQs	Recall 48%, precision 45% at optimal threshold	Suited to candidate generation
Causality	Ha et al., 2026 ³⁷	337 drug-level assessments	κ 0.368–0.641; "possible" F1 0.293	Not for independent decisions

Stage	Study	Data	Main results	Key takeaway
Causality	MARCA, 2026 ³⁸	1,562 pairs	Agreement 79.4%, κ 0.587	Low agreement on alternative causes
Safety boundary	Shi et al., 2026 ³⁹	47 drugs	All abstain without evidence under evidence gating	No source, no answer
Bias	Liu et al., 2025 ²⁵	FAERS personas	Disadvantaged groups predicted at higher risk	Minimise inputs

Three patterns emerge.

Pattern 1: the more closed the task, the better the performance. Extraction tasks, where the answer appears in the source text, can exceed 90% precision; coding tasks, which require selection from a controlled vocabulary, depend on whether the output space is constrained; causality assessment, which requires integrated judgement, shows clearly lower agreement, with intermediate categories especially difficult. This is consistent with the "computable PV" framework proposed by FDA researchers¹¹.

Pattern 2: architecture matters more than the model. Adding retrieval augmentation raised the same model's recall from 0.64 to 0.92¹⁸; evidence gating changed abstention from "rare" to "universal"³⁹; and restricting coding to a candidate set removes invented terms. What determines system reliability is often not which model was chosen but the constraints the model is placed under.

Pattern 3: low prevalence amplifies error. Norén et al. at the Uppsala Monitoring Centre note that many high-stakes applications target low-prevalence events, so apparently high accuracy may conceal limited practical value; unrealistic class balance and a lack of hard positives make test results overly optimistic. They recommend structured case-level review alongside statistical evaluation and cost-sensitive objectives that align model performance with operational value⁴⁰. In the Ludwig study, 84.9% of notes contained no adverse event¹⁶, a picture of low prevalence in practice.

Evaluation recommendations. On this basis we recommend that organisations design evaluation along five lines before introducing an LLM: (1) build test sets at realistic prevalence with hard positives and hard negatives; (2) report metrics field by field and category by category with confidence intervals, which the FDA draft guidance also asks for all performance estimates¹³; (3) evaluate the performance of the human–AI team, including review time and post-review error rate¹³; (4) measure consistency over repeated inference and record model versions and parameters; and (5) monitor continuously after deployment with periodic sampled re-evaluation.

Applications and Case Types

Asking three questions (what does the AI do, what does the human decide, and what record must remain) yields six types of PV scenario suited to LLMs and agents (Table 5). These are case types and do not refer to any specific company.

Table 5 | Six PV application scenarios and the division of labour.

Scenario	Work done by AI	Decision retained by humans	Records that must remain
Case intake and element extraction	Two-pass screening and extraction; fields bound to source text	Case validity; seriousness and expectedness	Source spans, extraction version, reviewer and time
MedDRA coding	Retrieve candidates, select and give reasons	Code confirmation, especially for serious and important medical events	Dictionary version, candidate list, final choice
Literature monitoring	First-pass screening of results; flagging potential cases	Whether an ICSR exists and must be reported	Search string, search date, time minimum criteria were first met
Digital platform monitoring	Post classification, deduplication, source labels	Whether to enter formal case processing	Source list, collection frequency, triage outcome
Signal assessment support	Case series assembly, timelines, supporting and opposing evidence	Signal validation, priority and follow-up action	Statistical computation records, evidence list, signed assessment
Aggregate report drafting	Constrained narratives generated from computed tables	Approval and submission of report content	Data snapshot, query version, generation and edit history

In the published literature the first two scenarios have the most substantial evidence^{16,18,20}; signal assessment support has an industry feasibility study¹⁹; generating safety summaries from computed results has a rerunnable workflow demonstration³⁴; and adverse event coding in clinical trials has a practice report from a Chinese company²⁰. In real-world monitoring of anti-amyloid- β therapy, Fong et al. compared expert review with an LLM-assisted workflow on 20,123 clinical documents from 46 patients: the LLM-assisted workflow identified every positive case (sensitivity 1.00, positive predictive value 0.897), while expert review had a sensitivity of 0.886 (positive predictive value 1.00)⁴¹. The comparison is telling: the model is less likely to miss, the human less likely to be wrong, and the sensible design combines them.

Challenges and Limitations

Hallucination and "plausible errors". Errors in PV often look plausible: a term that exists in medical language but not in the dictionary, or a dose close to but different from the source. Hakim et al. designed a set of guardrails for drug safety, including detecting anomalous documents, checking that drug names and adverse event terms are correct, and conveying uncertainty in generated content⁴². Guardrails should be deterministic checks, not the "opinion" of another model.

Bias and input contamination. Clinically irrelevant personal information can change model judgements²⁵. Inputs should be minimised to the task, and model call records should be reviewable.

Analytical misjudgement. Hammad et al. catalogue recurring analytical pitfalls in drug safety assessment: misinterpreting conditional probabilities, conflating association with causation, inappropriate denominators and comparators, neglect of background rates and confounding, Simpson's paradox and over-interpretation of exploratory findings. AI tools lacking transparency, bias assessment and clinical oversight may amplify rather than reduce these problems⁴³.

Disproportionality analysis has inherent limits, namely reporting bias and the lack of an exposure denominator, and the fluent explanations generated by LLMs may cause those limits to be overlooked.

Version drift and reproducibility. Model versions, prompts, retrieval corpora and MedDRA versions all change. The FDA draft guidance treats life-cycle maintenance as a dedicated topic¹³, and the FDA–EMA principles call for periodic re-evaluation to address data drift¹⁴. Without version locking, a past judgement cannot be replayed during inspection.

Insufficient evaluation evidence. Existing studies are mostly retrospective, with little external validation or prospective implementation research^{9,10,44}. Organisations should not treat published metrics as the expected performance in their own setting.

Data protection. Health information is sensitive personal information, and processing it requires a specific purpose and sufficient necessity. For most pharmaceutical companies, private deployment, de-identification and data minimisation are preconditions for using LLMs.

Table 6 | Main failure modes and corresponding controls.

Failure mode	Typical symptom	Control
Invented terms	Output of non-existent LLTs	Candidate-set constraint plus dictionary validation block
Field guessing	Fields populated without mention in the source	Fields must bind a source span or stay empty
Missed cases	Potential cases not identified	High-recall first pass plus stratified sampled review
Closed-book assertion	Risk conclusions without evidence	Evidence gating: no source, no answer
Bias	Irrelevant attributes affect judgement	Input minimisation; periodic fairness testing
Step skipping	Moving on before review is complete	Process order enforced by the system
Drift	Performance changes with versions	Version locking; changes trigger re-evaluation
Untraceability	Past judgements cannot be replayed	Tamper-evident audit logs and run records

Viewpoint and outlook

This section presents the opinions of the Great Unicorn Intelligence research team, derived from the public evidence cited above. They are not established facts.

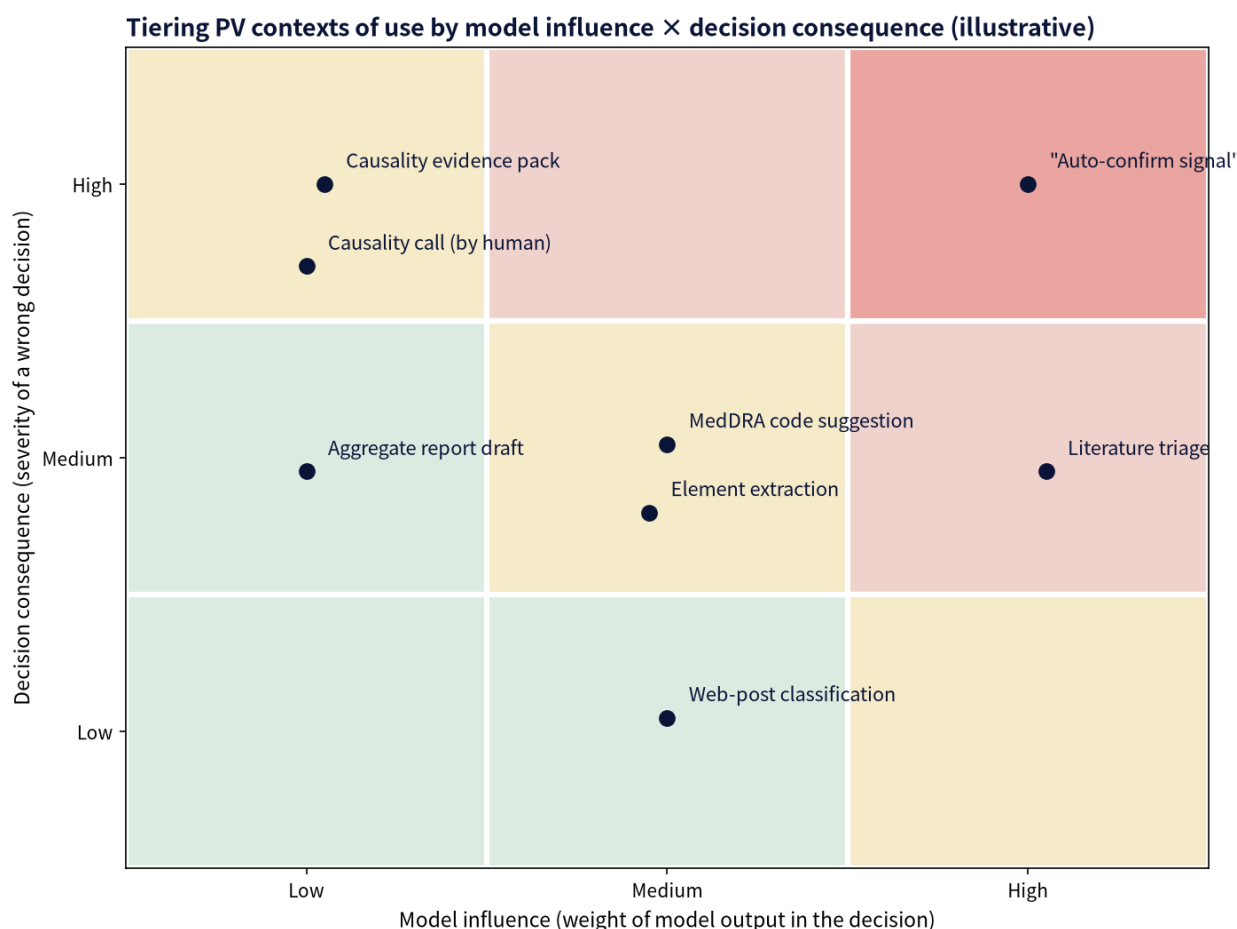
View 1: LLMs write; code, people decide

The evidence in the “Core Methods and Technical Analysis” and “Comparison and Evaluation” sections shows that LLMs perform well at extraction grounded in source text and at constrained writing, show limited agreement in causality assessment that requires integrated judgement, and are unreliable at mental arithmetic and assertion from memory. We therefore argue for a dividing line: reading (extraction, classification, retrieval) and writing (summaries, narratives, drafts) go to the model; computing (disproportionality statistics) goes to deterministic code; judging (case validity, code confirmation, signal validation, causality conclusions) goes to named people. This agrees with the

hybrid-architecture view of FDA researchers¹¹.

View 2: gates by influence × consequence

The model risk matrix of the FDA draft guidance¹³ can be applied directly to tiering PV use cases (Fig. 4). Literature screening, for example, has high model influence (the model decides which articles reach human readers), but every result is seen by a person, so decision consequence is medium. A draft causality assessment has high decision consequence, so the model may provide only sourced materials and the conclusion must be made by a person. We recommend that each context of use be tiered separately before deployment, and that the tier determine the review rate, the sampling scheme and whether human-on-the-loop oversight is permitted.



Framework from the FDA 2025 draft guidance; placements are the authors' opinion; red zone not recommended

Fig. 4 | Tiering PV contexts of use by model influence × decision consequence. Framework from the FDA draft guidance¹³; scenario placement is the authors' opinion and illustrative.

Our approach: governance in the workflow

According to the Great Unicorn Intelligence website, in its pharmacovigilance signal detection solution for pharmaceutical companies, disproportionality analyses such as PRR and ROR are computed by deterministic code; report text is understood and classified with the help of DH S1, the company's integrated inference-and-training LLM platform; and the deliverables are a candidate signal list and traceable computation records for PV staff to evaluate. Each release on the SAM platform generates a run manifest that binds the data snapshot, code version, parameter set, gate

sign-offs, model package and monitoring plan, so that the run can be replayed in full during inspection; the solution is deployed privately throughout⁴⁵. Agents such as the one that organises pharmacovigilance cases propose but do not decide, and run under the Harness governance engine of the enterprise agent platform. Entry control determines, before the model starts, whether a task may be handled by an agent, and re-checks authorisation on every tool call. Run control requires skills to execute in order without skipping steps, locks model versions, and uses a tool gateway and policy gates to block out-of-bounds calls. Exit control uses an independent verifier that re-checks results from structured data only, presents approvers with an approval package containing evidence both for and against, keeps a complete trail and supports emergency stop⁴⁶ (Fig. 5). The website also lists "drug safety determination" as a high-risk application that must retain sign-off by an authorised person, a human intervention channel and an emergency shutdown mechanism⁴⁶.

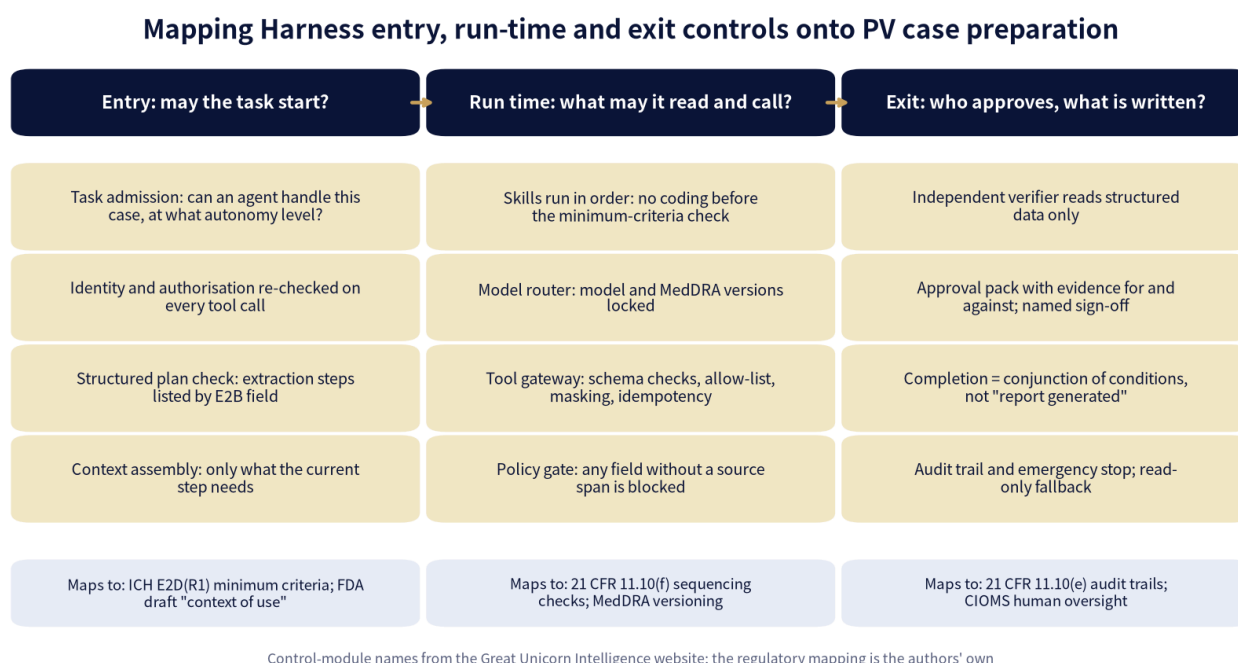


Fig. 5 | Mapping Harness entry, run and exit controls onto PV case processing. Control modules from the Great Unicorn Intelligence website⁴⁶; correspondence to regulatory requirements compiled by the authors.

Testable predictions for 2–3 years

Table 7 | Predictions, time frames and tests. Opinion.

No.	Prediction	Time frame	Test
P1	No fully automated case-level causality assessment will gain regulatory acceptance; model–expert agreement in the "possible" category will remain below inter-expert agreement	To end of 2028	Track category-level agreement in peer-reviewed studies and regulatory documents
P2	Evidence gating ("no source, no answer") becomes the default design of LLM tools for drug safety and appears in major guidance or industry standards	To end of 2028	Search subsequent CIOMS, ICH, FDA and EMA documents and major vendor documentation

No.	Prediction	Time frame	Test
P3	The share of PV LLM studies reporting human–AI team performance (review time, post-review error rate) rises clearly, exceeding half in journals such as <i>Drug Safety</i>	2027–2028	Systematic annual count of studies in relevant journals
P4	Once EU GVP Module VI is revised in line with E2D(R1) and the FDA AI guidance is finalised, inspections will set specific expectations for validation records, versions and audit trails of AI tools	To 2028	Track EMA/PRAC work plans and FDA guidance status ^{3,13}

Conclusion

What LLMs and agents in pharmacovigilance must really solve is not "can they replace people" but "how do people get to spend their time on judgement". The public evidence shows that on closed tasks such as extraction and coding, LLMs can already substantially reduce the burden of reading and data entry; in signal detection, statistics should be computed by deterministic code while the model groups, assembles and explains; and in causality assessment, the model is suited to preparing materials while conclusions must be made by named people. The CIOMS, ICH, FDA and EMA documents point to the same set of engineering requirements: a clear context of use, oversight tiered by risk, traceable results and replayable processes.

Our recommendations come down to four sentences. Constrain outputs to structures. Bind conclusions to source text. Leave statistics to code. Leave decisions to qualified people. When these four hold, LLMs can genuinely raise efficiency in pharmacovigilance rather than add risk.

References

1. European Medicines Agency. Guideline on good pharmacovigilance practices (GVP) Module IX – Signal management (Rev 1). EMA/827661/2011 Rev 1 (2017). https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-gvp-module-ix-signal-management-rev-1_en.pdf
2. 国家药品监督管理局. 国家药监局关于发布《药物警戒质量管理规范》的公告（2021年第65号）. 中国政府网 (2021). https://www.gov.cn/zhengce/zhengceku/2021-11/29/content_5654764.htm
3. International Council for Harmonisation. ICH E2D(R1): Post-Approval Safety Data: Definitions and Standards for Management and Reporting of Individual Case Safety Reports. Step 4, adopted 15 September 2025 (2025). https://database.ich.org/sites/default/files/ICH_E2D%28R1%29_Step4_FinalGuideline_2025_0819.pdf
4. International Council for Harmonisation. ICH E2B(R3): Electronic Transmission of Individual Case Safety Reports (ICSRs) – Data Elements and Message Specification – Implementation Guide. EMA scientific guideline page (2016). <https://www.ema.europa.eu/en/ich-e2b-r3-electronic-transmission-individual-case-safety-reports-icsrs-data-elements-message-specification-implementation-guide-scientific-guideline>
5. CIOMS Working Group XIV. Artificial Intelligence in Pharmacovigilance. Council for

- International Organizations of Medical Sciences, Geneva (2025). <https://doi.org/10.56759/cdob6397>
6. MedDRA MSSO. Introductory Guide, MedDRA Version 29.1 (September 2026). (2026). https://alt.meddra.org/files_acrobat/intguide_29_1_English.pdf
 7. International Conference on Harmonisation. ICH E2E: Pharmacovigilance Planning. Step 4 version, 18 November 2004 (2004). https://database.ich.org/sites/default/files/E2E_Guideline.pdf
 8. Salvo, F., Crupi, L. & Cholle, C. Language models in pharmacovigilance: Applications, promises and limits. *Thérapie* 81, 140–146 (2026). <https://doi.org/10.1016/j.therap.2025.12.003>
 9. Kim, T. Y., Oh, W.-S. & Jeong, D.-H. Large Language Models in Adverse Drug Reaction Detection and Pharmacovigilance: A Systematic Review of Current Applications, Challenges, and Future Directions. *Diagnostics* 16, 15 (2026) (2026). <https://doi.org/10.3390/diagnostics16152435>
 10. El Sayed, I., Bedwany, R. N., Salama, M. & Eltrass, A. S. H. Integration of artificial intelligence into pharmacovigilance: A systematic review. *Research in Social and Administrative Pharmacy* (2026). <https://doi.org/10.1016/j.sapharm.2026.09.001>
 11. Wu, L., Xu, J., Dang, O. & Ball, R. Does generative AI mean the "end of history" for pharmacovigilance automation? *towards a framework for the future of human-AI systems. Frontiers in Drug Safety and Regulation* 6, 1846339 (2026). <https://doi.org/10.3389/fdsfr.2026.1846339>
 12. European Medicines Agency. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023, adopted 9 September 2024 (2024). https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf
 13. U.S. Food and Drug Administration. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products. Draft Guidance for Industry and Other Interested Parties, January 2025 (2025). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological>
 14. U.S. Food and Drug Administration & European Medicines Agency. Guiding Principles of Good AI Practice in Drug Development. January 2026 (2026). <https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development>
 15. U.S. Code of Federal Regulations. 21 CFR 11.10 – Controls for closed systems. Electronic Code of Federal Regulations. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11/subpart-B/section-11.10>
 16. Ludwig, D., Wang, M., Buchanan, J. & Trinh, T. A Large Language Model for Extracting Post-marketing Adverse Drug Events from Clinical Notes in the Electronic Health Record. *Drug Safety* 49, 1181–1198 (2026). <https://doi.org/10.1007/s40264-026-01682-6>
 17. Plasek, J. M. et al. Relationship extraction for adverse drug events in clinical notes using large language models: a retrospective cross-sectional study. *eBioMedicine* 133, 106516 (2026). <https://>

18. Ma, J. et al. Intelligent Framework for Adverse Drug Event Identification Using Large Language Models and Retrieval-Augmented Generation: Development and Evaluation Study. *Journal of Medical Internet Research* 28, e84086 (2026). <https://doi.org/10.2196/84086>
19. Warner, J., De Souza Teodoro, L. H., Jardim, A. P. & Albera, C. Intelligent Automation Improved Efficiency in Pharmacovigilance Safety Signal Assessment. *Clinical Pharmacology & Therapeutics* (2026). <https://doi.org/10.1002/cpt.70409>
20. Yan, C. et al. Application of Artificial Intelligence in MedDRA Coding: A Practical Exploration from Clinical Data Management Perspective. *Therapeutic Innovation & Regulatory Science* 60, 1059–1067 (2026). <https://doi.org/10.1007/s43441-026-00953-9>
21. Vandenhende, F. et al. Performance of the SafeTerm AI-Based MedDRA Query System Against Standardised MedDRA Queries. Preprint at <https://arxiv.org/abs/2512.07552> (2025).
22. Vandenhende, F. et al. Automated Generation of Custom MedDRA Queries Using SafeTerm Medical Map. Preprint at <https://arxiv.org/abs/2512.07694> (2025).
23. Zhou, Y. et al. Quality-aware multi-source data fusion and enhancement for Medical Concept Normalization using Large Language Models. *Journal of Biomedical Informatics* 182, 105093 (2026). <https://doi.org/10.1016/j.jbi.2026.105093>
24. Xue, Z. et al. AskGLP1: A Knowledge Graph LLM Agent for Exploring GLP-1 Experience from Online Patient Narratives. Preprint at <https://arxiv.org/abs/2610.34987> (2026).
25. Liu, S., Zhang, S. & Bala, I. Robust or Suggestible? Exploring Non-Clinical Induction in LLM Drug-Safety Decisions. arXiv preprint arXiv:2510.13931 (2025). <https://arxiv.org/abs/2510.13931>
26. Evans, S. J., Waller, P. C. & Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety* 10, 483–6 (2001). <https://doi.org/10.1002/pds.677>
27. van Puijenbroek, E. P. et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety* 11, 3–10 (2002). <https://doi.org/10.1002/pds.668>
28. Bate, A. et al. A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology* 54, 315–21 (1998). <https://doi.org/10.1007/s002280050466>
29. DuMouchel, W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. *The American Statistician* 53, 177–190 (1999). <https://doi.org/10.1080/00031305.1999.10474456>
30. Norén, G. N., Hopstadius, J. & Bate, A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. *Statistical Methods in Medical Research* 22, 57–69 (2013). <https://doi.org/10.1177/0962280211403604>
31. Bright, M., Kontsioti, E., Pirmohamed, M., Zhou, Y. & Maskell, S. MCSG: A Method for

- Simultaneous Disproportionality Analysis and Background Rate Estimation in Large Pharmacovigilance Databases. *Drug Safety* 49, 481–491 (2026). <https://doi.org/10.1007/s40264-025-01632-8>
32. Balogh, O. M. et al. Contrastive learning of adverse events to provide effective and interpretable vector representations for machine-assisted pharmacovigilance. *Briefings in Bioinformatics* 27, 5 (2026) (2026). <https://doi.org/10.1093/bib/bbag463>
 33. Kefala, M., Painter, J. L., Bukhari, S. T. & Sessa, M. Development of a European Union Time-Indexed Reference Dataset for Assessing the Performance of Signal Detection Methods in Pharmacovigilance using a Large Language Model. Preprint at <https://arxiv.org/abs/2603.26544> (2026).
 34. Calle, X., Mendez, N. & Garin-Muga, A. Pharmacovigilance Assistant: An Agentic Workflow for Reproducible Drug Safety Summaries. *Studies in Health Technology and Informatics* 336, 42–46 (2026). <https://doi.org/10.3233/SHTI260105>
 35. Heckmann, N. S. et al. Biomedical Large Language Models and Prompt Engineering for Causality Assessment of Individual Case Safety Reports in Pharmacovigilance. *Pharmaceutical Research* 43, 2761–2774 (2026). <https://doi.org/10.1007/s11095-026-04121-4>
 36. Heckmann, N. S. et al. Optimizing Large Language Models for Causality Assessment in Pharmacovigilance: Developing a Performance Metric as Objective for Bayesian Hyperparameter Optimization. Preprint at <https://arxiv.org/abs/2607.03704> (2026).
 37. Ha, Y. M. et al. Large Language Models for World Health Organization-Uppsala Monitoring Centre Drug-Adverse Event Causality Assessment Using Food and Drug Administration Adverse Event Reporting System Cases: Comparative Performance Study. *Journal of Medical Internet Research* 28, e93237 (2026). <https://doi.org/10.2196/93237>
 38. Lopez, D. R., Boccia, A. & Sessa, M. Multi-Agent Retrieval-Augmented Causality Assessment (MARCA) of Individual Case Safety Reports in Pharmacovigilance. Preprint at <https://arxiv.org/abs/2610.33279> (2026).
 39. Shi, J., Ma, Y., Liu, Y. & Shi, B. Confident but unsupported: auditing large language models against supplied evidence boundaries in drug-induced liver injury assessment. *Frontiers in Public Health* 14, 1904062 (2026). <https://doi.org/10.3389/fpubh.2026.1904062>
 40. Norén, G. N., Meldau, E.-L. & Ellenius, J. Critical Appraisal of Artificial Intelligence for Rare-Event Recognition: Principles and Pharmacovigilance Case Studies. *Drug Safety* 49, 747–760 (2026). <https://doi.org/10.1007/s40264-026-01649-7>
 41. Fong, A. et al. Real-world pharmacovigilance for anti-A β therapies using a large language model. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 18, 3 (2026), e70412 (2026). <https://doi.org/10.1002/dad2.70412>
 42. Hakim, J. B. et al. The Need for Guardrails with Large Language Models in Medical Safety-Critical Settings: An Artificial Intelligence Application in the Pharmacovigilance Ecosystem. Preprint at <https://arxiv.org/abs/2407.18322> (2024).

43. Hammad, T. A. & Rochon, J. Analytic Misjudgment of Drug Safety Evidence and Causality: From the Prosecutor's Fallacy and Simpson's Paradox to Artificial Intelligence. *Drug Safety* 49, 1119–1133 (2026). <https://doi.org/10.1007/s40264-026-01683-5>
44. Schreier, O. et al. Application of Language Models for the Analysis of Adverse Drug Events in Pharmaceutical Research and Development: Scoping Review. *JMIR AI* 5, e77732 (2026). <https://doi.org/10.2196/77732>
45. 弘犀智能科技（深圳）有限公司 / Great Unicorn Intelligence. 行业解决方案·制药 (Pharmaceutical Solutions). Official website, accessed 2026-10-29 (2026). <https://www.hxallai.com/solutions/pharma/>
46. 弘犀智能科技（深圳）有限公司 / Great Unicorn Intelligence. 企业智能体平台·Harness 治理引擎 (Enterprise Agent Platform and Harness Governance Engine). Official website, accessed 2026-10-29 (2026). <https://www.hxallai.com/platform/agent/>

Acknowledgements We thank the Great Unicorn Intelligence Research Team, Tsinghua University, and Professor Qiang Wang's research team at Xi'an Jiaotong University.

Competing interests Qiang Wang and the Great Unicorn Intelligence Research Team are affiliated with Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd., whose products and views are discussed in this article.

Additional information Correspondence and requests for materials should be addressed to the Great Unicorn Intelligence Research Team (research@hxallai.com).

Copyright Notice: © 2026 Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. All rights reserved. Any reposting, quotation in commentary, or adaptation must credit the authors (Qiang Wang and Great Unicorn Intelligence Research Team) and cite the source (Great Unicorn Intelligence, <https://www.hxallai.com>).