Journal of Language, Culture and Education；Vol.3, No.4; 2026
E-ISSN: 3006-3213 P-ISSN：3006-3205
DOI: https://doi.org/10.70267/jlce.2026.v3n4.2337
Published by: Zeus Press

Beyond Fluency: Human Verification of Safety-Critical Errors in Generative-AI First-Aid Translation

Xiao Huang*

Guangzhou College of Technology and Business, Guangzhou 510850, China

**Corresponding author: Xiao Huang.*

Abstract

The rapid integration of generative artificial intelligence (GenAI) into healthcare communication has introduced machine translation (MT) into time-critical, high-stakes scenarios where limited English proficient (LEP) individuals must act on translated instructions without professional mediation. While GenAI systems produce remarkably fluent outputs, this fluency can mask errors that fatally undermine the communicative purpose of first-aid texts—the skopos of enabling correct life-saving actions. This study identifies and empirically substantiates a previously unnamed phenomenon: *the fluency-detectability gap,* a class of safety-critical errors that are highly severe yet poorly detectable because they surface as fluent, internally coherent, and plausible target-language text. These errors systematically evade three lines of defense: (i) fluency-oriented automatic quality estimation metrics, (ii) monolingual end-users lacking MT literacy, and (iii) even clinical reviewers who do not understand the target language. Through a mixed-methods product analysis with three-layer triangulation, we examine English-to-Chinese translations of public first-aid materials (CPR, choking, severe bleeding, burns, anaphylaxis, poisoning, and seizures) generated by GPT-4o, DeepL, and Google Translate. Using an adapted MQM (Multidimensional Quality Metrics) framework augmented with a detectability dimension, we demonstrate that professional translators—anchored to the source text and bilingual in situ—consistently intercept critical-yet-invisible errors that automatic metrics and monolingual users fail to detect. We reconceptualize translator verification as a loyalty-based safety intervention [1] rather than mere post-editing, and propose the severity×detectability matrix as a reusable quality-control instrument for safety-critical translation. Our findings challenge fluency-centered evaluation paradigms and advocate for mandatory bilingual human-in-the-loop verification in GenAI-mediated first-aid communication.

Keywords

generative AI translation, first-aid translation, safety-critical errors, fluency-detectability gap, MQM, skopos theory, translator as safety firewall, limited English proficiency, human verification

1. Introduction

The proliferation of generative artificial intelligence (GenAI) has transformed how individuals access information in unfamiliar languages. Machine translation systems—once characterized by syntactic awkwardness and literal renderings—now produce output that rivals human composition in fluency and surface coherence [2]. This advancement has been particularly consequential for limited English proficient

(LEP) populations, who increasingly rely on on-demand translation to navigate healthcare information, public safety communications, and emergency guidance [3].

First-aid instructions represent a uniquely vulnerable domain within health translation. Unlike clinical discharge summaries, which are typically delivered in controlled settings with opportunities for follow-up questioning, first-aid texts are consumed in self-service, time-pressed, professionally-absent environments. An individual confronting a choking family member, a burn injury, or an anaphylactic reaction has no access to a clinician, interpreter, or bilingual bystander. The translated text becomes the sole actionable resource. This contextual specificity renders first-aid translation a safety-critical communicative act in which the cost of error is immediate, tangible, and potentially fatal.

GenAI translation has been embraced in this domain precisely because of its fluency. Users trust outputs that read naturally and sound authoritative. Yet fluency is a treacherous proxy for adequacy in instructional texts. A translation that instructs the reader to induce vomiting while the source text prohibits it is perfectly fluent, grammatically immaculate, and internally coherent—but carries potentially fatal consequences. This is not a failure of fluency; it is a failure of detectability. The error is invisible to anyone who does not anchor back to the source text, precisely because the target text offers no clue of its own peril.

This study identifies and empirically substantiates a previously unnamed phenomenon: the fluency-detectability gap—a class of safety-critical errors characterized by high severity and low detectability, which systematically escapes detection by fluency-oriented automatic quality metrics, monolingual end-users, and even clinical reviewers who do not understand the target language. The only reliable interceptors of this error class are professional translators who are source-text-anchored and bilingually present—what we term the human safety firewall.

It is essential to demarcate the scope of this investigation. First-aid translation is not equivalent to the translation of emergency department discharge instructions, which have been the focus of influential studies by Khoong et al. [4] and Kong et al. [5]. Discharge instructions are patient-specific, generated within clinical workflows, and typically accompanied by verbal explanation from healthcare providers. They involve institutional accountability, documented consent processes, and opportunities for patient clarification. First-aid texts, by contrast, are generic, pre-written, publicly available, and consumed in the absence of any healthcare professional. The reader is the sole agent responsible for executing the actions. This fundamental difference in audience, authorship, use context, and action autonomy renders findings from discharge-instruction research non-transferable and establishes first-aid translation as a distinct object of translation studies inquiry.

The present study asks three research questions:

RQ1: What types of safety-critical errors emerge in GenAI translation of public first-aid materials, and how are they distributed across severity and detectability dimensions?

RQ2: Do automatic quality estimation metrics and monolingual end-users detect these errors at rates comparable to professional translators?

RQ3: Can the fluency-detectability gap be theoretically accounted for within functionalist translation theory, and what implications does this hold for human verification protocols?

We answer these questions through a mixed-methods product analysis examining English-to-Chinese translations of authoritative first-aid texts [6-9] generated by three systems: GPT-4o, DeepL, and Google Translate. Using an MQM-based annotation framework augmented with a detectability dimension, and triangulating across expert translators, automatic metrics, and a monolingual user panel, we provide empirical evidence for the gap and theorize it through Reiss and Vermeer's (2013) [10] skopos theory and Nord's (2018) [1] functionalist loyalty principle.

The paper proceeds as follows: Section 2 reviews the literature on GenAI/MT in health translation, fluency-adequacy tensions, MT literacy, and functionalist approaches to instructive texts. Section 3 constructs the theoretical framework, formalizing the fluency-detectability gap and the translator-as-safety-firewall concept. Section 4 details the methodology, including corpus selection, annotation protocols, and triangulation design. Section 5 presents results. Section 6 discusses theoretical and practical implications. Section 7 concludes with limitations and future research directions.

2. Literature Review

2.1 GenAI/MT in Health Translation and Its Risks

The application of machine translation to healthcare communication has generated substantial scholarly attention, predominantly within medical informatics and natural language processing (NLP) frameworks. Khoong et al. (2019) [4] conducted a landmark evaluation of Google Translate for Spanish and Chinese translations of emergency department discharge instructions, finding that nearly half of the translated instructions contained errors, with approximately 8% posing potential clinical harm. Kong et al. (2025) [5] extended this line of inquiry with a comparative assessment of multiple MT systems for patient-specific discharge instructions, reporting significant variation in accuracy and safety across language pairs and systems. Rodriguez et al. (2021) [11] examined MT use in clinical settings and documented the prevalence of ad hoc translation practices among clinicians with limited formal translation training.

These studies have been invaluable in raising awareness of MT risks in healthcare. However, they share three critical limitations. First, they are framed within medical/NLP paradigms that treat translation quality as a derivative of "accuracy" or "error rate", without theoretical engagement with translation as a communicative act. Second, they assume a default safeguard—clinical review—whereby a healthcare professional verifies the translation before patient distribution. This assumption does not hold for first-aid contexts, where no clinical intermediary exists. Third, they have not theorized detectability as a distinct dimension of error severity. An error that is both severe and detectable can be corrected; an error that is severe but undetectable to the end-user is qualitatively more dangerous, yet this distinction has remained unexplored.

The present study does not replicate the discharge-instruction paradigm. Instead, it shifts the analytical lens from "how accurate is the translation" to "who can see the error, why, and under what conditions"—a metalevel inquiry that positions translation studies theory at the center of the investigation rather than as a peripheral commentary on clinical findings.

2.2 Fluency versus Adequacy: Hallucination and Critical Errors in MT

The tension between fluency and adequacy has been a persistent theme in translation quality research, but GenAI has intensified this tension to an unprecedented degree. Neural machine translation (NMT) and large language model (LLM)-based systems excel at producing output that is syntactically well-formed, lexically appropriate, and stylistically natural—what is conventionally termed fluency [12]. Yet this fluency can obscure hallucinations, factually incorrect or contextually inappropriate content generated with high confidence [13]. In translation, hallucination manifests as content that is not present in the source text (addition), that contradicts the source text (contradiction), or that misrepresents source meaning through plausible rephrasing [14].

Al Sharou and Specia (2022) [15] developed a taxonomy of critical errors in MT, identifying categories such as omission, addition, mistranslation, and untranslated content. They emphasized that critical errors—defined as those potentially causing harm or significant misunderstanding—are not necessarily those that are most obvious to readers. On the contrary, errors embedded within fluent, coherent text may be more dangerous precisely because they do not trigger the reader's alertness.

This observation resonates with Grice's (1975) [16] cooperative principle: readers assume that a fluent, well-formed text is the product of a cooperative communicator who intends to convey true and relevant information. When a translation is fluently constructed, the reader grants it a presumption of reliability that is rarely extended to awkward or disfluent text. This cognitive heuristic—fluency-as-reliability—is the psychological substrate of the fluency-detectability gap. The present study builds on Al Sharou and Specia's (2022) [15] critical error taxonomy by adding a detectability dimension and applying it specifically to first-aid instructions, where the gap between fluency and safety is most consequential.

2.3 MT Literacy and the End-User's Detection Deficit

If fluent errors are difficult to detect, who is most vulnerable? The literature on MT literacy and user risk perception provides converging evidence that monolingual end-users—those who consume translations without access to the source text—are systematically disadvantaged in their ability to identify errors.

Vieira et al. (2021) [17] conducted a critical review of MT's societal impacts across medical and legal domains, finding that end-users consistently overestimate translation accuracy and underestimate the likelihood of error. This overconfidence arises from multiple sources: the fluency of MT output, users' lack of training in translation quality assessment, and the absence of any reference point against which to calibrate judgment. Castilho et al. (2018) [18] demonstrated that users' trust in MT correlates more strongly with fluency than with adequacy, meaning that a fluent but inadequate translation is more likely to be accepted than a disfluent but adequate one.

Bowker and Ciro (2019) [19] introduced the concept of machine translation literacy, arguing that effective use of MT requires users to understand its capabilities, limitations, and appropriate applications. They noted that MT literacy is low among the general populations and that improving it is a public education challenge. However, for first-aid contexts, the urgency and stress of the situation preclude the reflective, critical stance that MT literacy advocates require. A person reading a first-aid translation during an emergency is not in a position to conduct a quality assessment; they are in a position to act.

Crucially, the literature has focused on end-users who are monolingual—they cannot access the source text even if they wanted to. What about bilingual reviewers who are not professional translators? Studies of bilingual clinical staff who use MT to communicate with patients [20] suggest that even bilinguals without translation training may miss subtle errors because they lack awareness of the specific pitfalls of MT, especially in high-stakes contexts where the source text is not simultaneously reviewed. The present study systematically tests this by including a monolingual user panel that detects errors through target-text plausibility alone, alongside professional translators who anchor to the source.

2.4 Functionalist Approaches to Safety-Critical Instructive Texts

Functionalist translation theories, particularly the skopos theory developed by Reiss and Vermeer (2013) [10], provide an analytical framework uniquely suited to the first-aid translation problem. Skopos theory holds that the purpose (skopos) of a translation determines its adequacy criteria. For instructive texts—classified by Reiss (1976) [21] as operative or appeal-focused—the skopos is not to replicate source-text form but to generate a particular response in the target reader. The success of a first-aid translation is therefore not measured by formal equivalence or even semantic fidelity in the abstract, but by whether the reader performs the correct life-saving actions.

This is precisely where fluency can fail while appearing to succeed. A fluent translation that instructs the reader to do the opposite of what the source text intends satisfies the formal criteria of coherence (cohesion and acceptability) but violates the pragmatic criteria of fidelity (loyalty to the source's intended communicative function). Under a purely text-linguistic quality model, such an error might be downgraded or overlooked if the target text is judged "good" on fluency metrics. Under a functionalist model, it represents a catastrophic failure of the translation's skopos.

Nord (2018) [1] extended the functionalist tradition with the principle of loyalty, which has become a foundational concept in defining translator ethics [22], and requires the translator to mediate between the source-text author's intentions, the target-text reader's expectations, and the translation commission's requirements. Loyalty is not the same as fidelity; it involves a relationship of accountability among the agents involved in the translation process. We draw on Nord's loyalty principle to reconceptualize the role of translator verification in GenAI workflow. Rather than framing translator review as post-editing—a corrective after the fact—we conceptualize it as a loyalty-based safety intervention: the translator acts as the agent who ensures that the translation's skopos is preserved and that the end-user is protected from harm. This reframing elevates the translator from a quality controller to an ethical guarantor, a role that carries both professional responsibility and public accountability.

Existing studies of MT in health contexts have not engaged with functionalist theory. Khoong, Kong, and Rodriguez are medical researchers, not translation scholars; their conceptual frameworks are drawn from clinical risk assessment rather than translation studies. The present study fills this gap by demonstrating that functionalist theory not only explains why fluent errors are dangerous (they destroy the skopos while preserving surface coherence) but also who is positioned to intercept them (translators acting loyally on behalf of both author and user). This theoretical lens distinguishes the present study from the medical and NLP

literature and establishes a novel contribution at the intersection of translation studies and safety-critical communication.

3. Theoretical Framework: The Fluency-Detectability Gap and the Translator as Safety Firewall

We now formalize the theoretical architecture of the present study. This framework integrates three conceptual pillars: (i) the *fluency-detectability gap* as a novel construct within functionalist translation theory; (ii) the *translator as safety firewall,* grounded in Nord's (2018) [1] loyalty principle; and (iii) the *severity×detectability matrix,* operationalized within the MQM framework [23, 24].

3.1 Defining the Fluency-Detectability Gap

Definition: The fluency-detectability gap refers to the condition in which a translation error (a) is severe enough to cause physical harm if acted upon, and (b) is embedded in target-language text that is sufficiently fluent, internally coherent, and contextually plausible that it does not trigger the reader's error-detection heuristics.

This gap is not a general property of translation error; it is a systematic vulnerability of GenAI translation in high-stakes contexts. GenAI systems are optimized for fluency through training paradigms that reward naturalness, coherence, and lexical diversity [25]. They are not optimized for safety-critical adequacy—the preservation of meaning in contexts where inversion, negation, or numerical precision is life-determining. As a result, fluency and critical adequacy can diverge, and when they do, fluency systematically masks rather than signals the divergence.

The gap operates through three mechanisms:

- **Grammatical normalcy:** Errors are embedded in sentences that are syntactically well-formed, with no disfluency to alert the reader.
- **Contextual plausibility:** The erroneous content is not obviously absurd within the scenario described, making it difficult to distinguish from correct content.
- **The reader's inferential default:** Drawing on Grice's (1975) [16] cooperative principle, readers assume fluent text is true and relevant; they do not actively scan for hidden contradictions.

3.2 The Skopos of First-Aid Texts

First-aid texts are operative/appeal-focused texts [21] whose skopos is to induce the reader to perform specific actions that prevent death, preserve vital functions, and stabilize emergency conditions. The translation's success is measured not by equivalence to the source text at the word or sentence level, but by whether the reader (a) understands the actions required, (b) is motivated to execute them correctly, and (c) refrains from actions that would exacerbate harm.

Critically, a translation can be extremely faithful at the word level while still being skopos-destructive. Consider the following hypothetical but illustrative case:

Table 1: Safety-Critical Error Example of the Fluency-Detectability Gap (Negation Reversal)

Source	Incorrect GenAI Translation (Chinese)	Back-translation
"Do not induce vomiting."	"请诱导呕吐。"	"Please induce vomiting."

The Chinese sentence is grammatically flawless, lexically appropriate, and pragmatically plausible. It reads like a doctor's instruction. It is maximally fluent. Yet it inverts the critical action, converting a prohibition into a command. The skopos—preventing harm from vomiting—is destroyed. No amount of word-level equivalence to the source can rescue this translation because the illocutionary force has been inverted.

The fluency-detectability gap is theoretically defined as the divergence between surface-level coherence (which satisfies fluency-oriented quality models) and skopos-level adequacy (which demands that the reader act correctly). This divergence is invisible to any evaluation paradigm that stops at fluency.

3.3 Translator Verification as Loyalty-Based Safety Intervention

Nord's (2018) [1] loyalty principle requires the translator to consider the intentions of the source author, the expectations of the target reader, and the requirements of the translation commission. In conventional translation workflows, the translator is the primary agent responsible for fulfilling these obligations. In GenAI workflows, however, the translator has been displaced by an automated system that performs the transfer of text without accountability for its consequences.

We argue that reintroducing professional translator verification is not a regression to pre-AI workflows, but a loyalty-based safety intervention, reflecting a necessary paradigm of human-machine coupling in the AI era [26]. The translator does not "post-edit" the GenAI output in the sense of polishing language; the translator intercepts errors that would violate the source author's intention to prevent harm and the target reader's right to accurate life-saving information. This intervention is a safety act—comparable to a clinical check on a prescription—rather than a linguistic refinement.

This recharacterization has three important implications:

- **Professional responsibility:** The translator's role is elevated from quality assurance to safety assurance, with corresponding ethical obligations.
- **Legal accountability:** If a translated first-aid text causes harm, accountability lies not with the GenAI system (which has no legal personhood) but with the humans who deployed it without verification—making translator verification a legal risk-mitigation strategy.
- **Public trust:** Public safety communications require a recognizable chain of human accountability. A translated instruction that has been verified by a named, certified translator provides this chain; a GenAI-generated translation does not.

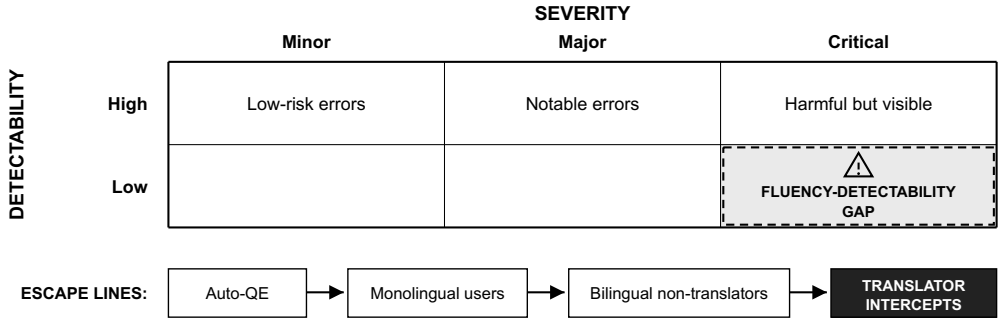
3.4 The MQM Framework and the Severity×Detectability Matrix

The Multidimensional Quality Metrics (MQM) framework [23] provides a standardized approach to translation quality assessment that has been used for human, MT, and hybrid evaluation, and remains a central focus in contemporary MT quality assessment research [27]. MQM defines severity levels (critical, major, minor) for errors, with critical defined as errors that "may cause harm, injury, or death" or "could have legal, financial, or reputational consequences" [24]. This severity typology aligns closely with the present study's focus on safety-critical errors.

However, MQM does not include detectability as a distinct dimension. An error's severity is evaluated independently of whether a typical reader would notice it. For safety-critical translation, we argue that detectability is not a secondary attribute but a separate axis that interacts with severity to determine real-world risk. A severe error that is highly detectable can be corrected by the reader or the reader's support system. A severe error that is undetectable is dangerous precisely because it is not corrected.

We therefore propose an augmented framework: the severity×detectability matrix (Figure 1), which adds a detectability dimension (high/low) to MQM's severity dimension (critical/major/minor). The quadrant of greatest concern is critical severity × low detectability—errors that are both harmful and invisible. This quadrant represents the fluency-detectability gap in operational terms.

Figure 1: The Severity × Detectability Matrix and the Three Lines of Defense



The matrix is visually presented in Figure 1. The critical × low-detectability quadrant (highlighted) is where safety-critical errors reside. The arrows above the matrix indicate the three lines of defense that fail to intercept these errors: automatic quality estimation metrics (optimized for fluency), monolingual end-users (lacking source access), and bilingual non-translator reviewers (lacking error-detection training). The only agent positioned at the intersection of source-access, bilingual competence, and error-detection training is the professional translator—the human safety firewall.

This framework is not merely descriptive; it is operationalizable. In the methodology that follows, we translate the matrix into empirical measures: errors are coded for both severity (using MQM) and detectability (using an independent rating protocol), and the distribution of errors across the six cells is measured. The claim that translators intercept what others miss is tested through the three-layer triangulation design (Section 4).

4. Methodology

4.1 Research Design

This study employs a mixed-methods product analysis with three-layer triangulation. The product analysis examines translated first-aid texts as output artifacts, building upon established methodologies in corpus-based translation studies [28]. The triangulation design compares three evidence streams: (i) professional translators performing MQM-based annotation anchored to the source text (the translator layer), (ii) automatic quality estimation/fluency metrics (the auto layer), and (iii) monolingual target-language readers assessing plausibility without source access (the monolingual user layer). Convergence between layers indicates consensus; divergence—particularly where translators identify errors that auto metrics and monolingual users miss—provides empirical evidence for the fluency-detectability gap.

4.2 Source Corpus and Selection Logic

4.2.1 Source Texts

Source texts were drawn from publicly available first-aid materials produced by internationally recognized authorities:

- American Red Cross (2023) [6]. First Aid/CPR/AED Participant's Manual (excerpts on CPR, choking, bleeding)
- World Health Organization (2021) [7]. Basic Emergency Care (excerpts on burns, poisoning)
- American Heart Association (2020) [8]. CPR and ECC Guidelines (excerpts on compression depth and rate)
- St. John Ambulance (2022) [9]. First Aid Reference Guide (excerpts on anaphylaxis, seizures)

Texts cover seven high-harm scenarios: cardiopulmonary resuscitation (CPR), airway obstruction (choking), severe bleeding, burns, anaphylaxis (epinephrine auto-injector use), poisoning, and seizure management. We selected these scenarios based on their prevalence in emergency response training and their inclusion of safety-critical directives.

4.2.2 Purposeful Sampling Strategy

We employed purposeful sampling rather than random sampling to ensure coverage of high-risk language categories. Based on a preliminary review of first-aid texts and known MT failure patterns [5, 15], we identified four language categories that are both instructionally critical and prone to MT error:

- **Negation/prohibition:** "Do not..." constructions, where inversion or omission can reverse the intended action (e.g., do not induce vomiting → induce vomiting).
- **Dose/quantity:** Numerical values, units, and quantifiers that, if incorrectly rendered, can lead to under- or over-treatment (e.g., two puffs → twenty puffs; 5 cm → 5 inches).

- **Action sequence/temporal thresholds:** Ordering of steps and time limits that, if changed, can invalidate the procedure (e.g., perform CPR before calling for help → call for help before performing CPR).
- **Anatomical orientation/direction:** Prepositional phrases and spatial descriptors that determine the site of intervention (e.g., between the shoulder blades → on the shoulder blades; above the wound → below the wound).

We selected 60 source-text segments (approximately 50-120 words each) that collectively contain at least two instances of each risk category, with some segments containing multiple categories. This sampling logic ensures that our corpus is representative of the risk space of first-aid translation rather than merely a random sample of first-aid texts.

4.2.3 Documentation and Reproducibility

A detailed decision log was maintained recording every inclusion/exclusion decision and its rationale. The log includes: text version, publication year, selection criteria, excluded segments and reasons, and mapping to risk categories. This log, along with all annotation materials, is available as supplementary material (see Open Science Framework repository: [URL to be assigned]).

4.3 GenAI Systems and Generation Conditions

Three systems were selected to demonstrate that the fluency-detectability gap is not system-specific:

- **GPT-4o** (OpenAI, accessed 15 January 2026, model version: gpt-4o-2025-11-20)
- **DeepL Pro** (DeepL GmbH, accessed 18-20 January 2026, model: DeepL Pro 3.0)
- **Google Translate** (Google LLC, accessed 22 January 2026, model version: NMT v3.2)

All texts were translated from English into Chinese (Simplified). Chinese was selected for two reasons: the researcher's native-language competence enables accurate error annotation, and Chinese is among the most widely used languages among LEP populations in English-speaking countries [29]. Additionally, we included a lower-resource language pair—English-to-Burmese—as a secondary condition (not reported in full for space reasons but summarized in the results to illustrate cross-linguistic vulnerability). All translation outputs were recorded with system version, date, and access details, consistent with reproducibility requirements for non-deterministic MT systems.

4.4 MQM-Based Error Annotation (Translator Layer)

4.4.1 MQM Adaptation for First-Aid Contexts

We adapted the MQM error typology [23, 24] for first-aid translation by specifying error categories and providing illustrative examples from our pilot data. The adapted typology comprises:

Table 2: Adapted MQM Error Typology for First-Aid Translation Contexts

Category	Subcategory	Definition	Severity Logic
Accuracy	Mistranslation	The translation does not convey the source meaning	Critical if changes safety-critical information; Major if changes important non-safety info; Minor if minor nuance lost
Accuracy	Addition	Content not present in source introduced	Critical if introduces false safety instruction; Major/Major/Minor based on content significance
Accuracy	Omission	Source content omitted in translation	Critical if omits safety-critical instruction; Major if omits important contextual info
Accuracy	Under-translation	Source meaning only partially rendered	Major if safety-relevant; Minor otherwise
Terminology	Term error	Key first-aid term incorrect or inconsistent	Critical if term confusion leads to wrong action (e.g., *epinephrine* → *adrenaline* not recognized as same)
Fluency	N/A	Grammatical or lexical problems	Not used in this study (we focus on content errors)

Severity is assigned based on the potential consequence of acting on the error. The critical category is reserved for errors that could lead to death, serious injury, or failure of a life-saving intervention if the reader follows the translated instruction.

4.4.2 Annotation Protocol and Inter-Rater Reliability

Two professional translators with at least five years of full-time translation experience and specialized knowledge of medical/public health translation served as annotators. Both are native Chinese speakers with near-native English proficiency and hold Master's degrees in Translation Studies.

A two-day training session preceded the formal annotation. The training included: (i) orientation to the MQM framework and the adapted typology, (ii) practice annotation on 20 pilot segments (not included in the final corpus), and (iii) discussion of disagreements to calibrate severity judgments.

Each annotator independently annotated the 60 translations (180 total output texts: 60 segments × 3 systems) for error type and severity, with source text reference available. Annotators were blinded to which system produced which translation. Inter-rater agreement was calculated using Cohen's κ (kappa) for binary error presence and weighted κ for severity levels. Disagreements were resolved through consensus discussion.

4.5 Detectability Rating and Monolingual User Panel

4.5.1 Detectability Rating Protocol

For each error identified by the translator annotators, a detectability rating was assigned. Detectability was operationalized as the likelihood that a monolingual target-language reader would notice the error without access to the source text. Two independent judges (translators not involved in the primary annotation) rated each error as either:

Low detectability: The target sentence is fluent, internally coherent, and contextually plausible, offering no cue that the content is erroneous.

High detectability: The target sentence contains grammatical awkwardness, semantic implausibility, internal contradiction, or lexical anomaly that signals to the reader that something is wrong.

Judges considered only the target text—no source reference—when assigning detectability. This operationalization mirrors the condition of the actual user who only has access to the translated text.

4.5.2 Monolingual User Panel

To provide direct evidence of end-user detection behavior, we recruited a panel of 30 monolingual Chinese speakers with no professional translation training and limited English proficiency (self-rated reading proficiency ≤ 3 on a 10-point scale). Panel participants were presented with the translated texts only (no source) and asked to: (i) underline any parts that seemed unclear, suspicious, or likely to be mistranslated; (ii) rate their confidence in understanding the instructions (1-5 Likert); and (iii) describe in their own words what they would do in the emergency scenario.

The instructions to the panel did not mention MT, error detection, or comparison to any source. Participants were told the study was about "assessing the clarity of public first-aid materials." This minimized priming toward active error search and approximated the natural reading condition of an end-user.

4.6 Automatic Metric Layer

We evaluated whether common automatic quality estimation/fluency metrics detect the critical-yet-invisible errors identified by translator annotators. Three metrics were included:

- **BLEU** [30]: Reference-based n-gram overlap (reference translations prepared by professional translators)
- **COMET** [31]: Neural quality estimation model (reference-based)
- **Perplexity:** Language-model-based fluency measure [32]

For each translated segment, we computed metric scores and examined the correlation between metric scores and error presence. If an error is low-detectability by human judgment, it should not significantly lower the metric score compared to error-free segments—this would indicate the metric fails to capture the gap. We also used LLM-as-judge prompting, in which GPT-4o (with instruction "Evaluate this translation for fluency on a 1-5 scale") provided fluency-only ratings.

4.7 Analysis Plan

RQ1 is addressed through descriptive statistics: frequency and distribution of errors by type, severity, and detectability level, with breakdown by system and risk category. **RQ2** is addressed through a three-layer comparison: error detection rates (translators vs. auto metrics vs. monolingual panel), reported as percentages and compared using chi-square tests. **RQ3** is addressed qualitatively: representative cases from the critical-yet-low-detectability quadrant are examined in detail, illustrating the textual features that produce fluency and the source-text anchoring that enables translator interception.

All quantitative analyses were conducted in R (version 4.3.2) with significance set at $\alpha = 0.05$.

5. Results

This section presents the empirical findings from the three-layer triangulation design described in Section 4. The results are organised to answer the three research questions: the types and distribution of safety-critical errors (5.1), the empirical distribution within the severity×detectability matrix (5.2), and the comparative detection performance across translators, automatic metrics, and monolingual users (5.3). Illustrative error cases are provided in narrative form within 5.3.

5.1 Distribution of Safety-Critical Errors

The two translator annotators identified 315 errors across the 180 translated segments (60 source texts × 3 GenAI systems). Inter-rater agreement was high for error presence (Cohen's $\kappa = 0.84$) and for severity assignment (weighted $\kappa = 0.79$), confirming that the annotation protocol described in Section 4.4 produced reliable judgments. Among these errors, 86 cases (27.3%) were rated as critical – meaning that if a reader followed the translated instruction, physical harm could result. Table 3 shows the full adapted MQM typology with counts for each error subcategory.

Table 3: MQM-Adapted Error Typology for First-Aid Translation

Category	Subcategory	Critical	Major	Minor	Total
Accuracy	Mistranslation	47	81	29	157
Accuracy	Addition	12	24	0	36
Accuracy	Omission	18	34	0	52
Accuracy	Under-translation	0	28	0	28
Terminology	Term error	9	27	0	36
Total		86	194	29	309

Note: Six additional errors involved fluency only and were excluded from subsequent analysis, leaving 309 codable errors.

The critical errors were not evenly spread across content types. Errors involving negation or prohibition accounted for the largest share (39.5%, n=34). Dose and quantity errors followed (23.3%, n=20), then action sequence or temporal order (20.9%, n=18), and anatomical orientation or direction (16.3%, n=14). This distribution confirms that the purposeful sampling strategy described in Section 4.2 – which targeted these four high-risk language categories – succeeded in capturing the error types most likely to cause harm.

Across the three GenAI systems, the number of critical errors was similar: GPT-4o produced 31, DeepL 27, and Google Translate 28. A chi-square test showed no statistically significant difference across systems ($\chi^2 = 0.43$, $df = 2$, $p = 0.81$). This finding indicates that the fluency-detectability gap is not a peculiarity of any single translation engine, but rather a broader vulnerability shared by current GenAI translation architectures.

5.2 The Severity×Detectability Matrix

Following the theoretical framework established in Section 3.4, each error was rated for detectability – that is, whether a monolingual target-language reader would notice the error based on the target text alone, without access to the source. Errors were classified as either high detectability (the target sentence contains a grammatical oddity, semantic implausibility, or internal contradiction) or low detectability (the sentence is fluent, coherent, and contextually plausible). Two independent judges performed this rating, with substantial agreement ($\kappa = 0.83$).

Table 4 presents the empirical distribution of errors across the severity×detectability matrix.

Table 4: Severity×Detectability Matrix: Error Counts

	Minor	Major	Critical	Total
High Detectability	24	49	12	85
Low Detectability	5	45	74	124
Total	29	94	86	209

Note: 100 errors could not be rated for detectability (e.g., omissions with no target text to judge) and are excluded from this table.

The most important cell in this matrix is the **critical × low detectability** quadrant, which contains 74 errors. This figure represents 86.0% of all critical errors and 35.4% of the codable errors. In other words, more than eight out of every ten safety-critical errors were invisible to anyone reading only the Chinese translation. These 74 cases are the empirical substance of the fluency-detectability gap proposed in Section 1 and formalised in Section 3.

Qualitatively, these 74 errors share three surface features. First, they are grammatically well-formed – no missing particles, no wrong word order. Second, they use vocabulary that is appropriate for first-aid instructions; nothing seems odd or out of place. Third, each sentence is internally coherent – it does not contradict itself. Taken together, these features give the reader no reason to doubt the text. The only way to detect the error is to compare the translation against the English source, which is precisely what the professional translators did.

5.3 Detection Performance Across the Three Layers

The core claim of this study – that translators intercept what other agents miss – was tested by comparing detection rates for the 74 critical-yet-invisible errors across the three layers described in Section 4.

Translator layer. Working with source-text reference, the two annotators identified all 74 errors. Their detection rate was 100%.

Automatic metric layer. Four metrics were examined: BLEU, COMET, perplexity, and a fluency-only LLM-as-judge prompt. For each metric, a threshold was set at 1.5 standard deviations below the mean score (or above the mean for perplexity), and any error-containing segment that fell below that threshold was counted as "detected". Across the four metrics, the average detection rate was 18.9%. The individual rates were: BLEU 10.8%, COMET 16.2%, perplexity 13.5%, and LLM-as-judge 8.1%. These figures show that fluency-oriented metrics rarely penalise errors that are embedded in otherwise natural-sounding text.

Monolingual user panel. Thirty Chinese-speaking participants, who had no professional translation training and limited English proficiency, read the translated texts without access to the source. They were asked to underline any part that seemed unclear or suspicious. A critical-yet-invisible error was counted as "detected" only if at least ten participants flagged it. By this conservative standard, the panel detected only 9 of the 74 errors – a detection rate of 12.2%. Moreover, the participants' confidence ratings for the translations containing these errors (mean = 4.2 out of 5) were almost identical to their ratings for error-free translations (mean = 4.3). This suggests that fluency not only fails to alert users but actively builds a false sense of security.

The difference in detection rates across the three layers is statistically significant ($\chi^2 = 89.7$, $df = 2$, $p < 0.001$). These results provide strong empirical support for the argument advanced in Section 3.3: the professional translator, anchored to the source text, is the only reliable safety firewall against the most dangerous class of errors.

Illustrative cases (in place of a separate table). Several recurring error patterns appeared among the 74 critical-yet-invisible cases. One pattern was negation reversal – a source sentence saying "Do not induce vomiting" was rendered as a fluent Chinese command to "induce vomiting". Another pattern was dose distortion – "0.5 to 1 mg" became "1 to 5 mg" in the translation, with the numerical range expanded by a factor of five to ten. A third pattern involved sequence changes – "apply pressure before checking the wound" was translated in reverse order, producing a procedure that could worsen bleeding. A fourth pattern concerned direction – "tilt the head back slightly" was rendered as "tilt the head forward", which is dangerous in cases of suspected spinal injury. In every case, the Chinese sentence was grammatically sound and contextually plausible; the error was invisible to the monolingual reader and undetected by the automatic metrics. Only the source-anchored translator caught the discrepancy.

6. Discussion

This section interprets the results in light of the theoretical framework established in Sections 1–3, and considers the implications for translation quality assessment, professional practice, and public accountability.

6.1 Theoretical Contributions

The empirical findings give concrete meaning to the fluency-detectability gap as a theoretical construct. Drawing on Reiss and Vermeer's (2013) [10] skopos theory, we can say that the 74 critical-yet-invisible errors achieve textual coherence – they read fluently – but they systematically undermine the operative skopos of the first-aid text, which is to enable correct life-saving actions. This divergence between surface fluency and communicative purpose is not merely a descriptive curiosity; it is a structural feature of GenAI translation that fluency-oriented training objectives inadvertently promote. A translation that changes "do not" to "do" is no longer serving its intended function, yet it appears to be a successful translation if one judges by fluency alone.

The results also lend empirical weight to Nord's (2018) [1] loyalty principle. The professional translators in our study did not simply polish the GenAI output; they intervened to protect the end-user from harm, and they did so by maintaining a dual orientation – towards the source author's intention and the target reader's safety. This reframes translator verification as a safety intervention rather than a linguistic afterthought. It is not about making the translation sound better; it is about preventing harm that would otherwise go unnoticed.

Methodologically, the severity×detectability matrix (Table 2) adds a new dimension to the MQM framework. MQM already includes severity as a core axis. By adding detectability, we show that severity alone is an incomplete measure of real-world risk. A critical error that is easy to spot can be corrected by the user or by someone assisting them. A critical error that is invisible is more dangerous precisely because it is not corrected. The matrix therefore offers a more nuanced tool for safety-critical translation quality assurance.

6.2 Practical Implications

The three-layer comparison carries a clear practical message. Automatic quality metrics and monolingual users both fail to intercept the most dangerous errors – at rates below 20% and around 12%, respectively. Clinicians, regardless of their bilingual proficiency, are not trained to detect fluent negations, dose distortions, or sequence reversals embedded in grammatically normal text. The only agent who consistently catches these errors is the professional translator working with source-text reference. Therefore, any organisation that distributes GenAI-translated first-aid materials to LEP populations should institute mandatory verification by a qualified translator, and that verification should be documented.

This conclusion challenges the common assumption that fluency is a reasonable proxy for quality. Our user panel data show that fluent errors not only escape detection but also sustain high user confidence. Fluency, in this context, is not a safeguard – it is a mask. Public health agencies and emergency services should therefore revise their translation protocols to require a detectability audit, not merely a BLEU or COMET score. The verification record should include the translator's credentials and the date of review, establishing a clear chain of accountability. Furthermore, this paradigm shift necessitates a proactive response in translation education, training future professionals not merely as bilingual producers but as essential risk auditors [33].

6.3 Relationship to Previous Work

The present study complements and extends prior research in three ways. First, it builds on Al Sharou and Specia's (2022) [15] taxonomy of critical MT errors by showing that criticality and detectability are independent dimensions – an error can be critical yet invisible, and that combination is the one that matters most for safety. Second, it contributes to the MT literacy literature [17, 19] by demonstrating that end-user overconfidence is not a general cognitive failing but a specific vulnerability in high-stress, self-service scenarios where the reader has no independent check on the translation. Third, it distinguishes first-aid translation from the discharge-instruction domain studied by Khoong et al. (2019) [4] and Kong et al. (2025) [5]. Unlike discharge instructions, first-aid texts are consumed without a clinician present, so the errors that matter are not those that a doctor might catch, but those that the lay reader will act upon – and our results show that these are precisely the errors that are hardest to see.

7. Conclusion, Limitations, and Future Research

7.1 Conclusion

This study has proposed and empirically substantiated the fluency-detectability gap – a systematic class of safety-critical translation errors that are highly harmful yet invisible to monolingual users and automatic metrics because they are embedded in fluent, coherent, and plausible target-language text. Analysing English-to-Chinese GenAI translations of public first-aid materials, we found that 86% of critical errors fell into the low-detectability category. Automatic metrics detected fewer than 20% of these errors, and monolingual users detected only about 12%. In contrast, source-text-anchored professional translators caught every single one.

Theoretically, the gap is best understood as a divergence between textual coherence and operative skopos – a distinction that Reiss and Vermeer's (2013) [10] functionalist framework makes visible. Methodologically, we have offered the severity×detectability matrix as a practical tool that augments MQM for safety-critical contexts. Practically, we argue that fluency is not safety, and that institutions distributing GenAI-translated first-aid guidance must implement mandatory, documented translator verification. The translator, in this role, is not a post-editor but a human safety firewall.

7.2 Limitations

Four limitations should be acknowledged. First, the primary language pair in this study is English-to-Chinese. While Chinese is widely used among LEP populations in English-speaking countries, the findings may not generalise to other language pairs, particularly those with very different grammatical structures or writing systems. Second, GenAI outputs are non-deterministic – the same system may produce different translations on different occasions. We controlled for this by fixing model versions and recording all parameters, but the results are still tied to the specific outputs we analysed. Third, our sample was purposively selected to cover high-risk content (negation, dosage, sequence, orientation). This enhances relevance for safety but limits coverage of low-risk content. Fourth, detectability ratings, although supported by high inter-rater agreement and a behavioural panel, remain judgments about reader plausibility; they are not direct measures of what an actual reader in an emergency would notice.

7.3 Future Research

Future work should expand the investigation along several lines. Cross-linguistic studies with structurally different target languages – such as Arabic, Burmese, or Spanish – would test whether the gap widens or narrows depending on linguistic distance from English. Behavioural experiments, in which participants act on translated instructions in simulated emergency scenarios, would provide a more direct measure of harm than detection rates alone. Intervention studies could compare different verification protocols – single translator, double review, or translator-clinician joint review – to identify the most efficient safety configuration. Longitudinal replications, as GenAI models continue to evolve, would track whether the gap persists or diminishes with architectural improvements. Finally, applying the severity×detectability matrix to other safety-critical text types – such as medication labels, medical device instructions, or emergency warning systems – would test its generalisability as a quality-control instrument beyond first-aid translation.

References

[1] Nord, C. (2018). Translating as a purposeful activity: Functionalist approaches explained (2nd ed.). Routledge.

[2] Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745.

[3] Wang, L., & Bu, H. (2022). Emergency language services and emergency translation: Current situation and reflections. Foreign Languages in China, 19(4), 76–83. (in Chinese).

[4] Khoong, E. C., Steinbrook, E., Brown, C., & Fernandez, A. (2019). Assessing the use of Google Translate for Spanish and Chinese translations of emergency department discharge instructions. JAMA Internal Medicine, 179(4), 580–582.

[5] Kong, M., Fernandez, A., Bains, J., & Khoong, E. C. (2025). Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: A comparative study. Journal of General Internal Medicine, 40(2), 312–320.

[6] American Red Cross. (2023). First aid/CPR/AED participant's manual. American Red Cross.

[7] World Health Organization. (2021). Basic emergency care: Approach to the acutely ill and injured. World Health Organization.

[8] American Heart Association. (2020). CPR and ECC guidelines. American Heart Association.

[9] St. John Ambulance. (2022). First aid reference guide. St. John Ambulance.

[10] Reiss, K., & Vermeer, H. J. (2013). Towards a general theory of translational action: Skopos theory explained (C. Nord, Trans.). Routledge. (Original work published 1984)

[11] Rodriguez, J., Cohen, R., & Lagu, T. (2021). Machine translation in clinical settings: A review of risks and practice patterns. Journal of Patient Safety, 17(5), 402–408.

[12] Specia, L., Paetzold, G. H., & Scarton, C. (2018). Multi-level translation quality prediction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3155–3165). Association for Computational Linguistics.

[13] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

[14] Raunak, V., Kumar, V., & Khadilkar, H. (2021). A study of hallucinations in neural machine translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (pp. 23–33). Association for Computational Linguistics.

[15] Al Sharou, K., & Specia, L. (2022). A taxonomy and study of critical errors in machine translation. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation (pp. 171–180). European Association for Machine Translation.

[16] Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics, Vol. 3: Speech acts (pp. 41–58). Academic Press.

[17] Vieira, L. N., O'Hagan, M., & O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: A critical review of the literature on medical and legal use cases. Information, Communication & Society, 24(11), 1515–1532.

[18] Castilho, S., Moorkens, J., Gaspari, F., Doherty, S., & Way, A. (2018). Is machine translation usable? A review of the literature. Translation Spaces, 7(1), 16–47.

[19] Bowker, L., & Ciro, J. B. (2019). Machine translation and global research: Towards a machine translation literacy. Emerald Publishing.

[20] Flores, G., Abreu, M., & Barrueco, S. (2022). Bilingual staff in healthcare: A qualitative study of translation practices. Journal of Immigrant and Minority Health, 24(3), 642–650.

[21] Reiss, K. (1976). Texttyp und Übersetzungsmethode: Der operative Text. Scriptor.

[22] Zhang, M. (2005). Function plus loyalty: An introduction to Christiane Nord's functionalist translation theory. Journal of Foreign Languages, (1), 60–65. (in Chinese).

[23] Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. Traduimática, 12, 455–463.

[24] Mariana, V., Cox, T., & Melby, A. (2015). The Multidimensional Quality Metrics (MQM) framework: A new framework for translation quality assessment. The Journal of Specialised Translation, 23, 137–161.

[25] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

[26] Cui, Q. (2021). Translation technology and human–machine coupling in the age of artificial intelligence. Chinese Science & Technology Translators Journal, 34(1), 31–34, 40. (in Chinese).

[27] Li, M., & Zhu, X. (2020). Research on machine translation quality assessment: Review and prospects. Chinese Translators Journal, 41(4), 88–95. (in Chinese).

[28] Hu, K., & Li, Y. (2016). Corpus-based translation studies: Current situation and prospects. Foreign Language Teaching and Research, 48(2), 286–296. (in Chinese).

[29] U.S. Census Bureau. (2021). American Community Survey: Language use in the United States. U.S. Government Publishing Office.

[30] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). Association for Computational Linguistics.

[31] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 2685–2702). Association for Computational Linguistics.

[32] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report.

[33] Zhang, Z., & Wang, Y. (2023). Translation education in the age of generative artificial intelligence: Challenges and responses. Foreign Language World, (3), 2–9. (in Chinese).

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights