scientific reports

OPEN

The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings

Joe B. Hakim[1], Jeffery L. Painter[2], Darmendra Ramcharran[3], Vijay Kara[4], Greg Powell[2], Paulina Sobczak[5], Chiho Sato[6], Andrew Bate[4,7,9]✉ & Andrew Beam[8,9]

Large language models (LLMs) are useful tools with the capacity for performing specific types of knowledge work at an effective scale. However, LLM deployments in high-risk and safety-critical domains pose unique challenges, notably the issue of "hallucinations", where LLMs can generate fabricated information. This is particularly concerning in settings such as drug safety, where inaccuracies could lead to patient harm. To mitigate these risks, we have developed and demonstrated a proof of concept suite of *guardrails* specifically designed to mitigate certain types of hallucinations and errors for drug safety, with potential applicability to other medical safety-critical contexts. These guardrails include mechanisms to detect anomalous documents to prevent the ingestion of inappropriate data, identify incorrect drug names or adverse event terms, and convey uncertainty in generated content. We integrated these guardrails with an LLM fine-tuned for a text-to-text task, which involves converting both structured and unstructured data within adverse event reports into natural language. This method was applied to translate individual case safety reports, demonstrating effective application in a pharmacovigilance processing task. Our guardrail framework offers a set of tools with broad applicability across various domains, ensuring LLMs can be safely used in high-risk situations by eliminating the occurrence of key errors, including the generation of incorrect pharmacovigilance-related terms, thus adhering to stringent regulatory and quality standards in medical safety-critical environments.

The integration of large language models (LLMs) into the fabric of numerous applications has positioned them as instrumental in navigating the complex challenges in biology and medicine[1]. The breadth of their application, combined with their rapid evolution, has created anticipation that LLMs will be near-universal solvers across the biomedical landscape[1–3]. Yet, alongside this growing optimism, there is an increasing cognizance of their limitations that may impede their applicability in specific areas of scientific inquiry. Prominently, the phenomenon of "hallucinations"—instances of generating baseless information—stands as a pivotal concern[4]. This phenomenon is a byproduct of the mechanisms underpinning LLMs, which rely implicitly on internally stored "memories" for response generation, without explicit grounding in verifiable facts[5]. LLMs also face challenges in communicating the uncertainties of their outputs to end-users effectively. Though measures of uncertainty can sometimes be quantified, validating the trustworthiness of LLM outputs remains a challenge[6,7] including within the biomedical domain[8].

In contexts where inaccuracies can result in severe consequences, particularly in decision-making processes affecting patient safety, the issue of LLM hallucinations and omission of key information[9] becomes acutely significant[10]. One critical domain is drug safety, also known as pharmacovigilance (PV), which involves the ongoing surveillance for adverse events (AEs) linked to pharmaceutical medicines and vaccines[11]. Given the limitations of pre-market trials in fully characterizing a drug or vaccine's safety profile, PV relies on the collection and analysis of spontaneously reported AEs, vital for continued assessment of a product's benefit-risk. The reported information is transcribed into an Individual Case Safety Report (ICSR) which serves as

[1]Harvard-MIT Department of Health Sciences and Technology, Cambridge, MA, USA. [2]GSK, Durham, NC, USA. [3]GSK, Providence, RI, USA. [4]GSK, London, UK. [5]GSK, Warsaw, Poland. [6]GSK, Tokyo, Japan. [7]London School of Hygiene and Tropical Medicine, London, UK. [8]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [9]Andrew Bate and Andrew Beam have contributed equally to this work. ✉email: andrew_beam@hms.harvard.edu

the standardized international framework for AE reporting, encompassing a vast array of information sourced globally in varied formats and demanding timely review and processing. The non-random process by which ICSRs are collected, coupled with the prevalence of incomplete or erroneous data, underscores the necessity for clinical review to unearth potential safety signals for further exploration and serve as the primary data source to formally evaluate a potential causal association[12]. Consequently, a challenge within PV lies in the efficient parsing of extensive, noisy, and often incomplete domain-specific textual data, some of which may be contradictory, to identify safety signals meriting additional investigation.

The propensity of LLM to hallucinate and omit key details presents a considerable hazard if applied naively within the PV domain, which is inherently safety-critical. For instance, an LLM might erroneously suggest that an ICSR details a serious AE such as liver failure while this is not mentioned in the source report, potentially signaling a false-positive safety concern and diverting resources from legitimate safety investigations. Moreover, understanding how LLMs are integrated with human end-users becomes essential, as human-mediated oversight systems will likely remain indispensable for certain tasks within safety-critical applications for the foreseeable future.

Preventing and mitigating hallucinations involves the implementation of "guardrails" around LLMs to shape and restrict their output. While the term *guardrail* lacks a precise definition in this context, it is here understood as a series of constraints applied to either an aspect of the LLM or its output to ensure adherence to predefined criteria. One approach involves "structural guardrails," defined as mechanisms ensuring model outputs maintain a consistent structure (e.g., CSV, XML, JSON)[13], thus obviating the need for further processing of free text to extract pertinent information.

This paper focuses on "semantic guardrails," aimed at verifying the accuracy of LLM output by checking for biased or problematic content and coding errors. These guardrails may be "hard," offering clear binary outcomes, or "soft," providing probabilistic assessments regarding the potential error in the output. Within pharmacovigilance, such guardrails are pivotal in enforcing the avoidance of errors that may impact safety decisions resulting in patient harm analogous to medical "never events" incidents in clinical practice contexts identified by U.S. and U.K medical organizations as wholly preventable and unacceptable[14].These never events, deemed intolerable and preventable, have the potential to lead to significant harm or mortality and usually trigger comprehensive investigations to avert recurrence. Examples include severe allergic reactions to contraindicated medications or dosing errors and are "serious incidents that, due to the provision of systemic protective barriers at a national level, are completely preventable and should have been preemptively addressed by all healthcare providers"[15], analogous to guardrails. Hence, to function within safety-critical domains like PV, semantic guardrails must ensure the absolute prevention of defined "never event" errors that have the potential to adversely impact pharmacovigilance decision-making[14,16].

In our investigation, we introduce a comprehensive set of both hard and soft semantic guardrails designed to enable LLMs to function within the high-risk, safety-critical environment of PV. Focusing on the complex and expansive data processes integral to PV, our research specifically addresses the challenge of processing multilingual ICSR intake and analogous processing within a real-world PV system. This encompasses a text-to-text task that involves both structured to unstructured data conversion and translation. Our guardrails were specifically tested on the task of transforming Japanese language ICSRs (combined unstructured and structured, tabular data with numeric codes for various biomedical concepts) into English narrative text for subsequent analysis by safety professionals.

We identified multiple potential failure modes for LLMs within this context and engineered a series of guardrails to mitigate these risks (Fig. 1). We implemented a hard semantic guardrail to address model outputs with generated drug or vaccine names not present in the source text, utilizing existing drug safety dictionaries and tools to ensure consistency of key drug- and vaccine-related information between the source text and the LLM-generated English narrative. Additionally, we incorporated two soft semantic guardrails to communicate the model's uncertainty regarding the quality and accuracy of both the input text and its final translation, thereby flagging instances potentially requiring further human review. While our study concentrates on a critical, real-world case in PV, we posit that the framework developed herein holds relevance across a multitude of medical safety-critical domains.

Methods

A schematic of the workflow is presented in Fig. 1, including processing the ICSRs, the LLM tasks, creation of standards and the evaluation of LLM generated case reports, the sequential guardrail processing, and the evaluations of the guardrails.

Data acquisition

The dataset utilized in this study was sourced from GSK's global safety database as part of a collaboration by providing Harvard University, Cambridge, Massachusetts, USA, access on a privately maintained, secure server equipped with advanced graphics processing units (two 80 GB A100s). This dataset encompasses over 2 decades of ICSRs, with more than 4 million cases available for review. For the purposes of our assessment, the analysis concentrated on the original ICSRs as submitted to GSK, prior to any form of human review. This excluded any subsequent modifications or additional data reported post-initial submission, including follow-up details.

Analysis of individual case safety reports

Overall, we use ICSRs combined with important data fields in PV such as the level of seriousness of the adverse event to produce single chunks of text that are used as the "source" text to be fed into the LLM. Specifically, spontaneously reported AEs are transcribed into an ICSR which serves as the standardized international framework for AE reporting. A valid ICSR for entry into the GSK Global safety database is comprised of four

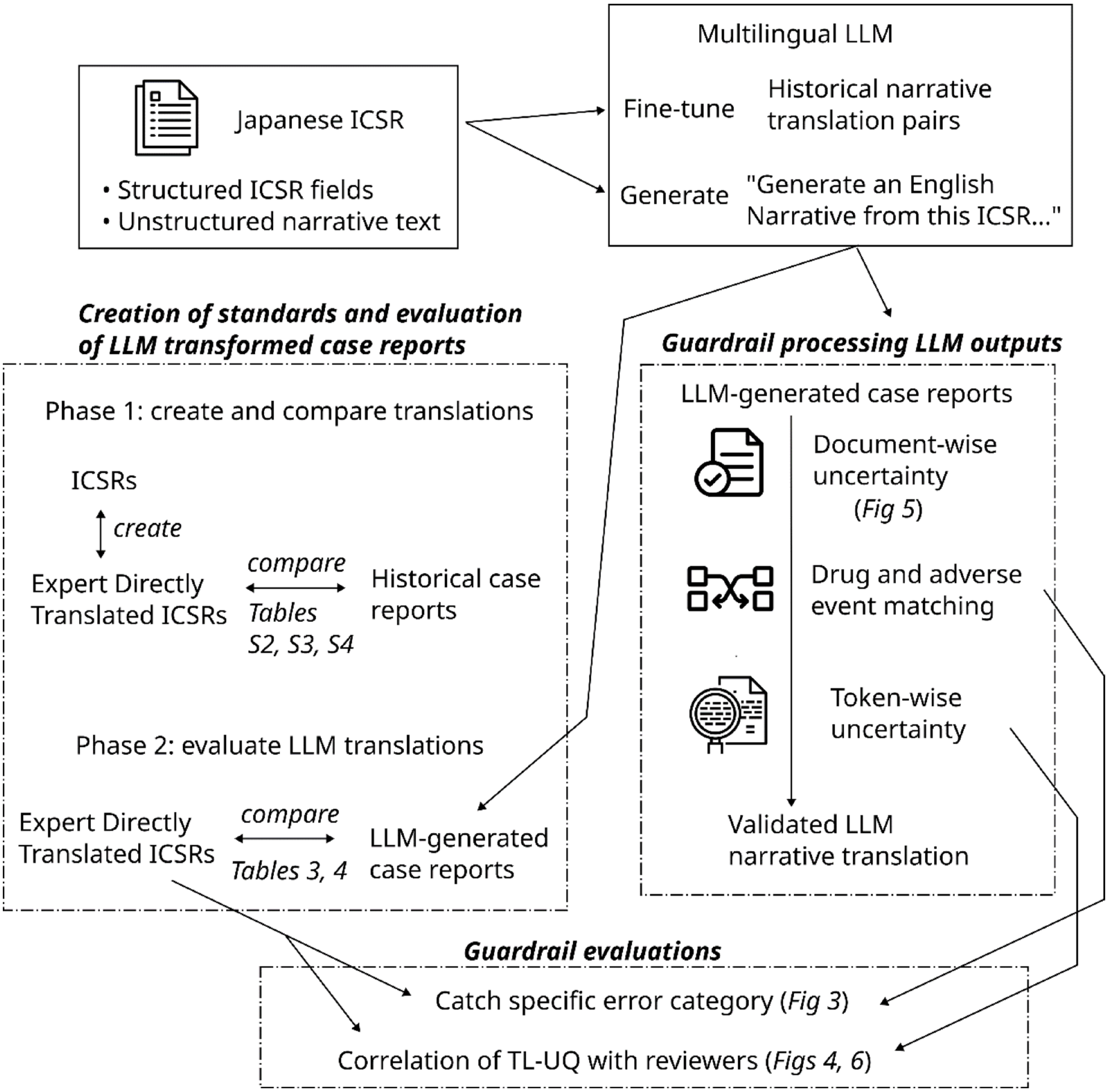

Japanese ICSR
• Structured ICSR fields
• Unstructured narrative text
Multilingual LLM
Fine-tune
Historical narrative translation pairs
Generate
"Generate an English Narrative from this ICSR..."
Creation of standards and evaluation of LLM transformed case reports
Phase 1: create and compare translations
ICSRs
create
Expert Directly Translated ICSRs
compare
Tables S2, S3, S4
Historical case reports
Phase 2: evaluate LLM translations
Expert Directly Translated ICSRs
compare
Tables 3, 4
LLM-generated case reports
Guardrail processing LLM outputs
LLM-generated case reports
Document-wise uncertainty (Fig 5)
Drug and adverse event matching
Token-wise uncertainty
Validated LLM narrative translation
Guardrail evaluations
Catch specific error category (Fig 3)
Correlation of TL-UQ with reviewers (Figs 4, 6)

Fig. 1. Graphical summary of the large language model (LLM) workflow. We used extra structured fields and unstructured narrative texts from individual case safety reports (ICSRs), along with historical matched language examples, to fine-tune an LLM. We added a specific task prefix, and generated an English narrative from a Japanese ICSR, and finally checked this process via several guardrails: the document-level uncertainty, drug and adverse event matching, and token-level uncertainty guardrails (see Methods section).

essential elements: (1) at least one identifiable reporter; (2) an identifiable patient; (3) at least one suspect adverse reaction; and (4) at least one GSK suspected product[17]. Pharmaceutical entities often accumulate reports in large volumes from various data partners. Whenever feasible, these reports are exchanged using standardized E2B XML documents[18], which offer structured fields alongside narrative descriptions of each case. For our study, we treated the entirety of information initially available as a singular data point. This approach included aggregating additional structured fields such as the country of the primary source, country of occurrence, level of seriousness (including death, life-threatening situations, hospitalization, disability, congenital anomalies), relevant dates, details of the reporter (name, organization, country), patient demographics (age, sex), primary reaction of the patient, and the implicated medicinal product.

Development of a multilingual corpus for LLM pretraining

We constructed a multilingual corpus of ICSRs to serve as the dataset for text-to-text fine-tuning of an LLM. To achieve this, we aligned the raw text from the submitted ICSRs (source text) with the human-generated summaries provided by a third-party contractor (original standard target text) to create text pairs in four languages: Japanese, Spanish, French, and German. These languages were selected due to their prevalence in the database and, particularly for Japanese, the complexity they present in translation tasks. While our analysis primarily concentrates on Japanese due to the high number of ICSRs available in this language, the LLMs are designed for multilingual application. To integrate the additional structured fields, we prefixed each one to the source text of the ICSR in the format:

field_name_1: field_value_1; field_name_2: field_value_2.

To fine-tune an LLM or employing it in text generation, we also prefixed a brief instruction indicating the specific task for the model, such as "Translate the following Japanese case report into English narrative text" for translations from Japanese to English. Our pretraining corpus was enriched further with direct translation pairs from the OPUS-100 corpus[19], a comprehensive multilingual translation dataset covering 100 languages, thus furnishing additional examples for model fine-tuning on translation tasks involving parallel language sets. The volume of pretraining examples is detailed in Table 1.

Development of the ICSR translation LLM

Model fine-tuning and generation

A graphical flow chart describing this process is available as Supplemental Fig. S1.

In our study, we conducted an evaluation of three LLMs with parameter sizes ranging from 700 million to 7 billion: mt5-xl, mpt-7b-instruct, and stablelm-japanese. The criteria for selecting these models included the relevance of their initial pretraining objectives, the scale of the models, and the computational resources required for their operation. These models underwent further fine-tuning for translation tasks, utilizing a corpus composed of 131,037 examples from ICSRs and texts from the OPUS-100 dataset (Table 1). The training process was applied uniformly across Japanese, Spanish, French, and German, adopting a split of 70% for training, 15% for validation, and 15% for testing. This distribution ensured a balanced representation of languages and sources (ICSR vs. OPUS-100) within each set. Fine tuning was done by iterating over the training corpuses input-output pairs (either the structured-unstructured PV pairs or the direct translation pairs). For each, the model used either the joined structured-unstructured ICSR or the source language as input, and its output was compared to the target language report or target language direct translation.

Generation was done for each of the ICSRs, including those in validation and testing. This involved running the beam search algorithm for each input, and simply storing those for downstream computing of metrics or human evaluation tasks.

For the generation phase, we evaluated a variety of hyperparameters. Utilizing beam search[20], we experimented with different settings for the temperature (0.5, 0.7, 1.0, 2.0) and beam counts (3, 10, 25). Additionally, in our application of contrastive search[21], adjustments were made to the α values (0.2, 0.6, 0.9) and the top-k selections (4, 8, 16, 64). The optimal set of generation hyperparameters was determined based on the BLEU score[22] performance on the validation set, ultimately selecting a contrastive search configuration with $\alpha = 0.2$ and top-k = 16.

Model evaluation

In our initial assessments, we concentrated on evaluating the Japanese translation quality, a task of significant relevance in PV due to the human resources required with securing proficient translators for Japanese drug safety data. We conducted comparative analyses of the three models, utilizing per-token perplexity as a metric on a validation subset comprising 7820 ICSRs, which constitute approximately 13% of the total Japanese ICSRs in our dataset. For the best performing model, we further explored its translation capabilities by applying standard machine translation evaluation metrics, including the BLEU score[22], SACRE-BLEU score[23], and word error rate[24].

Expert human evaluation of the target text

After finalizing our model, we performed a comprehensive evaluation aimed at assessing its efficacy in translating cases that were originally documented in Japanese. This analysis involved 210 cases, all sourced from Japan and initially documented in Japanese. The selection of these cases was governed by a predefined set of criteria. Our goal was to achieve an even distribution across various product categories, with our sample evenly divided among vaccines, general medicines, and specialized products, like those in oncology. Priority was given to serious cases that had been subjected to in-depth analysis upon their reception, thus offering a comprehensive insight into potentially critical incidents. Additionally, we sought to maintain a balanced representation of products across these categories. The cases spanned the entire 20-year period for which we had data, ensuring

	Number of ICSRs
Pretraining examples (GSK), total	131,037
ICSRs in Japanese	58,855
ICSRs in Spanish	13,264
ICSRs in German	30,370
ICSRs in French	28,548
Japanese direct translation pairs (OPUS-100)	10,000
Spanish direct translation pairs (OPUS-100)	10,000
German direct translation pairs (OPUS-100)	10,000
French direct translation pairs (OPUS-100)	10,000

Table 1. Numbers of individual case safety reports (ICSRs) and direct translation pairs.

temporal representativeness. Finally, our case selection employed random sampling within these specific strata to reflect the overall distribution of Clinical Utility Score for Prioritization (CUSP) scores[25] found in our entire ICSR database. This methodology was designed to secure a broad and diverse representation in the completeness of the cases under review.

Phase 1: establishment of high-quality baseline translations

The first phase was dedicated to creating a baseline foundation of high-quality translations. Each of the 210 Japanese ICSRs, available in the database as previous translations into English by an external contractor, was subjected to a thorough review by two independent PV experts fluent in both Japanese and English. This double-blind review not only verified the translations for accuracy in comparison to source text and fluency, but also established a robust English "ground truth" for further comparative analysis. No adjudication review was required for this comparative assessment as, as the review comments of both reviewers were made available for phase 2 reviewers, with the ability to seek clarity where required to support phase 2 review.

The outcomes from this phase's evaluation are detailed in the supplementary materials, with Tables S2, S3, and S4 offering a juxtaposition of the initial standard target texts against the evaluations conducted by the bilingual PV specialists.

Phase 2: evaluation of LLM translations against established baseline

In the next phase, we assessed the LLM-generated English translations against the "ground truth" translations derived in Phase 1 of the experiment. This assessment was carried out by PV experts proficient in English, with experience in safety evaluations. Employing a carefully designed evaluation framework, they conducted independent dual reviews of each translation, incorporating both a detailed five-category assessment system (Table S1 for category specifics) and binary evaluation criteria (Table 4). In instances of binary evaluation, the presence of any noted error category, observed even once, warranted its marking, with evaluators having the option to detail the specific nature of the error. Moreover, the experts assessed the clinical acceptability of each processed ICSR for reporting to regulatory agencies. Any discordance among the evaluations was resolved by an additional independent senior expert, this need is underpinned by the low inter-rater agreement between clinical experts when evaluating the same drug-event ICSR cases has been well documented in the medical literature[26–29]. For the four-category criteria, evaluations were made on a five-point Likert scale[30], with ratings ranging from 1 (least favorable) to 5 (most favorable), as detailed in Table S1 in the supplement for the definitions of each rating level.

To streamline the evaluation, a custom web application was created, affording the reviewers the ability to methodically compare translations side-by-side and to log their assessments using dropdown menus and open-ended text fields. A screen capture of this web application's graphical user interface is available in Supplemental Fig. S2. Cases were randomly distributed among a team of reviewers to minimize the potential for individual reviewer and selection bias. This application was designed with tracking capabilities for capturing individual evaluator responses, and it was programmed to automatically signal for independent expert adjudication should discrepancies between reviewers emerge.

LLM guardrails for ICSR translations

We developed one hard and two soft semantic guardrails for this application, as described below in order of application in the ICSR processing pipeline:

Document-level uncertainty quantification (DL-UQ)

This soft guardrail identifies submitted documents that are unlikely to be ICSRs reports (based on statistical probabilities as reported by a model, as opposed to using the 4 aforementioned validation criteria for ICSRs). To support potential automation of ICSR intake, this guardrail detects documents unlikely to be an AE report and prevents any LLM processing of these reports. The DL-UQ guardrail first creates a document level embedding by performing an average pooling operator to the token-level embeddings created using the source language encoder LLM. Next, a k-nearest neighbors' Euclidean distance is calculated between the embedding for the submitted document and a cache of ICSR embeddings created using the same methodology from the training data. This distance is a measure of uncertainty according to the LLM as it measures how anomalous a new submission is relative to the documents the model has seen before and can be used to automatically discard a submission or flag it for review. A distance threshold can be tuned to achieve a desired trade-off between sensitivity and specificity.

MISMATCH (drug and AE mismatching)

This hard guardrail enforces a "never" event by identifying drug names that appear in either the source text or target text but not both, indicating that a drug name has been either mistranslated or hallucinated. This kind of error represents a so-called "never event" because incorrectly identifying a drug in an ICSR could have dire safety consequences and should be avoidable. To implement this guardrail, we matched (with regular expressions) both the source and target texts for any mentions of drugs; similarly, this was implemented for AEs. Then, we used two dictionaries (a custom in-house drug dictionary from the global safety database, and MedDRA, Medical Dictionary for Regulatory Activities[31]; 28 K preferred terms) to find the matching terms, and whether the set difference had any elements corresponding to unmatched terms. The dictionary matches allowed generic-trade name associations for drugs. Note: this guardrail did not match terms that are slightly misspelled drug names or AEs, since those are not matched by the regular expression-based text matching comparison with terms in the dictionaries. If there was a mismatch, this hard guardrail would trip and the eventual integrated system would route outside of the standard case processing and for further adjudication, either through post-processing or human-in-the-loop assessment and correction.

Token-level uncertainty quantification (TL-UQ)
This soft guardrail identifies potential LLM errors at word and sub-word levels. Each token in the vocabulary is assigned a log probability by the LLM, and we take the entropy of this multinomial distribution as the token-level uncertainty score. Intuitively, the more entropy in the predictive distribution of the next token, the "less certain" the model is in generating that specific token.

Guardrail assessments

We assess each of the guardrail types in different ways. For the DL-UQ guardrail, we compare the score of real inputs to fake inputs. For the MISMATCH guardrail by counting the types of instances where the model's outputs different from humans'. For the TL-UQ guardrail, we produce visualizations that flag each tokens' span with its correlated score, and also comparing the level of overall TL-UQ flagging to the human rated error rates in certain categories.

We assessed each guardrail as follows:

For DL-UQ, using the train validation split described above (see "Data pre-processing" and "Model evaluation"), we sampled 80 example texts from the training and validation sets, and produced a score for each. We then injected a sample of 25 "extraneous samples", which included 14 Japanese Wikipedia articles, 7 Japanese fake case reports (in a similar format as the original case reports), 2 Japanese texts that have nothing to do with PV, and 2 non-Japanese texts. We plotted the numeric score for each example to evaluate the separation and reported the area under the receiver operator curve (AUROC) for a discrimination between validation and extraneous samples.

For MISMATCH, the primary evaluation of this guardrail was whether the specific targeted "never event" is always flagged when the target text contains that error. To this end, we used the human evaluators' flagged drug errors as the exemplar never events on a (programmatically) randomly selected sample of 20 cases. We calculated the fraction of cases caught by the MISMATCH guardrail where the human evaluators indicated a drug name had been hallucinated spontaneously. The MISMATCH guardrail was also useful for the other categories. We divided its fixes in the "generic-trade name" category by how the specific drugs mentioned were flagged by the MISMATCH guardrail itself ("fixed by mismatch guardrails") or by a separate system that we added to check if generic-trade names match by looking for parentheses ("direct generic-trade name checking"). For this, we used the existing pairs of generic-trade names from the dictionary afterwards. We divided fixes in the "drug spelling issues" section by whether they were directly fixed by the guardrails ("fixed by mismatch guardrails") or not fixed, which happened when the same drug had both correct and incorrect spellings in the LLM generated output ("multiple mentions"). In these cases, the guardrail did not find the misspelled drug by matching the text in the LLM output in English, and it did not detect that the drug mention in the Japanese source text is unmatched, because the drug is also spelled correctly. The frequency of the MISMATCH guardrail flagging individual drug and AE names is evaluated by a "missrate", which is a measure of the frequency of these erroneous outputs that are not fixed by this guardrail. A missrate of 1.0 indicates that, for example, the source text contained a number of drugs or AEs that are not matched to any translated terms in the target text. In the standard translations used to train the model, we expect this to be 0.0, so any missrates > 0.0 are due to a limitation of the drug or AE lists, or a misspelling of these terms.

For evaluation of the TL-UQ guardrail, we showed a qualitative example of a visualization flagging spans of uncertain text. In that example, we correlated the flagged spans with a human evaluator's assessment of specific errors in that case. Spans were flagged by differing intensities of text highlighting, from least to most, corresponding to the 10th percentile, 5th percentile, and 1st percentile most entropic predicted tokens. Quantitative evaluations were conducted by stratifying each reviewed case by "Is the case clinically accurate", "Wrong name or information", and "Incorrect AE/Wrong outcome" and assessing the case entropy score (an average of the individual token entropies) for each case in each category.

Results

Translation model development and evaluation

We considered three LLMs, that at the time of beginning this study, were performant multilingual models that could run given local resources on our internal servers: mt5-xl, MPT-7B, and stablm-japanese. We first assessed how well each could translate ICSRs from Japanese to English without any task-specific fine-tuning and then assessed this ability when the models were fine-tuned with ICSR data (Table 2).

These results indicate that none of the base models are suitable for translation "off the shelf" (Table 2). Fine-tuning improved all models by a significant margin and only mt5-xl reached a suitable perplexity after fine-tuning (Table 2). This is most likely due to this base model being pretrained explicitly on Japanese text, while the others likely only encountered Japanese text during their initial pretraining in an extremely small number of instances. On this basis, we decided to move forward with the mt5-xl model for further evaluation.

Traditional metrics of machine translation quality for mt5-xl show a BLEU score of 0.39, which is considered to be associated with relatively high-quality translations[32], as are the Sacre-BLEU score of 0.44 and the word error rate of 0.73.

Preliminary feasibility study

In a pilot assessment of 20 translations, we found that 16/20 (80%) were deemed acceptable overall. Supplemental Table S6 shows a breakdown of the kinds of errors identified by human experts on this pilot dataset. The most common kinds of mistakes were miscellaneous errors, which includes misspellings and grammatical errors.

	Perplexity
Base model	
mt5-xl	2.72×10^3
mpt-7B instruct	2.20×10^7
Stablelm-Japanese	1.09×10^6
Fine-tuned models	
mt5-xl	1.43
mpt-7B instruct	113
Stablelm-Japanese	131

Table 2. Per-token perplexity scores on held out data in the validation set, before fine-tuning (base model) and after fine-tuning on a parallel language corpus (fine-tuned models).

	Score				
Evaluation criteria	5	4	3	2	1
Is the original translation provided by the human clear?	119 (56.7%)	72 (34.3%)	16 (7.6%)	3 (1.4%)	0 (0%)
Is LLM translation clear?	32 (15.2%)	98 (46.7%)	56 (26.7%)	21 (10.0%)	3 (1.4%)
Is the LLM translation complete?	82 (39.0%)	70 (33.3%)	37 (17.6%)	21 (10.0%)	0 (0%)
Is the information in the LLM translation correct?	19 (9.0%)	68 (32.4%)	91 (43.3%)	32 (15.2%)	0 (0%)
Is there unnecessary or extraneous information in the LLM translation?	97 (46.2%)	78 (37.1%)	28 (13.3%)	3 (1.4%)	4 (1.9%)
Amount of key* (drug safety related) information in the LLM translation not present in the source text	108 (51.4%)	48 (22.9%)	36 (17.1%)	11 (5.2%)	7 (3.3%)

Table 3. Phase 2 frequencies of each error type in large language model (LLM) generated target text, as determined by human drug safety experts. A score of 5 in each category means the target text was essentially without error, 4 indicates minor errors that do not affect interpretation, 3 indicates errors that might have a small impact on interpretation, 2 indicates an error that would change interpretation, and 1 indicates an unacceptable error. See Supplementary Table S1 for a mapping of the score to the specific questions presented to the human reviewers.

Phase 1 evaluations: evaluation of existing standard produced target text

Using the rubric in Supplemental Table S1, the reviewers evaluated the quality of the original standard supplied target texts. In Table S2, we report summary statistics evaluating whether the standard supplied target text sufficiently captured the same meaning as the source texts. Supplemental Table S3 reports the human-assessed clarity of the source texts and incorporates a two-reviewer system to get an inter-rater agreement in this metric.

Both rater 1 and rater 2's median scores were 4.0 (mostly clear and easy to read). Calculating the inter-rater agreement using Cohen's Kappa, and quadratic weights, gave a Kappa of 0.542. The interpretation of this is typically domain-specific and variable with the number of categories, but in this case represents a much better than random association between the raters and shows consistency in the clarity of the source material.

Supplemental Table S4 reports the Phase 1 reviewers' assessment of the translation accuracy between the standard provided source text and the target text. The human evaluations from the Phase 1 component, in which we checked the set of original data fed into the model (the source ICSR plus the "ground truth" translation), show errors and other issues with the input data. The columns represent, e.g. for "Added information", that there was additional text in the standard provided source text relative to the target text.

Phase 2 evaluation: expert assessment of LLM produced translations

We evaluate the LLM produced translations via the human reviewer's Likert-like criteria (Table 3) and binary criteria (Table 4). When compared to the existing human translation on the same source text in the database, slightly fewer cases had "perfect" (5) clarity scores when generated by LLM (45% vs 56% with human translation). For most categories, the translations were rated as 3 or higher, indicating that they were generally considered acceptable. The notable exception concerned correctness of the LLM translation, which was rated 2 for 12.6%, indicating significant errors that would affect the interpretation (Table 3).

Following the global assessment of the suitability of the translations, a fine-grained assessment was performed by PV experts proficient in English to detect the presence of different error categories in the target text. Adjudication was required by an independent senior PV expert if discordance amongst the evaluations were identified. The most common areas of discordance were "Nonsensical phrases" (46%) and "Wrong dates/times" (42%), and a summary of adjudication is included in the supplemental Table S5.

Low inter-rater agreement between clinical experts when evaluating the same drug-event ICSR is not uncommon in pharmacovigilance. The results (Table 4) showed the LLM translation had errors in categories including dates/times (60% error rate), drug names (59% error rate), AEs (66% error rate), and 62% had nonsensical phrases, including grammatical errors. In the "Other errors" category, the most frequent were typos in drug names, missing causality information, repeated information, incorrect specification of concomitant

Error category	Number (%)
Source contains contradictions	30 (14%)
LLM contains contradictions	86 (41%)
Wrong drug name or information	127 (60%)
Wrong dosage	34 (16%)
Wrong dates/times	135 (64%)
Incorrect/missing AE/wrong outcome	149 (71%)
Rechallenge/dechallenge errors	13 (6%)
TTO issues	48 (23%)
Nonsensical phrases	135 (64%)
Other errors	157 (75%)
Is the case clinically accurate?	73 (35%)

Table 4. Phase 2 fine-grained assessment of the presence of any error in the target text for several important error categories. *LLM* large language model, *AE* adverse event, *TTO* time to onset.

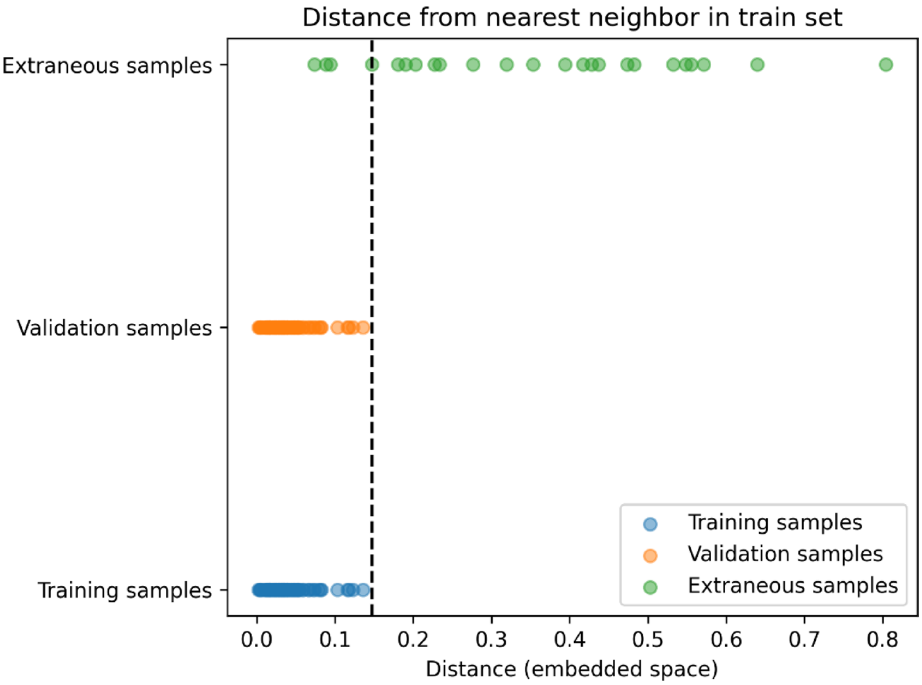


Fig. 2. The distribution of document-level uncertainty scores in extraneous, validation, and training samples. The vertical bar represents the minimum validation sample score that is greater than all the validation and training samples.

medications, incorrect inferred indications, incorrect or missing batch number, and other errors that overlapped with those in other categories (e.g. wrong date).

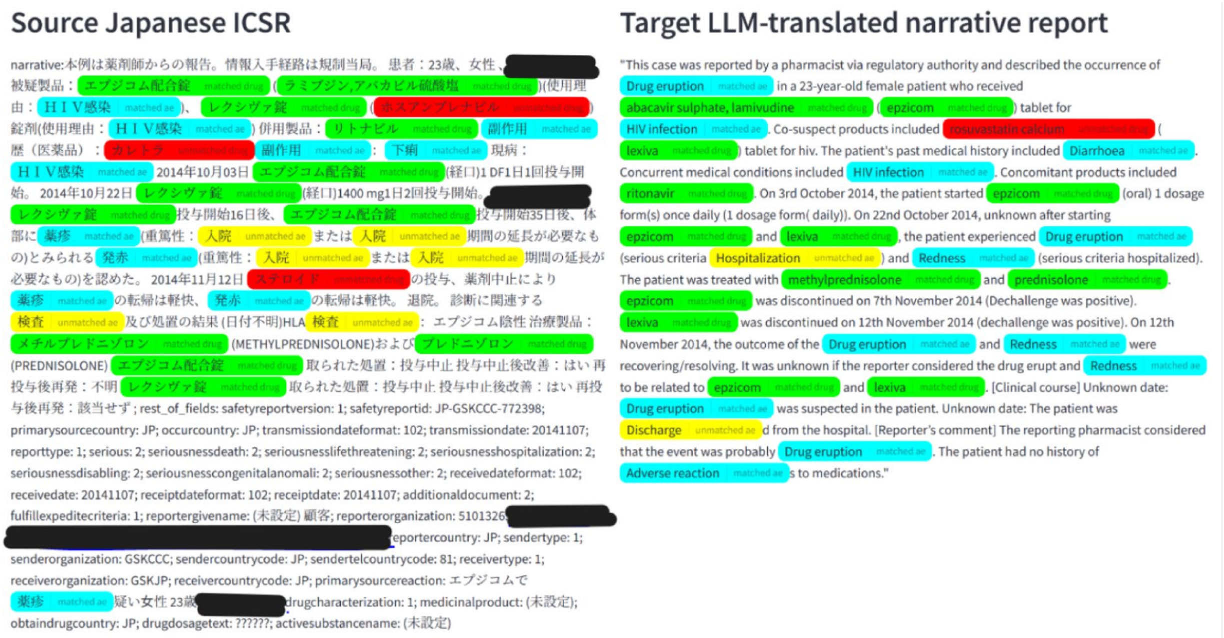
Assessment of DL-UQ guardrail

The DL-UQ metric was applied to training, validation, and non-ICSR Japanese documents. Figure 2 shows that the non-ICSR documents typically had higher distances to their closest training sample in embedding space and, with three counterexamples, can be discriminated from training and validation examples without training.

The distribution demonstrates separation of the assigned scores for the different categories of cases. Although not completely separated, the separation of the validation and extraneous samples corresponds to an AUROC in the validation data of 0.80.

Assessment of MISMATCH guardrail

Figure 3 shows an interface that illustrates the drug and AE MISMATCH guardrail. With this interface, unmatched entities are quickly highlighted, allowing downstream users of this system to understand the specific mismatches that would lead the system to re-route the case to automatic or human-in-the-loop adjudication, and for qualified users to identify and resolve specific issues. For a quantitative evaluation, we report the success



Source Japanese ICSR

narrative:本例は薬剤師からの報告。情報入手経路は規制当局。 患者：23歳、女性、[redacted] 被疑製品：エプジコム配合錠 matched drug (ラミブジン,アバカビル硫酸塩 matched drug)(使用理由：ＨＩＶ感染 matched ae)、 レクシヴァ錠 matched drug (ホスアンプレナビル unmatched drug) 錠剤(使用理由：ＨＩＶ感染 matched ae) 併用製品：リトナビル matched drug 副作用 matched ae 歴（医薬品）：カレトラ unmatched drug 副作用 matched ae ： 下痢 matched ae 現病：ＨＩＶ感染 matched ae 2014年10月03日 エプジコム配合錠 matched drug (経口)1 DF1日1回投与開始。 2014年10月22日 レクシヴァ錠 matched drug (経口)1400 mg1日2回投与開始。[redacted] レクシヴァ錠 matched drug 投与開始16日後、 エプジコム配合錠 matched drug 投与開始35日後、体部に 薬疹 matched ae (重篤性：入院 unmatched ae または 入院 unmatched ae 期間の延長が必要なもの)とみられる 発赤 matched ae (重篤性：入院 unmatched ae または 入院 unmatched ae 期間の延長が必要なもの)を認めた。 2014年11月12日 ステロイド unmatched drug の投与、薬剤中止により 薬疹 matched ae の転帰は軽快、 発赤 matched ae の転帰は軽快。 退院。 診断に関連する 検査 unmatched ae 及び処置の結果 (日付不明)HLA 検査 unmatched ae ： エプジコム陰性 治療製品：メチルプレドニゾロン matched drug (METHYLPREDNISOLONE)および プレドニゾロン matched drug (PREDNISOLONE) エプジコム配合錠 matched drug 取られた処置：投与中止 投与中止後改善：はい 再投与後再発：不明 レクシヴァ錠 matched drug 取られた処置：投与中止 投与中止後改善：はい 再投与後再発：該当せず ; rest_of_fields: safetyreportversion: 1; safetyreportid: JP-GSKCCC-772398; primarysourcecountry: JP; occurcountry: JP; transmissiondateformat: 102; transmissiondate: 20141107; reporttype: 1; serious: 2; seriousnessdeath: 2; seriousnesslifethreatening: 2; seriousnesshospitalization: 2; seriousnessdisabling: 2; seriousnesscongenitalanomali: 2; seriousnessother: 2; receivedateformat: 102; receivedate: 20141107; receiptdateformat: 102; receiptdate: 20141107; additionaldocument: 2; fulfillexpeditecriteria: 1; reportergivename: (未設定) 顧客; reporterorganization: 5101326[redacted]eportercountry: JP; sendertype: 1; senderorganization: GSKCCC; sendercountrycode: JP; sendertelcountrycode: 81; receivertype: 1; receiverorganization: GSKJP; receivercountrycode: JP; primarysourcereaction: エプジコムで 薬疹 matched ae 疑い女性 23歳[redacted]drugcharacterization: 1; medicinalproduct: (未設定); obtaindrugcountry: JP; drugdosagetext: ??????; activesubstancename: (未設定)

Target LLM-translated narrative report

"This case was reported by a pharmacist via regulatory authority and described the occurrence of Drug eruption matched ae in a 23-year-old female patient who received abacavir sulphate, lamivudine matched drug (epzicom matched drug) tablet for HIV infection matched ae . Co-suspect products included rosuvastatin calcium unmatched drug (lexiva matched drug) tablet for hiv. The patient's past medical history included Diarrhoea matched ae . Concurrent medical conditions included HIV infection matched ae . Concomitant products included ritonavir matched drug . On 3rd October 2014, the patient started epzicom matched drug (oral) 1 dosage form(s) once daily (1 dosage form(daily)). On 22nd October 2014, unknown after starting epzicom matched drug and lexiva matched drug , the patient experienced Drug eruption matched ae (serious criteria Hospitalization unmatched ae) and Redness matched ae (serious criteria hospitalized). The patient was treated with methylprednisolone matched drug and prednisolone matched drug . epzicom matched drug was discontinued on 7th November 2014 (Dechallenge was positive). lexiva matched drug was discontinued on 12th November 2014 (dechallenge was positive). On 12th November 2014, the outcome of the Drug eruption matched ae and Redness matched ae were recovering/resolving. It was unknown if the reporter considered the drug erupt and Redness matched ae to be related to epzicom matched drug and lexiva matched drug . [Clinical course] Unknown date: Drug eruption matched ae was suspected in the patient. Unknown date: The patient was Discharge unmatched ae d from the hospital. [Reporter's comment] The reporting pharmacist considered that the event was probably Drug eruption matched ae . The patient had no history of Adverse reaction matched ae s to medications."

Fig. 3. Illustration of guardrails filtering matched and unmatched drug terms and adverse event (AE) terms in the original Japanese ICSR and the LLM produced English case report. Text spans in blue indicate AEs that were successfully matched between the two texts while spans in yellow indicate AEs that were unmatched. Spans in green represent matched drugs while spans in red represent unmatched drugs. The section "narrative" in the Source Japanese ICSR text (the first word) precedes the unstructured text data, and every field after the heading "rest_of_fields" (shown within the text) encompass the rest of the fields.

rate of the MISMATCH guardrail in identifying one "never event", a subset of human evaluator-identified drug issues, in a randomly selected set of 20 cases from the 210 that were evaluated. As can be seen in Fig. 4, all instances of the never event, "spontaneously hallucinated drug names", were correctly identified by the MISMATCH guardrail.

In addition to the error type which we termed a never event, "spontaneously hallucinated drug names", Fig. 4 shows other error categories, including "dictionary incompleteness issues" and "drug spelling issues". Since these guardrails were designed based on known translated maps of generic and trade name drug pairs between English and Japanese, limitations in these dictionaries (that PV experts can spot) lead to guardrail-unaddressed but human-spotted errors, as seen in the "dictionary incompleteness issues". For the "drug spelling issues", the above mentioned phenomena of multiple mentions makes this specific area difficult to address with this current version of guardrails, and as such there were instances of misses in that category as well.

The missrates for the MISMATCH guardrail are summarized in Supplemental Fig. S3 (comparing the model's outputted target text to the source text) and Supplemental Fig. S4 (comparing the original standard's target text to the source text). There were significant amounts of cases with a high ratio of unmatched adverse events, despite using the original standard source translations. Notwithstanding the reviewers' noted imperfection of those translations (see the Phase 1 section), the difference could be explainable by the increased number of ordinary words that are found in AEs. Additionally, comparing these figures to the Fig. 4, we note in addition to adverse events having higher missrates overall, drugs have significant missrates in categories beyond what we describe as "spontaneously hallucinated drug names".

Assessment of TL-UQ guardrail

An example of a visualization of TL-UQ is shown in Fig. 5 and highlights the distribution of the entropy score, which may facilitate efficient and targeted human-in-the-loop review. Figure 6 shows the distributions of TL-UQ uncertainty scores, stratified by clinical accuracy, wrong drug or information, and incorrect/missing AE/wrong outcome. Mann–Whitney U tests (using a Bonferroni correction with n = 9 trials) revealed significant differences in in the "clinical accuracy" stratification (Yes vs. No, adj. *p*-value of 0.0031, Yes/No vs. No, adj. p-value of 0.043) and the "wrong drug" stratification (Yes/No vs. No, adj. *p*-value of 0.028). In each of these cases, the trend was the more "correct" direction. More clinically accurate, less incorrect drugs/AEs trended towards a higher uncertainty score, implying that entropy correlates negatively with the model's human evaluated performance. The observed trend of higher entropy scores correlating with more clinically accurate outputs and fewer incorrect drug mentions may be interpreted as counterintuitive. However, one possible explanation is that the distribution of entropy scores reflects inappropriate model confidence, where the model is more confident in its predictions for cases it is more likely to get wrong. Further investigation is needed to fully understand this pattern, but the results suggest that entropy scores, even at the token level, can provide a useful, if subtle, signal of the model's likely correctness on a given case (see Fig. 5 for example).

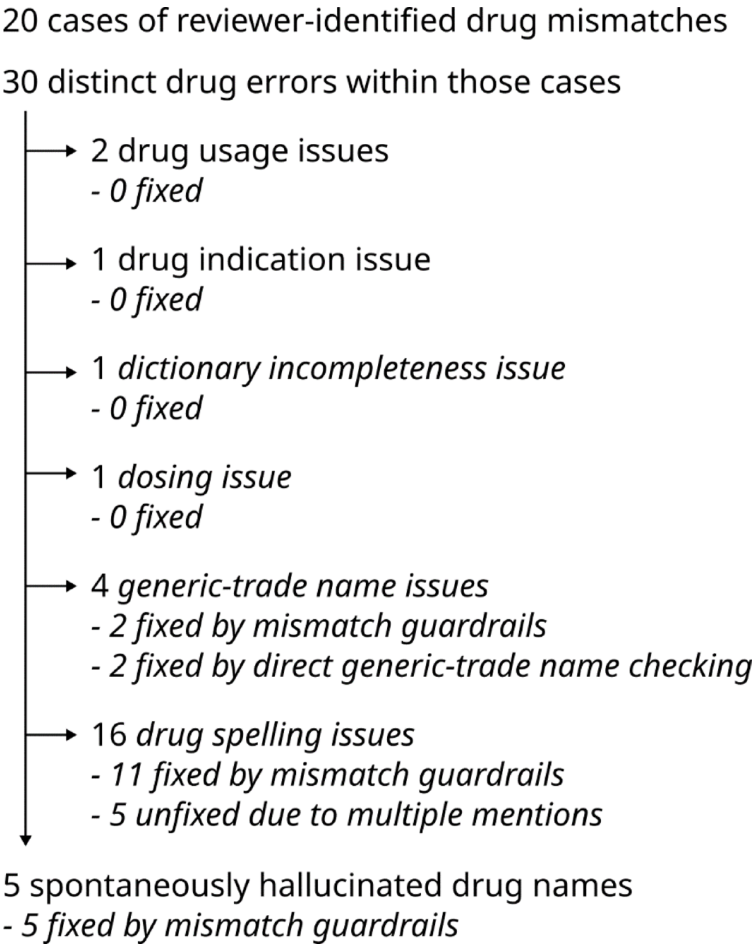
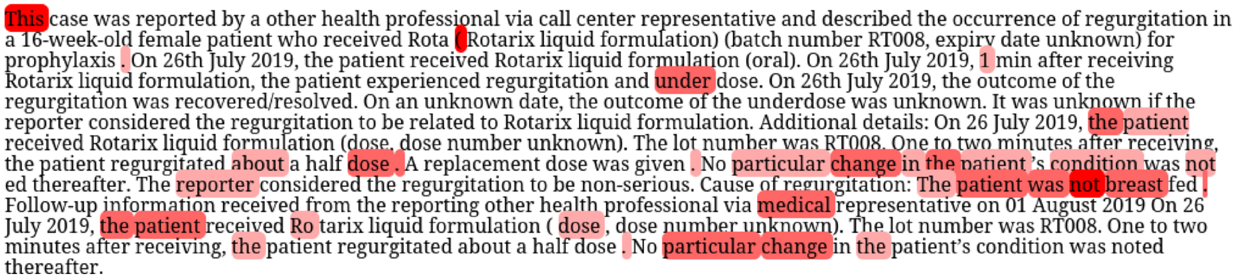

20 cases of reviewer-identified drug mismatches
30 distinct drug errors within those cases
2 drug usage issues
- 0 fixed
1 drug indication issue
- 0 fixed
1 dictionary incompleteness issue
- 0 fixed
1 dosing issue
- 0 fixed
4 generic-trade name issues
- 2 fixed by mismatch guardrails
- 2 fixed by direct generic-trade name checking
16 drug spelling issues
- 11 fixed by mismatch guardrails
- 5 unfixed due to multiple mentions
5 spontaneously hallucinated drug names
- 5 fixed by mismatch guardrails

Fig. 4. Counts of reviewer-identified drug error categories and mismatch guardrail fixes thereof. For each category, counts are given indicating which of the errors had been flagged.



This case was reported by a other health professional via call center representative and described the occurrence of regurgitation in
a 16-week-old female patient who received Rota (Rotarix liquid formulation) (batch number RT008, expiry date unknown) for
prophylaxis . On 26th July 2019, the patient received Rotarix liquid formulation (oral). On 26th July 2019, 1 min after receiving
Rotarix liquid formulation, the patient experienced regurgitation and under dose. On 26th July 2019, the outcome of the
regurgitation was recovered/resolved. On an unknown date, the outcome of the underdose was unknown. It was unknown if the
reporter considered the regurgitation to be related to Rotarix liquid formulation. Additional details: On 26 July 2019, the patient
received Rotarix liquid formulation (dose, dose number unknown). The lot number was RT008. One to two minutes after receiving,
the patient regurgitated about a half dose . A replacement dose was given . No particular change in the patient 's condition was not
ed thereafter. The reporter considered the regurgitation to be non-serious. Cause of regurgitation: The patient was not breast fed .
Follow-up information received from the reporting other health professional via medical representative on 01 August 2019 On 26
July 2019, the patient received Ro tarix liquid formulation (dose , dose number unknown). The lot number was RT008. One to two
minutes after receiving, the patient regurgitated about a half dose . No particular change in the patient's condition was noted
thereafter.

Fig. 5. Example flagged spans using the TL-UQ guardrail. Differing levels of red highlighting correspond to increasing relative scores: least color saturation: between 10th percentile and 5th percentile scores for the whole text. Medium color saturation: between 5th and 1st percentile scores. Least color saturation: 1st percentile and above scores.

Discussion

Our investigation represents a significant step in the application of LLMs within PV, a field where accuracy and safety are paramount. We have explored one of the first integrations of LLMs into the PV workflow, particularly focusing on translating Japanese ICSRs to English. Through the deployment and critical assessment of both hard and soft semantic guardrails, our work confronts the critical challenges associated with LLMs, namely the propensity for hallucinations and the inherent uncertainties associated with model predictions. These approaches are complementary and therefore should be used in conjunction with other strategies to improve the quality of LLM outputs (e.g., temperature adjustments and prompt engineering). Even with safety-critical applications, there is variability in tolerance to inaccurate outputs: the impact of some issues could be so significant that

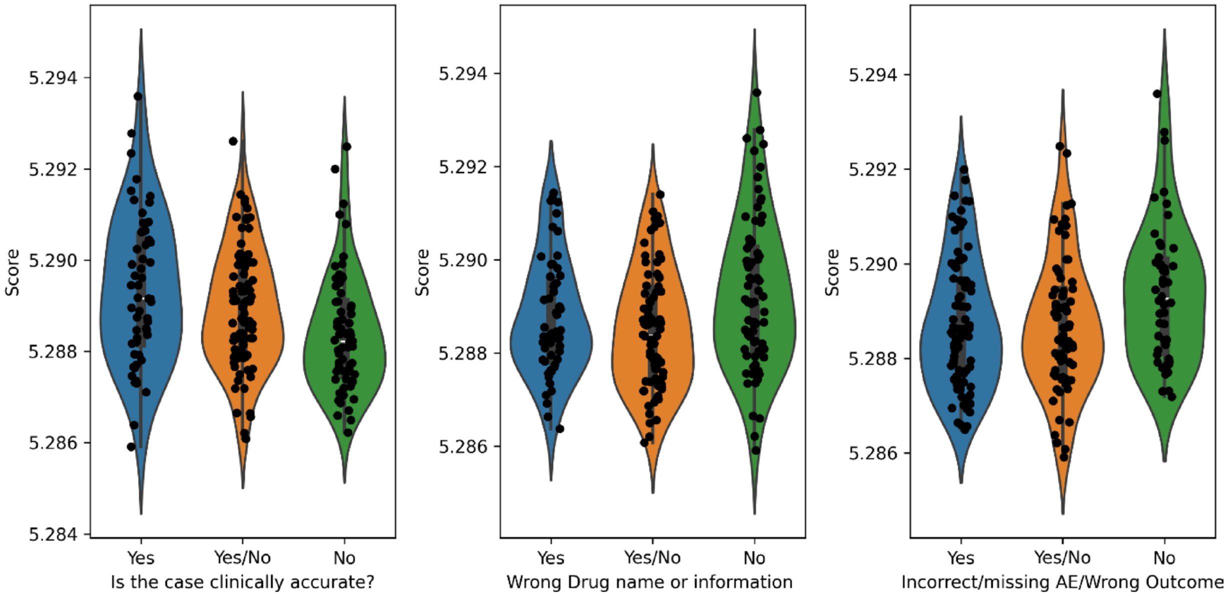


Fig. 6. TL-UQ distributions. Stratifying each reviewed case by "Is the case clinically accurate", "Wrong name or information", and "Incorrect AE/Wrong outcome" and reporting entropy score distributions.

safeguards are needed. In the context of LLM usage, safeguards could be guardrails in addition to or even before full human review. Our findings reveal that strategic guardrail applications effectively mitigate the risk of "never event" errors, with our MISMATCH guardrail successfully identifying every instance of hallucinated drug names in our translated texts from a carefully chosen case sample, although other drug error categories were not universally caught, such as wrong indications and due to dictionary limitations. We anticipate in routine usage as part of quality systems the ability to articulate a priori that certain errors cannot occur. We also note that some erroneous hallucinations could be so problematic that even if human review corrected them, the risk of wrongly recalling them as true outputs could still be problematic: the ability to remove such errors prior to human review holds advantages.

Furthermore, we introduced both document-level and token-level uncertainty guardrails to facilitate a process that incorporates human oversight. The document-level guardrail serves to screen out irrelevant text, reducing unnecessary LLM processing at the ICSR intake stage, whereas the token-level guardrail flags segments of the generated text that exhibit low confidence. These measures immediately make outputs look less definitive and enable the rigorous verification of LLM outputs by skilled human evaluators, who can further investigate and rectify potential inaccuracies. Specifically, the token-level guardrail is designed to highlight areas of high entropy—signifying considerable uncertainty—for thorough review, thereby addressing potential inaccuracies extending beyond specific entities such as drug names or AEs. This approach adds to the burgeoning methodologies aimed at quantifying and communicating model uncertainties to users, supporting human-in-the-loop review and mitigation of risks.

To our knowledge, this project is the first of its kind to develop and implement a range of guardrails for an LLM within the medical safety-sensitive environment of PV. As we look forward, we envision LLMs playing an increasingly central role in this sector, with ongoing improvements enhancing their precision and reliability. Nonetheless, the concept of never events, and its potential extrapolations into other medical safety critical areas, underscores a continuous need for robust guardrails like those we have developed here. The combination of LLMs with these safeguards offers a foundational model for their responsible and efficacious application in PV and beyond.

On the topic of scalability to other domains: the underlying approach is a set of ontology enriched guardrails supporting a text-to-text LLM transformations. Although this approach is still maturing in pharmacovigilance, other safety critical domains such as natural language machine learning tasks in medicine (report generation), and even non-safety critical domains in which strict accuracy could be a "nice to have", for example in consumer facing LLM applications. Areas such as processing electronic health records, for instance, could use very similar guardrails to what we propose: since some systems include structured and unstructured data, and some tasks include producing natural language reports from these sources, MISTMATCH checking of e.g. the diagnostic concepts in the source and target could function as a guardrail in that domain.

Comparison to other hallucination detection systems: The field of generative AI has evolved quickly, and there are now other tools that serve to limit hallucinations in generated language output. For instance, the field of retrieval augmented generation (RAG) has aimed to ground LLM outputs by augmenting the context and pretraining with a knowledge base. The guardrails approach to reduce hallucination is complementary to RAG and other techniques for improving LLM reliability, which are particularly necessary in high-risk contexts; RAG and guardrail frameworks should therefore be used in tandem. Our hard guardrails are intended to completely prevent hallucinations and false negative outputs. In addition, soft guardrails are not designed to directly reduce the error rate of LLM outputs, but instead, serve to enhance the human-computer interaction by enabling more effective human review of uncertain segments or entire input/output instances as part of human-in-the-loop review and oversight processes.

How this is used by non-experts:

Since this pipeline aims to be useful to non-machine learning experts with domain expertise in pharmacovigilance, and in general safety critical domains, there are some additional gaps that can be closed in terms of usability and user experience. The uncertainty scores and TL-UQ distributions, for instance, are in their current form raw numbers that should be calibrated for downstream tasks based on the deployment target. For example, in the uncertainty score, a cutoff value of approximately 0.9 in this trial (Fig. 2) would catch the extraneous samples but only a minority of the training or validation samples, so a downstream piece of software that "alerts" users at this threshold would make this more usable.

Computational resources required to maintain:

The computational resources needed to maintain the hard guardrails are relatively minimal, since only relatively inexpensive operations such as text matching and looking up entries in databases were needed. The uncertainty quantification soft guardrail did entail running computational tasks similar to model inference, so this remains as limited as the end users' ability to use the LLM's inference in the first place.

Limitations

Our work also has several limitations that should be addressed prior to widescale deployment of guardrail frameworks. We focused initial evaluations of hard guardrails on the problem of drug name hallucinations, but there are other kinds of errors that are classified as never events, like misinterpreting exposure outcomes of dechallenge/ rechallenge and AEs. Furthermore, although we did not solve for drug misspellings, this represents a type of error that may be addressed on case intake prospectively, while it could also be resolved by using structured data elements, retrospectively. Further work will extend the list of PV never events and their encoding in the system. Lastly, token-level uncertainty guardrails represent an area of evolving research and will likely continue to improve as the research field produces more solutions to quantify and informatively convey LLM output uncertainty.

In the guardrails as presented in this work, there are two main improvement levers. Firstly, creating accurate underlying ontologies (such as the drug translation pairs, etc.) themselves is a challenging research task. For example, work to expand existing databases of drugs and side effects[33] is ongoing, which when incorporated in our guardrails will more completely catch rare or under documented drugs or side effects. The second lever we believe will directly fall out of using more modern LLMs, which might (although this is subject to future experimentation) be able to more accurately ascribe uncertainty when appropriate, per that guardrail.

Data availability

The datasets generated and/or analysed during the study are not publicly available via GSK as they include sensitive, proprietary post-marketing adverse event data with individually identifiable information. As a private company, GSK must comply with data privacy regulations worldwide and contractual arrangements in place for the information sharing between GSK and Harvard. However whilst we cannot directly share individual case safety reports (ICSRs), worldwide unique case ID numbers (WUCINs) can be provided upon reasonable request (www.safetyinnovation@gsk.com) to request the ICSRs from the Regulatory Authorities. Information relating to AI model characteristics is available at reasonable request from the primary author.

Code availability

We've included the primary analysis code used to parse the human reviewer data and summarize the performance of our system in the attached URL: https://github.com/jlpainter/llm-guardrails/. This is a public GitHub repository with no restrictions to access.

Received: 29 May 2025; Accepted: 25 June 2025
Published online: 31 July 2025

References

1. Tang, L. et al. Evaluating large language models on medical evidence summarization. *Npj Digit. Med.* **6**(1), 158 (2023).
2. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N. *et al.* Large Language Models Encode Clinical Knowledge (2022). http://arxiv.org/abs/2212.13138.
3. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**(1), 141 (2023).
4. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., *et al.* Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models (2023). http://arxiv.org/abs/2309.01219.
5. McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M. & Steedman, M. Sources of Hallucination by Large Language Models on Inference Tasks (2023). http://arxiv.org/abs/2305.14552.
6. Wagle, S., Munikoti, S., Acharya, A., Smith, S. & Horawalavithana, S. Empirical Evaluation of Uncertainty Quantification in Retrieval-Augmented Language Models for Science (2023). http://arxiv.org/abs/2311.09358.
7. Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J. & Hooi, B. Can Llms Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in Llms (2023). http://arxiv.org/abs/2306.13063.
8. Bolton, W. J., Poyiadzi, R., Morrell, E. R., Bueno, G. V. B. G. & Goetz, L. RAmBLA: A Framework for Evaluating the Reliability of LLMs as Assistants in the Biomedical Domain. http://arxiv.org/abs/2403.14578 (2024).
9. European Medicines Agency. Guideline on Good Pharmacovigilance Practices (GVP): Module VI – Collection, Management and Submission of Reports of Suspected Adverse Reactions to Medicinal Products (Rev 2). EMA/873138/2011 Rev 2, 28 July 2017. European Medicines Agency. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-good-pharmacovigilance-practices-gvp-module-vi-collection-management-and-submission-reports-suspected-adverse-reactions-medicinal-products-rev-2_en.pdf

10. Bowen, J. & Stavridou, V. Safety-critical systems, formal methods and standards. *Softw. Eng. J.* **8**(4), 189–209 (1993).
11. World Health Organization. The Importance of Pharmacovigilance (2002).
12. Bate, A. & Evans, S. J. W. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol. Drug Saf.* **18**(6), 427–436 (2009).
13. Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., Meng, J., Ruan, W. & Huang, X. Building Guardrails for Large Language Models. http://arxiv.org/abs/2402.01822 (2024)
14. Health Service Journal. Guidance on implementing the never events framework (2009). https://www.hsj.co.uk/home/guidance-on-implementing-the-never-events-framework/5000691.article.
15. Anderson, J. E. & Watt, A. J. Using safety-II and Resilient healthcare principles to learn from never events. *Int. J. Qual. Health Care* **32**(3), 196–203 (2020).
16. National Quality Framework. List of SREs (2024). https://www.qualityforum.org/Topics/SREs/List_of_SREs.aspx.
17. European Medicines Agency. ICH E2B (R3) Electronic transmission of individual case safety reports (ICSRs): Data elements and message specification implementation guide, Scientific Guideline (2018).
18. Food and Drug Administration. E2B(R3) Electronic Transmission of Individual Case Safety Reports Implementation Guide: Data Elements and Message Specification; and Appendix to the Implementation Guide—Backwards and Forwards Compatibility (2022). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e2br3-electronic-transmission-individual-case-safety-reports-implementation-guide-data-elements-and
19. Zhang, B., Williams, P., Titov, I. & Sennrich, R. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation (2020). http://arxiv.org/abs/2004.11867
20. Graves, A. Sequence Transduction with Recurrent Neural Networks (2012). http://arxiv.org/abs/1211.3711.
21. Su, Y. et al. A contrastive framework for neural text generation. *Adv. Neural. Inf. Process. Syst.* **35**, 21548–21561 (2022).
22. Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. Bleu: A method for automatic evaluation of machine translation. in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* 311–18 (2002).
23. Post, M. A Call for Clarity in Reporting BLEU Scores (2018). http://arxiv.org/abs/1804.08771.
24. Klakow, D. & Peters, J. Testing the correlation of word error rate and perplexity. *Speech Commun.* **38**(1–2), 19–28 (2002).
25. Kara, V. et al. Finding needles in the haystack: Clinical utility score for prioritisation (CUSP), an automated approach for identifying spontaneous reports with the highest clinical utility. *Drug Saf.* **46**(9), 847–855 (2023).
26. Koch-Weser, J., Sellers, E. M. & Zacest, R. The ambiguity of adverse drug reactions. *Eur. J. Clin. Pharmacol.* **11**, 75–78. https://doi.org/10.1007/BF00562895 (1977).
27. Arimone, Y. et al. Agreement of expert judgement in causality assessment of adverse drug reactions. *Eur. J. Clin. Pharmacol.* **61**, 169–173. https://doi.org/10.1007/s00228-004-0869-2 (2005).
28. Arimone, Y. et al. Inter-expert agreement of seven criteria in causality assessment of adverse drug reactions. *Br. J. Clin. Pharmacol.* **64**(4), 482–488. https://doi.org/10.1111/j.1365-2125.2007.02937.x (2007).
29. Kosov, M., Maximovich, A., Riefler, J., Dignani, M. C., Belotserkovskiy, M. & Batson E. Interexpert agreement on adverse events' evaluation. Applied Clinical Trials Online (2016).
30. Likert, R. A technique for the measurement of attitudes. Archives of Psychology (1932).
31. Brown, E. G., Wood, L. & Wood, S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.* **20**(2), 109–117 (1999).
32. Google. Evaluating models|AutoML Translation Documentation. (2024). https://cloud.google.com/translate/automl/docs/evaluate.
33. Kuhn, M. et al. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**(D1), D1075–D1079 (2016).

Acknowledgements

The authors would like to thank the GSK staff from Japan safety and Global Safety who participated in phase 1 and 2 of the study; Phase 1: Asako Takata, Kohei Ogawa, Toshifumi Kimura, Tomoko Matsukawa, Kaoru Fujikura, Yoko Hijioka, Hiroko Toyota, Tamami Kaneko, Hiroki Nagahama, Takako Watanabe, Naoki Kaneko, Akiko Suhara; Phase 2: Tony Ning, Weronika Dardzinska, Marta Krzywdzinska-Kopacz, Joanna Kawałek, Chetan Sharma, Manju Uttam, Avinash Chaturvedula, Anupama Mathew, Aparna Jayachandra, Abhinaya Surender, Kaverappa M D, and Ewa Nowicka. In addition, students from the University of North Carolina, USA who supported Phase 2: Nathan Andert, Herbert Wan and Jackie Tan and GSK Global Safety Staff François Haguinet and Olivia Mahaux who supported with the multilingual assessments.

Author contributions

JBH, JLP, and A. Beam contributed to the study concept, data acquisition, data analysis, and data interpretation. DR, VK, and A. Bate contributed to the study concept, and data interpretation. GP contributed to data interpretation, and PS and CS contributed to the data analysis and data interpretation.

Funding

This research was funded by GlaxoSmithKline Biologicals S.A.

Declarations

Competing interests

All GSK co-authors (Jeffery L Painter, Darmendra Ramcharran, Vijay Kara, Greg Powell, Paulina Sobczak, Chiho Sato, Andrew Bate) receive GSK salary and some hold GSK stock and stock options. Andrew L Beam is a consultant for Generate Biomedicines and Flagship Pioneering, Inc and holds stock and stock options in Generate Biomedicines and FL 85, Inc. The rest of the co-authors (Joe B. Hakim) have no conflict of interest to state.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-09138-0.

Correspondence and requests for materials should be addressed to A.B.

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

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