Journal of Scientific and Human Dimensions
مجلة الأبعاد العلمية والإنسانية
ISSN: 3106-1702
Volume 2, Issue 3, 2026 Pages: 160-171
Journal homepage: https://jshd.com.ly/index.php/jshd/en

Beyond Translation Quality: A Risk-Based Framework for Monolingual Reliance on AI Translation

Dr. Hussin Mahmoud Shneb *
Department English Language, Faculty of Arts, Elmergib University, Khoms, Libya

ما وراء جودة الترجمة: إطار قائم على المخاطر لاعتماد المستخدمين أحاديي اللغة على الترجمة بالذكاء الاصطناعي

د. حسين محمود شنيب *
قسم اللغة الإنجليزية، كلية الآداب، جامعة المرقب، الخمس، ليبيا

*Corresponding author: hmshneb@elmergib.edu.ly

Received: May 26, 2026	Accepted: June 25, 2026	Published: July 23, 2026

Abstract:
AI translation has improved to the point where many monolingual professionals now use it directly in their work. Healthcare professionals, lawyers, teachers, businesses, and public organisations increasingly rely on AI-generated translations without involving a bilingual reviewer. Most research has focused on one question. How accurate are these translations? To answer this, researchers use evaluation methods such as Multidimensional Quality Metrics (MQM), Direct Assessment, and the Crosslingual Optimized Metric for Evaluation of Translation (COMET). However, these methods do not answer the question that matters most in practice: Can this translation be relied on safely, for this purpose, in this situation? This paper argues that translation quality and translation dependability are different concepts. Quality evaluates how well a translation has been produced, whereas dependability considers whether it is safe to rely on that translation when making a real decision. Addressing this problem requires a framework that supports decisions about using AI translations rather than simply measuring system performance. Drawing on Pym's (2015) theory of translation risk and Lee and See's (2004) theory of trust in automation, this paper proposes a four-tier framework supported by a decision tree and a practical checklist. The framework is illustrated across eight professional domains to show how it can guide decisions about the safe use of AI translation. Finally, the paper compares the framework with existing translation evaluation methods and outlines directions for future empirical validation.

Keywords: AI Translation; Translation Dependability; Translation Quality; Risk-Based Framework; Trust In Automation; Monolingual Reliance**.**

الملخص
شهدت الترجمة المعتمدة على الذكاء الاصطناعي تطورًا ملحوظًا، حتى أصبح كثير من المهنيين أحاديي اللغة يعتمدون عليها مباشرة في أداء أعمالهم. ففي مجالات مثل الرعاية الصحية، والقانون، والتعليم، والأعمال، والإدارات العامة، باتت الترجمات التي ينتجها الذكاء الاصطناعي تُستخدم دون مراجعة من شخص يتقن اللغتين. وقد انصب اهتمام معظم الدراسات السابقة على تقييم دقة هذه الترجمات وذلك بالاعتماد على أدوات مثل مقاييس الجودة متعددة الأبعاد (MQM)، والتقييم المباشر (Direct Assessment)، والمقياس المحسَّن متعدد اللغات لتقييم الترجمة (COMET). إلا أن هذه الأدوات لا تجيب عن السؤال الأكثر أهمية في الممارسة العملية. هل يمكن الاعتماد على هذه الترجمة بأمان لهذا الغرض وفي هذا السياق؟ تنطلق هذه الدراسة من أن جودة الترجمة واعتمادية الترجمة مفهومان مختلفان؛ فالجودة تقيس مدى جودة الترجمة من الناحية اللغوية، بينما تشير الاعتمادية إلى مدى أمان الاعتماد عليها عند اتخاذ قرار أو تنفيذ إجراء في الواقع. ومن ثم،

فإن معالجة هذه الإشكالية تتطلب إطارًا يساعد المستخدمين والمؤسسات على اتخاذ قرارات مدروسة بشأن استخدام الترجمة بالذكاء الاصطناعي، بدلاً من الاقتصار على تقييم أداء أنظمة الترجمة. واستنادًا إلى نظرية بيم (Pym, 2015) في أخطار الترجمة، ونظرية لي وسي (Lee & See, 2004) في الثقة بالأتمتة، تقترح هذه الدراسة إطارًا مفاهيميًا يتكون من أربعة مستويات، مدعومًا بشجرة قرار وقائمة مراجعة عملية. وقد جرى توضيح تطبيق هذا الإطار من خلال ثماني مجالات مهنية مختلفة، لبيان كيفية الاستفادة منه في توجيه القرارات المتعلقة بالاستخدام الآمن للترجمة بالذكاء الاصطناعي. كما تقارن الدراسة هذا الإطار بأساليب تقييم الترجمة المتداولة، وتطرح تصورًا للبحوث المستقبلية اللازمة للتحقق من صلاحيته من خلال الدراسات التطبيقية.

الكلمات المفتاحية: الترجمة بالذكاء الاصطناعي؛ اعتمادية الترجمة؛ جودة الترجمة؛ الإطار القائم على المخاطر؛ الثقة بالأتمتة؛ اعتماد المستخدمين أحاديي اللغة.

Introduction

Machine translation used to be something professional translators worked with, not something everyone else used directly. That has changed. Large language models now produce translations that are fluent enough, and often accurate enough, that people with no knowledge of the target language paste a message into a chat window and send whatever comes out. Recent evaluations show that general-purpose large language models (LLMs) can match or even beat dedicated translation engines for high-resource language pairs, though their performance is far less reliable for low-resource languages and specialised domains (Hendy et al., 2023; Jiao et al., 2023). The existing research response to this shift has been to measure quality more precisely, better metrics, larger human evaluations, finer error taxonomies. What that response has not done is answer the question a monolingual professional is actually asking in the moment of use, which is not "how good is this system in general" but "can I act on this translation, right now, for this decision." No existing evaluation tool was built to answer that question, because all of them describe system behaviour averaged across many sentences, not the safety of one specific use. Translation quality tells us whether a translation is good; translation dependability tells us whether it is safe to rely on. This paper contributes a framework intended to address that gap, not a better-quality metric, but a use-specific decision layer that sits on top of whatever quality the underlying system provides. The framework proposed in this paper is conceptual and has not yet been empirically validated. The tier classifications presented in the paper are therefore intended to illustrate how the framework can be applied rather than to represent empirically established categories. The paper proceeds as follows. Section 2 defines the key terms used throughout. Section 3 reviews three relevant areas of literature: AI translation evaluation, trust in automation, and translation risk. Section 4 examines the problem of monolingual reliance, while Section 5 explains why dependability needs to be considered beyond accuracy. Section 6 presents the proposed risk-based framework, including the four tiers, decision tree, and practical checklist. Section 7 compares the framework with existing evaluation approaches, and Section 8 illustrates its application across eight professional domains. Section 9 discusses implementation, empirical validation, and implications. Sections 10–12 address ethical considerations, low-resource languages, and limitations, before Section 13 concludes the paper.

2. Glossary of Key Concepts

- Quality: The degree to which a translation accurately conveys the meaning, register, and fluency of the source text. Translation quality is typically evaluated using established human and automatic evaluation methods, including Multidimensional Quality Metrics (MQM), Direct Assessment (DA), and the Crosslingual Optimized Metric for Evaluation of Translation (COMET).
- Dependability: Whether a specific translation can be safely relied upon by a specific user, for a specific purpose, in a specific context. Dependability requires translation quality but is not reducible to it.
- Trust: A user's willingness to rely on a system's output before that output can be independently verified.
- Reliance: The behavioural act of acting on a system's output, whether that reliance is warranted.
- Verification: A human or automated process used to check whether a translation is sufficiently accurate before it is acted upon. The method used should be appropriate and sufficiently reliable for the specific translation task. Examples include qualified bilingual review, back-translation, and terminology validation, although these methods may provide different levels of assurance.
- Monolingual reliance: Reliance on a translation by a user who cannot independently understand or verify the target-language output. In this paper, the term does not necessarily mean that the user speaks only one language. Rather, it refers to situations in which the user does not understand the target language well enough to judge whether the translation is accurate.

3. Related Work

3.1 AI Translation Evaluation

Modern machine translation systems, including the large language models (LLMs) now widely used for everyday translation, are based on the transformer architecture (Vaswani et al., 2017). Neural systems solved many of the fluency problems of older statistical models but introduced their own weaknesses. Koehn and Knowles (2017) identify several common weaknesses in machine translation. These include domain mismatch, limited training data, rare words, and long sentences. Many of these problems occur in the professional and technical texts that monolingual users often need to translate. LLMs are trained as general-purpose text generators rather than dedicated translation systems. As a result, their translation ability emerges from broader language training. They perform well on common languages and everyday texts but are less reliable for less-resourced languages and technical domains (Hendy et al., 2023; Jiao et al., 2023). Evaluation methods have also advanced. MQM asks trained reviewers to identify and classify specific translation errors (Lommel et al., 2014). DA asks crowd raters for a numeric score, shown to be consistent at scale (Graham et al., 2017). COMET predicts human judgments automatically (Rei et al., 2020), and xCOMET, an extended version of COMET, adds span-level error localisation (Guerreiro et al., 2024). Each represents a genuine advance in measuring quality. None of them, however, was designed to tell an individual user whether the one translation in front of them is safe to act on. All report performance aggregated over a batch of sentences or systems, not a verdict on a single use-event. This is the first of the gaps the framework proposed in this paper is intended to address.

3.2 Trust in Automation

This research introduced important concepts such as calibration and appropriate reliance. However, it was developed to explain trust in automation in general rather than AI translation specifically. As a result, it cannot distinguish between translations that are safe to rely on and those that pose a risk. A translation is fundamentally different from a warning light in a factory or an alert in an aircraft cockpit. Although this body of research explains how people develop trust in AI, it does not help users decide whether a specific AI-generated translation is safe to rely on in a particular context. More recent studies have extended this work but have not fully addressed the problem. Li et al. (2024) integrated research on interpersonal trust, trust in automation, and trust in AI into a unified framework. Küper and Krämer (2025) found that personality traits, such as an individual's natural tendency to trust others, influence reliance on AI. Zhang et al. (2024) also showed that mental workload and the way AI explains its outputs affect whether people rely on it appropriately. Although these studies deepen our understanding of trust in AI, they do not explain what makes an AI-generated translation safe or risky in a real translation context. This is the second gap that the present framework aims to address.

3.3 Translation Risk

Translation scholars have approached risk from a different perspective. Pym (2015) argued that translation is a form of risk management. According to him, translators make decisions by balancing three types of risk. The first is credibility risk, where a translation may appear inaccurate or unreliable, even if it is correct. The second is uncertainty risk, which arises when several translation options are possible and the translator must choose one. The third is communicative risk, where the intended meaning may not be understood by the target audience. This way of thinking is much closer to the challenges of translation than the general literature on trust in automation because it focuses on risks that are unique to translation. However, Pym's framework was developed for professional human translators. It assumes that the translator understands both the source and target languages and can judge the risks before making a translation decision. That assumption does not apply to the situation examined in this paper. Here, the person using the translation cannot understand the target language well enough to independently verify the translation. At the same time, there is no human translator making those decisions on the user's behalf. As a result, even the translation framework that comes closest to addressing this problem still assumes that someone who understands both languages is involved. In monolingual use of AI translation, that person is no longer part of the process. This is the third gap that the present framework seeks to address.

3.4 Integrating the Two Theories

On their own, neither of these theories fully addresses the problem. Pym (2015) explains the different types of risk that can arise in a translation, but he assumes that someone who understands both languages can recognise and manage those risks. Lee and See (2004), on the other hand, explain how people decide whether to rely on an automated system, but their framework is not specific to translation and does not explain what makes a translation risky. Rather than treating these theories as competing explanations, this paper brings them together. Pym's framework explains what the risks are in translation, while Lee and See's framework explains who must decide whether it is safe to rely on the translation. Together, they provide the conceptual foundation for the framework

proposed in Section 6. Table 1 shows how the two theories complement each other and why both are needed to understand monolingual reliance on AI translation.

Table 1. Comparison of the Two Theoretical Foundations of the Proposed Framework

Theoretical question	Pym (2015) — Translation Risk	Lee & See (2004) — Trust in Automation
What can go wrong?	Pym explains the main types of risk involved in translation. A translation may appear unreliable (credibility risk), require choosing one option when several are possible (uncertainty risk), or fail to communicate the intended meaning (communicative risk).	Lee and See do not focus on translation itself. Instead, they define failure as a mismatch between a system's actual reliability and the user's level of trust in it.
Who decides how much caution is needed?	Pym's framework assumes a professional bilingual translator who can read both languages and make informed decisions about risk while producing the translation.	Lee and See focus on the user, who decides whether to rely on an automated system even though they cannot directly observe how it works.
What does this paper add?	This paper extends Pym's framework to a new situation in which the user is monolingual and no human translator is involved in judging translation risks.	It also extends Lee and See's concept of calibrated reliance to a situation where users cannot verify the translation themselves because they do not understand the target language.

4. The Problem of Monolingual Reliance

The challenge for a monolingual user is simple to understand but difficult to solve. In the context of this paper, a monolingual user is someone who cannot understand the target language well enough to independently verify the AI-generated translation. The user may speak other languages, but they cannot judge whether the translation in the target language is accurate or appropriate. In traditional translation, this role is performed by someone who knows both languages. When AI is used by a monolingual user, that safeguard no longer exists. Modern machine translation systems and large language models make this problem even more challenging. They often produce translations that sound fluent and natural but still contain important errors. Research has shown that people are more likely to notice awkward or unnatural language than errors in meaning. However, errors in meaning may have more serious consequences because fluent language can make such errors difficult for users to detect (Martindale & Carpuat, 2018). This reflects the problem described by Lee and See (2004), users judge the quality of a system by the wrong signal. For monolingual users, fluency becomes the main reason for trusting a translation, even though fluent language does not guarantee that the translation is accurate. This is not only a theoretical concern. Khoong et al. (2019) examined Google Translate for hospital discharge instructions. Most translations were accurate, with about 92% of Spanish sentences and 81% of Chinese sentences translated correctly. However, (2%) of the Spanish sentence translations and (8%) of the Chinese sentence translations had the potential to cause clinically significant harm (Khoong et al., 2019). For a healthcare professional who cannot understand the target language, such errors may be difficult to detect independently before the translated instructions are used. Researchers have also tested tools designed to help users judge translation quality, but the results have been mixed. Zouhar et al. (2021) found that back-translation made users feel more confident without making their decisions more accurate. Mehandru et al. (2023) found that back-translation helped physicians detect more clinically important errors than confidence scores alone. Neither approach identified every dangerous error. More recent research has reached a similar conclusion. Carpuat et al. (2025) argue that helping users make reliable independent judgments about machine translation remains a major challenge. Xiao et al. (2025) argue that providing more information does not always improve decisions. It can instead increase users' mental workload.

5. Why Dependability Matters Beyond Accuracy

This paper focuses on dependability rather than accuracy because the two answer different questions. Accuracy tells us whether a translation correctly conveys the intended meaning. Dependability goes a step further by asking whether a particular translation can be safely relied on in a specific situation. Even a highly accurate system will sometimes make mistakes. A system that is correct 99% of the time will still produce errors. In low-stakes situations, those errors may have little impact. However, in legal, medical, or immigration settings, a single mistake can have serious consequences. An incorrect medication instruction, a mistranslated contract, or an error in an asylum application may cause harm, regardless of how accurate the system is on average. This highlights an important difference between accuracy and dependability. Accuracy measures performance across many translations, whereas dependability concerns a single decision made by one user at one moment. It is that

individual decision not the system's overall accuracy that determines whether relying on the translation is safe. For this reason, a highly accurate AI system may still fall into Tier 0 if the consequences of an error are unacceptable or the translation cannot be reliably verified. The framework therefore asks a different question from most translation evaluation methods. Rather than asking, "How accurate is this system overall?" it asks, "Is it safe to rely on this translation for this purpose, in this context, right now?" This shift represents the central contribution of the paper. It moves the discussion beyond improving translation quality alone and towards helping users make better decisions about when AI-generated translations can be relied upon safely.

6. A Risk-Based Dependability Framework

6.1 Two Dimensions

The framework is built around two operational questions. Stakes asks If this translation contains an error, how serious would the consequences be, and could they be reversed? This operationalises Pym's (2015) communicative risk. Verifiability asks whether an appropriate and sufficiently reliable method is available to detect an error before the translation is acted upon or causes harm. The required method may vary depending on the type of text, the context, and the level of risk. This operationalises the calibration problem described by Lee and See (2004). At its extreme, it also reflects Pym's (2015) concept of uncertainty risk. Without verification, no one, including the system itself, can know whether the translation is correct.

Translation Quality	**+**	**Use-Context Factors**	**→**	**Translation Dependability**
System-level, average performance across many sentences (MQM, DA, COMET, xCOMET)		Stakes of an error and whether the user can verify the output (this paper's framework)		Use-specific: safe for THIS person, THIS decision, NOW
Quality is necessary but not sufficient for dependability.				

Figure 1. Relationship between translation quality and translation dependability.

Figure 1 shows why high translation quality does not always mean a translation is safe to rely on. The same translation may be acceptable in one context but not in another. This depends on what is at stake and whether its accuracy can be verified. For this reason, translation quality should be distinguished from translation dependability. Translation quality is essential, but dependability also considers the context in which the translation is used. This distinction provides the conceptual foundation for the framework presented in Figure 1 and developed in the following sections.

6.2 Why Four Tiers?

Three questions may be raised about this framework. The first concerns the use of four tiers. Why four rather than three or six? Four tiers are proposed as a simple structure for representing the two key dimensions of the framework: stakes and verifiability. Using fewer tiers may make it more difficult to show how these two dimensions interact, while adding more tiers could make the framework unnecessarily complex. The four-tier structure therefore offers a practical balance between simplicity and explanatory power. However, as the framework has not yet been empirically validated, future research should examine whether four tiers provide the most appropriate structure or whether further refinement is needed. A second question is why the framework focuses on stakes and verifiability instead of other factors such as text length, subject area, or language pair. These factors are certainly important, but they influence translation decisions through stakes and verifiability rather than acting as separate dimensions. For example, a longer text is not inherently more risky. It becomes riskier because it increases the chance that an error will go unnoticed or because more important decisions may depend on it. Similarly, subject area and text type affect either the consequences of an error or the user's ability to verify the translation. Using two broad dimensions keeps the framework simple while still allowing these additional factors to be considered. Section 6.3 explains how text type and subject area are incorporated into the criteria for each tier. A third question is why existing risk scales, such as those used in safety engineering, were not adopted. These scales are designed to measure the severity of potential harm, but they were not developed for translation. More importantly, they do not consider whether a user is able to recognise an error before relying on the translation. For monolingual users, this is a critical issue. The consequences of an error depend on two factors. They depend on how serious the error is and whether the user can detect it before acting on the translation.

6.3 Operational Definitions and Classification Criteria

Table 2 gives concrete, checkable criteria for each tier, so that classification does not rest on intuition alone.

Table 2. Operational Definitions and Classification Criteria for the Four Dependability Tiers

Tier	When it applies	Can the translation be verified?	Recommended action
0 – Human translation required	An error here could be serious, or even impossible to undo, think legal, medical, financial, or safety consequences.	No. An appropriate and sufficiently reliable verification method is not available, or the available method does not provide enough assurance for this type of text and level of risk.	A qualified human translator must produce or approve the final text before use.
1 – Verified reliance	This tier also applies to high-stakes situations where translation errors could have serious consequences.	Yes. An appropriate and sufficiently reliable verification method is available for the specific task, such as qualified bilingual review or another suitable form of verification.	AI output may be used only after the verification step is completed and passed.
2 – Assisted reliance	An error here would be minor and fixable, and it would likely come to light naturally as the conversation or task continues.	A formal verification process is not necessary because the ongoing interaction allows misunderstandings to be identified and corrected naturally.	The AI translation can be used directly for reading or drafting, while remaining alert to signs of misunderstanding or confusion.
3 – Autonomous reliance	This tier applies when a translation error would have little or no meaningful consequence.	Verification is unnecessary because the effort required to check the translation would outweigh the potential impact of any error.	The AI-generated translation can be used without additional verification.

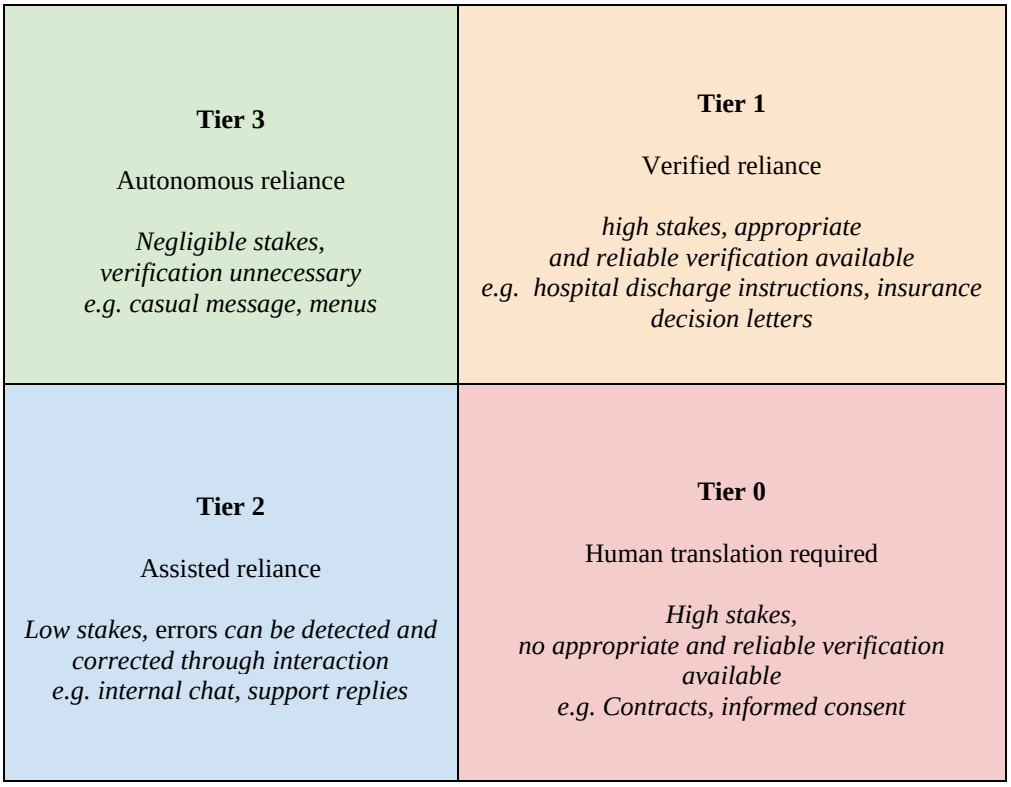

Figure 2. The risk-based dependability framework, plotted by stakes and verifiability.

Figure 2 Translation Dependability Matrix ***Reliable verification available or verification not required***

Low/negligible stakes

High stakes

6.4 Decision Tree

Figure 3 turns the criteria in Table 2 into a sequence of questions a non-specialist can work through for any given translation task, without needing to hold the full framework in mind at once.

Step 1 — If this translation contains an error, could the outcome be serious or irreversible?			
No (low stakes) Will an error likely be noticed and corrected in the next exchange?		**Yes (high stakes)** Is a bilingual check available (colleague, back-translation, terminology validation)?	
(Yes — noticed/corrected)	*(No/unsure)*	*(Yes — check available)*	*(No — no check available)*
Tier 3 Autonomous reliance	**Tier 2** Assisted reliance	**Tier 1** Verified reliance	**Tier 0** Human translation required

Figure 3. Decision tree for classifying a translation task by reliance tier.

6.5 Practical Checklist

Before relying on an AI translation for a professional purpose, ask:

- If this translation is wrong, can the error be undone? If not, treat it as high stakes.
- Is there an appropriate and reliable way to detect an error before it causes harm, such as review by a qualified bilingual person or another suitable verification method?
- Does the text involve a high-stakes area, such as legal, medical, immigration, financial, or safety-related communication? If so, classify it as Tier 1 when appropriate and reliable verification is available. If reliable verification is not available, classify it as Tier 0.
- Has a verification step been completed, not just made available, before the translation is used?
- If in doubt between two tiers, choose the more cautious one; the cost of an unnecessary check is almost always smaller than the cost of an undetected error.

7. Comparing the Framework to Existing Evaluation Approaches

The proposed framework complements existing translation evaluation approaches rather than replacing them. Methods such as MQM, DA, COMET, xCOMET, Quality Estimation (QE), and back-translation assess translation quality in different ways. The proposed framework addresses a different question: can a particular AI-generated translation be relied upon safely in its intended context, for a specific person and purpose?

Table 3. Comparison of Existing Translation Evaluation Approaches and the Proposed Dependability Framework

Approach	What it measures	What it cannot do	Relation to this framework
MQM (Lommel et al., 2014)	MQM provides a detailed evaluation of translation quality by identifying and classifying errors and rating their severity. The assessment is carried out by trained human evaluators across a set of translations.	It is designed for expert evaluation rather than real-time use. It cannot help a monolingual user decide whether the single translation in front of them is safe to rely on.	MQM can be useful for understanding the potential seriousness of translation errors, but it does not help users decide whether a specific AI-generated translation can be trusted at the moment.
DA (Graham et al., 2017)	DA measures overall translation quality by averaging ratings from many human evaluators across multiple translations.	The scores reflect the general performance of a translation system rather than the reliability of an individual translation.	DA provides evidence about how well a system performs on average, but it cannot determine whether one particular translation is suitable for a specific purpose.

COMET / xCOMET (Rei et al., 2020; Guerreiro et al., 2024)	COMET automatically predicts the quality of individual translations, while xCOMET can also identify where translation errors may occur.	These are predicted quality scores rather than confirmed evidence of correctness. A monolingual user still cannot verify whether the translation is actually accurate.	These tools may contribute to judgments about verifiability, but they cannot replace independent verification.
Quality Estimation (QE)	Quality Estimation provides a confidence score for a translation without requiring a reference translation.	Confidence does not always reflect accuracy. Zouhar et al. (2021) found that confidence scores increased users' trust without improving their ability to identify translation errors.	This supports the framework's argument that confidence alone is not enough. Reliable reliance requires verification, not simply a higher confidence score.
Back-translation	Back-translation checks a translation by translating it back into the original language and comparing the result with the source text.	Although it can help identify some errors, it is not completely reliable, requires additional time, and may not always be available (Mehandru et al., 2023).	Back-translation may contribute to Tier 1 verification when it is appropriate for the task and provides sufficient assurance. However, it should not automatically be treated as adequate verification for every high-stakes translation.

8. Illustrative Applications

To illustrate the framework's broader applicability, Table 4 applies it to eight representative professional translation scenarios spanning diverse domains.

Table 4. Applying the Risk-Based Dependability Framework Across Professional Domains

Domain	Example task	Stakes	Verifiability	Tier	Rationale
Healthcare	Hospital discharge instructions (Khoong et al., 2019)	High	Available through bilingual review or appropriate back-translation	**1**	Translation errors may cause patient harm. AI output may be used only after appropriate verification.
Legal	Cross-border vendor contract	High	Low	**0**	Errors may have serious and lasting legal consequences. A qualified human translator should produce or approve the final translation.
Immigration	Visa or asylum application narrative	High	Low	**0**	A mistranslation may affect a person's legal status or case. A qualified human translator should produce or approve the translation.
Police	Witness statement taken through AI translation	High	Low	**0**	Misrepresented testimony may affect legal outcomes and may be difficult to correct later. A qualified human translator or interpreter is required.
Business	Routine customer support reply	Low	High (dialogue continues)	**2**	Errors are usually minor and can be identified and corrected as the conversation continues. Formal verification is not normally required.

Education	Translated instructions for an overseas exchange student	Medium	Available through bilingual staff	**1**	Errors may have meaningful consequences, but bilingual staff can verify the translation before it is relied upon.
Insurance	Translated claims-denial letter for a policyholder	High	Available through bilingual review	**1**	Errors may affect appeal rights or important deadlines. AI output may be used only after appropriate verification.
Workplace communication	Informal message about a staff social event	Very low	Not necessary	**3**	An error would have little or no meaningful consequence. The AI-generated translation can be used without additional verification.

Note: The tier classifications presented in this table are illustrative applications of the proposed framework and have not yet been empirically validated.

Three clear patterns emerge from these examples. First, high-stakes areas such as healthcare, law, immigration, policing, and insurance fall into Tier 0 or Tier 1, regardless of how fluent the AI-generated translation appears. A translation that sounds natural is not necessarily safe to rely on. What matters is the potential consequence of an error and whether reliable verification is available. Second, verifiability distinguishes Tier 0 from Tier 1 in high-stakes situations. Legal, immigration, and police tasks in these illustrative scenarios fall into Tier 0 because appropriate and sufficiently reliable verification is assumed not to be available before the translation is used. Healthcare and insurance fall into Tier 1 because appropriate verification is available before the translation is used. Third, the business and workplace examples show how the framework applies to lower-stakes situations. Routine customer support falls into Tier 2 because misunderstandings can be identified and corrected as the conversation continues. An informal workplace message about a social event falls into Tier 3 because an error would have little or no meaningful consequence. Together, these examples show that the framework considers both the consequences of an error and the possibility of detecting it before harm occurs.

9. Implementation, Validation, and Implications

9.1 Practical Workflow

Figure 4 summarises how an organisation can operationalise the checklist and decision tree as a repeatable process rather than a one-off judgment call.

1	**2**	**3**	**4**	**5**	**6**
Identify the translation task	Assess stakes: reversible or not?	Assess verifiability: any check available?	Assign a tier (0–3)	Apply the tier's required safeguard	Proceed, or escalate to a human translator

9.2 Towards Empirical Validation

The framework proposed in this paper is conceptual and should be tested before it is used in practice. Future research could validate it in three stages. The first stage would examine whether different experts classify translation tasks in the same way. Professional translators, interpreters, and specialists in risk management could independently assign a range of realistic translation scenarios to the four tiers. Their classifications could then be compared to measure the level of agreement and assess the consistency of the framework. The second stage would investigate whether the framework helps users make better decisions about when to rely on AI-generated translations. Controlled experiments could test this directly. Participants using the framework could be compared with those who do not. Similar designs have already been used to evaluate back-translation and quality estimation (Mehandru et al., 2023; Zouhar et al., 2021). This would show whether the framework reduces inappropriate reliance on AI translations and helps users make better decisions. The third stage would test the framework in real organisations. Organisations using AI translation could classify their translation tasks according to the four tiers and follow the recommended safeguards. They could then monitor translation-related problems over time to see whether the framework reduces errors and supports safer use of AI translation. Together, these three stages would provide evidence of the framework's reliability, usefulness, and practical value. They would also show whether the framework can work effectively across different organisations and translation contexts.

9.3 Implications

The proposed framework has implications for AI developers, organisations, and translation studies. For AI developers, confidence scores should not be presented as proof that a translation can be trusted. A confidence score may provide useful information, but it cannot confirm that a particular translation is safe to rely on. In high-stakes settings, AI systems should provide practical ways to verify translations, such as back-translation or other suitable verification tools (Mehandru et al., 2023). Users should also be made aware of the limits of these tools. Previous research suggests that such tools may increase users' confidence without necessarily improving their ability to judge translation quality (Zouhar et al., 2021). This means that confidence alone should not be treated as evidence that an AI-generated translation is dependable. For organisations, decisions about when AI translation can be used should not be left to individual users alone. Different users may have different levels of experience with AI and may judge the same translation task differently. Organisations should therefore establish clear policies before AI translation is used. A tiered framework, such as the one proposed in this paper, could help organisations decide when AI translation can be used directly, when verification is needed, and when a qualified human translator should be involved. This would provide a more consistent approach to AI translation across an organisation and reduce the need for individual users to make difficult reliance decisions on their own. This is consistent with Lee and See's (2004) view that appropriate reliance on automation should be supported at the organisational level rather than depending only on individual judgement. For translation studies and translator education, the framework extends Pym's (2015) theory of translation risk to a different context. Pym's framework focuses on how professional bilingual translators manage risk while making translation decisions. The present framework considers situations in which a monolingual user relies directly on AI translation without a human translator being involved. This shift raises new questions about the role of professional translators in an increasingly AI-supported translation environment. As AI translation becomes more widely used, translators may increasingly be asked not only to produce translations but also to verify AI-generated output, assess translation risks, and advise organisations on when AI translation can be relied upon safely. Translator education may therefore need to give greater attention to these emerging roles alongside traditional translation skills.

10. Ethical Implications

The framework proposed in this paper should be used to support responsible decision making, not simply to reduce translation costs. There is a risk that organisations could classify too many tasks as Tier 2 or Tier 3 to justify greater use of AI translation. If this happens without carefully considering the level of risk, the purpose of the framework is undermined. The four tiers are intended to guide appropriate use of AI translation, not to encourage replacing human translators wherever possible. The framework also raises important questions about fairness. Many people who rely on translated information, including migrants, refugees, and individuals with limited English proficiency, are unable to verify whether a translation is correct. At the same time, they may be the people most affected when translation errors occur. The study by Khoong et al. (2019) on hospital discharge instructions illustrates how even a small number of translation errors can have serious consequences for patient safety. For this reason, organisations using this framework should view limited verifiability as a reason to apply greater caution, not as an opportunity to reduce costs or lower standards. The goal of the framework is to promote safer and more responsible use of AI translation, particularly in situations where the consequences of an error may be significant.

11. Low-Resource Languages

The framework is particularly relevant for languages and language varieties that are less well represented in machine translation systems and large language model training data, including some Arabic varieties and specialised Arabic domains (Hendy et al., 2023). As a result, translation quality may be less consistent and more difficult to predict. This does not change the two main questions of the framework; how serious would an error be and can it be reliably detected before it causes harm? However, when translation quality is less predictable, verification becomes even more important. The less certain the quality of the translation, the greater the need to identify errors before relying on the output. In practice, this means that professionals working with lower-resource language pairs should adopt a more cautious approach. Until the quality of AI translation for a particular language pair has been independently established, it is generally safer to place the task in a higher-risk tier than would be appropriate for a well-supported language pair. This approach helps reduce the risk of relying on translations whose quality has not yet been demonstrated consistently.

12. Limitations

This paper presents a conceptual framework that has not yet been tested empirically. The four tiers are intended to illustrate how the framework can be applied rather than to represent boundaries established through empirical evidence. As discussed in Section 9.2, future research is needed to evaluate and refine the framework. The

framework also assumes that the level of **stakes** and **verifiability** can be assessed before a translation is used. In practice, this may not always be possible. A translation that appears to involve only minor consequences may later prove to be much more important than expected. Likewise, a verification method that seems reliable, such as review by a bilingual colleague, may not always detect subtle translation errors. Finally, this paper focuses on written AI translation. It does not address spoken or simultaneous interpreting, where decisions must be made in real time and opportunities for verification are much more limited. These settings involve additional challenges that require separate investigation.

13. Conclusion

AI translation has reached a stage where many monolingual professionals use it directly, often without any bilingual review. This change has created a practical problem that existing translation research does not fully address. Most evaluation methods measure how well a system performs across many translations, but they do not answer the question that matters most in practice. Can this particular translation be relied on safely for this purpose, in this situation, by this user? This paper argues that answering this question requires looking beyond translation quality alone. It introduces the concept of translation dependability to focus on whether an AI-generated translation is safe to rely on in a real decision-making context. By bringing together Pym's (2015) theory of translation risk and Lee and See's (2004) theory of trust in automation, the paper proposes a four-tier framework based on two simple questions. How serious would the consequences of an error be and can the error be detected before it causes harm? These two questions provide a practical way for individuals and organisations to decide when AI translation can be used safely, when verification is needed, and when a qualified human translator remains essential. The value of this framework is not that it improves the accuracy of AI translation. Instead, it offers a different way of thinking about how AI translations should be used. As AI becomes part of everyday professional practice in healthcare, law, education, business, government, and many other fields, the challenge is no longer only to build better translation systems. It is also to help people make better decisions about when those systems can be relied upon. The framework presented here is intended as a starting point rather than a final solution. Future empirical research is needed to test its reliability, refine the boundaries between the four tiers, and evaluate its usefulness across different languages, organisations, and professional settings. Even so, it offers a practical foundation for future work by shifting attention from translation performance alone to the conditions under which AI-generated translations can be used responsibly. Ultimately, this paper argues that the future of AI translation should not be judged by accuracy alone. It should also be judged by how responsibly people can rely on AI-generated translations when making real decisions. As AI continues to reshape multilingual communication, knowing when not to rely on a translation may become just as important as producing a more accurate one.

References

[1] Carpuat, M., Asscher, O., Bali, K., Bentivogli, L., Blain, F., Bowker, L., Choudhury, M., Daumé III, H., Duh, K., Gao, G., Grissom II, A., Karpinska, M., Khoong, E. C., Lewis, W. D., Martins, A. F. T., Nurminen, M., Oard, D. W., Popovic, M., Simard, M., & Yvon, F. (2025). An interdisciplinary approach to human-centered machine translation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 22859–22879). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1164

[2] Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone? Natural Language Engineering, 23(1), 3–30.

[3] Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. T. (2024). xCOMET: Transparent machine translation evaluation through fine-grained error detection. Transactions of the Association for Computational Linguistics, 12, 979–995.

[4] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. arXiv. https://arxiv.org/abs/2302.09210

[5] Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. arXiv. https://arxiv.org/abs/2301.08745

[6] 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.

[7] Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation (pp. 28–39). Association for Computational Linguistics.

[8] Küper, A., & Krämer, N. (2025). Psychological traits and appropriate reliance: Factors shaping trust in AI. International Journal of Human-Computer Interaction, 41(7), 4115–4131.

[9] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.

[10] Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust. Frontiers in Psychology, 15, Article 1382693.

[11] Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica: Tecnologies de la Traducció, 12, 455–463.

[12] Martindale, M. J., & Carpuat, M. (2018). Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Vol. 1, pp. 13–25). Association for Machine Translation in the Americas.

[13] Mehandru, N., Agrawal, S., Xiao, Y., Gao, G., Khoong, E. C., Carpuat, M., & Salehi, N. (2023). Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 11633–11647). Association for Computational Linguistics.

[14] Pym, A. (2015). Translating as risk management. Journal of Pragmatics, 85, 67–80.

[15] 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.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[17] Xiao, Y., Hancock, C., Agrawal, S., Mehandru, N., Salehi, N., Carpuat, M., & Gao, G. (2025). Sustaining human agency, attending to its cost: An investigation into generative AI design for non-native speakers' language use. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (Article 345). Association for Computing Machinery.

[18] Zhang, Z. T., Ketenci Argın, S., Bilen, M. B., Urgun, D., Deniz, S. M., Liu, Y., & Hassib, M. (2024). Measuring the effect of mental workload and explanations on appropriate AI reliance using EEG. Behaviour & Information Technology. Advance online publication. https://doi.org/10.1080/0144929X.2024.2431055

[19] Zouhar, V., Novák, M., Žilinec, M., Bojar, O., Obregón, M., Hill, R. L., Blain, F., Fomicheva, M., Specia, L., & Yankovskaya, L. (2021). Backtranslation feedback improves user confidence in MT, not quality. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 151–161). Association for Computational Linguistics.