
Copyright © Authors, 2026

Article

AI-Mediated Translation and Epistemic Hierarchies: LLMs, Literary Language, and Translational Decision-Making in Joyce's *Ulysses*

Daniele Duscovich ^{1*}

¹ Alma Mater Europaea; daniele.duscovich@almamater.si

* Correspondence: daniele.duscovich@almamater.si

Abstract

Large language models (LLMs) are increasingly used to mediate linguistic and cultural exchange, yet their capacity to preserve linguistically and historically marked forms remains insufficiently understood. This study examines how AI-mediated translation negotiates linguistic difference and cultural expression through a comparative analysis of James Joyce's "Oxen of the Sun" from *Ulysses*. The analysis compares Enrico Terrinoni's professional Italian translation with translations generated by ChatGPT and Google Translate. An adaptation of the Multidimensional Quality Metrics (MQM) framework, developed specifically for the assessment of literary translation, is combined with qualitative close reading to examine accuracy, fluency, diachronic style, register, multilingualism, wordplay, and formal experimentation. The findings indicate a distinction between linguistic competence and strategic translational competence. While ChatGPT produces comparatively fluent and coherent Italian, it frequently normalises stylistically and historically marked forms, reducing the differentiation that structures Joyce's text. Google Translate exhibits more extensive difficulties in both semantic transfer and stylistic reconstruction, whereas Terrinoni's translation demonstrates more consistent strategic management of competing textual priorities. These findings suggest that LLM-mediated translation should be examined not only in terms of linguistic accuracy and fluency, but also in terms of how AI systems preserve, transform, or neutralise culturally and historically situated forms of expression. The study contributes to critical discussions of AI, language, and epistemic hierarchies by proposing literary translation as a site for examining the non-neutral mediation of linguistic and cultural difference.

Keywords: artificial intelligence; large language models; literary translation; language mediation; epistemic hierarchies

Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. LLMs and Literary Translation

The rapid development of large language models (LLMs) has significantly expanded the role of artificial intelligence in language mediation. Unlike earlier machine translation systems, which were primarily evaluated in terms of sentence-level adequacy and fluency, LLMs can process broader textual contexts and generate translations that are highly fluent, coherent, and context-sensitive. Recent research on document-level machine translation has shown that LLMs can effectively exploit information extending beyond the individual sentence, suggesting new possibilities for translation tasks in which meaning depends on wider textual contexts (Wang et al., 2023). These developments are particularly relevant to multilingual and internationally oriented contexts, where AI-mediated language practices increasingly shape how linguistic and cultural difference is encountered and interpreted.

Yet improvements in linguistic fluency do not necessarily correspond to an ability to preserve the cultural and epistemic dimensions of language. Literary texts provide a particularly revealing context in which to examine this problem. Literary translation cannot be reduced to the accurate transfer of propositional meaning. Literary texts frequently exploit stylistic variation, register, historical language, ambiguity, multilingualism, rhythm, sound patterning, and wordplay as constitutive elements of both form and meaning. In such contexts, translation involves not only the transfer of linguistic content but also the interpretation and reconstruction of textual effects.

This issue acquires a broader significance when translation is understood as a form of AI-mediated language and cultural representation. Choices concerning which linguistic features are retained, transformed, or normalised can affect the visibility of particular historical, cultural, and stylistic forms of expression. AI systems are therefore not necessarily neutral mediators between languages: their outputs may privilege certain forms of linguistic expression while reducing the visibility of others. This question is particularly relevant to international education, where multilingual communication and the mediation of cultural difference are central to cross-border academic and social interaction. The present study approaches literary translation as a controlled case through which these processes of AI-mediated linguistic selection can be examined.

MT research has increasingly recognised literary translation as a particularly demanding domain. Toral and Way (2018), in a comparative study of MT systems applied to literary texts, showed that neural MT could achieve substantial levels of quality and, in some cases, approach the quality of professional human translation. Their findings challenge the assumption that literary translation is necessarily beyond the capabilities of computational systems, while also highlighting the need for evaluation methods capable of accounting for the specific characteristics of literary texts. The question is therefore not simply whether MT can produce an adequate literary translation, but which dimensions of literary form and function are preserved, transformed, or lost in the process.

The relationship between machine translation and literary creativity further complicates the assessment of quality. Guerberof-Arenas and Toral (2020) found differences between human, machine-translated, and post-edited literary texts in terms of creativity and reading experience, suggesting that literary translation involves dimensions that extend beyond conventional measures of linguistic accuracy and fluency. This issue is particularly relevant to LLMs, whose capacity to generate stylistically varied and contextually coherent language may suggest a high degree of literary competence without necessarily guaranteeing consistent strategic control over stylistic choices.

James Joyce's *Ulysses* provides a particularly demanding case for examining this distinction. Joyce's novel is characterised by extensive linguistic experimentation, stylistic variation, multilingualism, historical registers, and formal play. These features are especially concentrated in the "Oxen of the Sun" episode, in which the prose undergoes a succession of stylistic transformations associated with different periods and traditions of English writing. Translating this episode therefore requires more than semantic transfer: the translator must negotiate changes in register and historical style while also responding to multilingual elements, formal experimentation, and wordplay.

This case makes it possible to examine a broader question concerning AI-mediated language: what happens to linguistically and culturally marked forms when an LLM is required to select among competing translational possibilities? The question is not limited to whether the resulting sentence is correct or fluent. It concerns which features of the source text remain visible in the target text, which are transformed, and which are effectively normalised. In this sense, literary translation offers a particularly sensitive site for investigating the relationship between linguistic competence, translational decision-making, and the mediation of cultural and epistemic difference.

In international education, where multilingual communication increasingly relies on AI-mediated language technologies, such processes may shape how students and scholars encounter linguistic and cultural difference across educational contexts.

1.2. Research Question and Contribution

Against this background, the present study investigates how LLM-mediated translation negotiates linguistic and stylistic difference in a highly experimental literary text. It focuses on the "Oxen of the Sun" episode of *Ulysses*, comparing an LLM-generated Italian translation with a professional human translation and a conventional machine translation. The analysis combines qualitative close reading with quantitative evaluation based on an adaptation of the Multidimensional Quality Metrics (MQM) framework. MQM provides a structured and extensible framework for describing translation-quality issues through categories and severity levels (Lommel et al., 2014). In the present study, the framework has been adapted by the author specifically for literary translation, with particular attention to stylistic phenomena that conventional quality assessment may not adequately capture.

The study addresses three research questions:

RQ1. How do LLMs translate stylistically heterogeneous, historically marked, and multilingual passages in *Ulysses*?

RQ2. What patterns of translational intervention characterise LLM-generated translations in comparison with human and conventional machine translation?

RQ3. To what extent can an MQM framework adapted for literary translation capture differences in the strategic management of linguistic and culturally marked forms?

The central hypothesis is that LLMs will demonstrate relatively strong performance in grammatical fluency and baseline semantic transfer while showing greater difficulty in situations requiring strategic management of stylistic heterogeneity. In particular, they are expected to normalise or flatten stylistically marked features, especially where translation requires the recreation of diachronic stylistic shifts, multilingual elements, or complex wordplay. Such limitations may therefore manifest not primarily as conventional mistranslations, but as insufficient, excessive, or inconsistent translational intervention.

The study makes two principal contributions. First, it proposes a methodological adaptation of MQM for literary translation that treats stylistic phenomena as sites of strategic intervention rather than exclusively as discrete translation errors. Second, it applies Levý's decision-making framework to the empirical study of LLM translation,

examining whether current systems can move beyond fluent linguistic production towards contextually and stylistically appropriate selection among competing translational possibilities. More broadly, the study contributes to discussions of AI, language, and epistemic hierarchies by examining how AI-mediated translation can preserve, transform, or normalise linguistically and culturally situated forms of expression. In doing so, it positions literary translation as a particularly sensitive case for investigating the broader implications of AI-mediated language practices in multilingual and internationally oriented contexts.

2. Translation as Decision-Making: Theoretical Framework

2.1. Literary Translation and Stylistic Complexity

Literary translation has long been understood as a form of mediation in which linguistic form and textual meaning cannot be considered entirely separable. Unlike functional or informational texts, literary works frequently derive part of their significance from the interaction between semantic content and stylistic organisation. Register, rhythm, syntax, lexical choice, sound patterning, ambiguity, and figurative language may therefore constitute meaning-bearing elements rather than merely decorative embellishments. Baker (2000) demonstrates that the stylistic profile of a literary translator can be investigated through recurrent linguistic patterns, while Boase-Beier (2004) emphasises the relationship between translation and the interpretation of stylistic choices.

This problem becomes particularly complex when the source text deliberately incorporates multiple historical or social varieties of language. In such cases, translation involves not simply finding lexical or syntactic equivalents but determining how differences between registers should be represented in the target language. The problem is especially acute in texts whose stylistic organisation is itself structurally significant. *Ulysses* represents an extreme example because its linguistic experimentation repeatedly foregrounds the relationship between historical language, literary convention, register, and textual form.

The “Oxen of the Sun” episode provides an especially suitable case for investigating these problems. The episode develops through a sequence of stylistic transformations that evoke different periods of English prose, moving from archaic and medieval forms through later literary and journalistic registers towards increasingly contemporary modes of expression. The stylistic progression is therefore not incidental to the narrative but constitutes one of the episode’s principal organising devices. A translation that renders individual propositions accurately may nevertheless substantially alter the function of the passage if it fails to reproduce the contrasts between the different stylistic stages.

From this perspective, literary translation quality cannot be reduced to semantic equivalence. House (2015) similarly argues that translation quality assessment requires attention to contextual and functional dimensions rather than exclusively to isolated linguistic correspondences. For experimental literary texts, this means that evaluation must account for the interaction between accuracy, fluency, register, stylistic variation, and textual function.

2.2. Translation Quality Assessment and MQM

The assessment of literary translation presents a methodological challenge because conventional quality metrics were frequently developed for purposes different from those of literary analysis. Automatic metrics based primarily on surface correspondence can provide useful information about particular dimensions of translation performance

but cannot independently determine whether a literary translation has preserved stylistic differentiation, rhetorical effects, or historically marked language.

The Multidimensional Quality Metrics (MQM) framework offers a potentially useful basis for addressing this problem. MQM was developed as an open and extensible framework for describing translation-quality issues through a shared taxonomy of issue types (Lommel et al., 2014). Its original purpose was to provide a common framework capable of describing different approaches to translation-quality evaluation and of facilitating comparison between human and machine translation.

An important advantage of MQM for the present study is its multidimensional structure. Rather than reducing quality to a single binary distinction between correct and incorrect translation, the framework allows different kinds of problems to be identified and classified separately. This makes it possible to distinguish, for example, problems of semantic accuracy from problems of fluency or stylistic realisation. Recent interdisciplinary work on AI in language education likewise stresses the importance of reliability, equity, and human-guided use (Imran et al., 2026).

The present study adapts this principle to literary translation. The categories Accuracy, Fluency, Style/Pastiche, and Terminology/Register are treated as analytically distinct dimensions, while severity levels distinguish minor deviations from major and critical problems. The adaptation is necessary because stylistic variation in *Oxen of the Sun* cannot always be treated as a conventional “error”. A target-text choice may be grammatically correct and semantically plausible while nevertheless weakening the stylistic contrast deliberately established by Joyce.

MQM therefore functions in this study not as an absolute measure of literary value but as a structured descriptive instrument. Its purpose is to make differences between translations systematically observable while leaving the interpretation of those differences to qualitative analysis.

2.3. Translation as Decision-Making

The second theoretical pillar of the study is Jiří Levý's conception of translation as a decision process. In his influential 1967 essay, Levý conceptualises translation as a succession of choices among alternative solutions. Each decision is conditioned by the situation created by previous decisions and contributes to the possibilities available at subsequent stages of the translation process (Levý, 1967). The original essay was published in *To Honor Roman Jakobson*, Volume 2, pp. 1171–1182.

Levý's model is particularly relevant to literary translation because it shifts the analytical focus from isolated equivalences towards the cumulative consequences of translational choices. A translator confronted with a stylistically marked expression does not encounter a single predetermined equivalent. Rather, multiple possible solutions may be available, each emphasising different aspects of the source text. Choosing one solution therefore simultaneously limits or enables subsequent choices.

This perspective is especially productive for the analysis of LLM-generated translation. An LLM may produce a linguistically plausible solution without necessarily reproducing the strategic relationship between individual choices across a passage. For example, the normalisation of an archaic lexical item may appear insignificant at sentence level but become consequential when similar normalisations occur repeatedly and progressively erase the stylistic differentiation of an entire section.

Levý's model consequently provides a conceptual bridge between translation quality and translational decision-making. MQM makes it possible to identify and classify observable differences between source and target texts, while Levý provides a theoretical basis for interpreting those differences as consequences of choices among alternative translational possibilities. The two frameworks therefore serve

complementary purposes: MQM supplies the descriptive and quantitative dimension, whereas Levý supplies the decision-theoretical dimension.

This distinction is important for the present study because the central question is not simply whether an LLM produces more or fewer errors than a human translator or an NMT system. It is whether the distribution and nature of its interventions reveal a different approach to stylistically complex translation problems. In this sense, the study treats observable translation deviations as potential evidence of underlying patterns of translational decision-making rather than as isolated failures of linguistic competence.

2.4. From Linguistic Competence to Strategic Competence

The combination of these perspectives suggests a distinction between linguistic competence and strategic translational competence. Linguistic competence concerns the ability to generate grammatically acceptable, semantically plausible, and contextually coherent target-language text. Strategic competence, by contrast, concerns the ability to determine which aspects of the source text require preservation, transformation, compensation, or deliberate deviation in accordance with the broader function of the translation.

This distinction is particularly significant for LLMs because their outputs can exhibit a high degree of linguistic fluency while still displaying systematic patterns of normalisation. A fluent translation may therefore conceal substantial losses at the level of stylistic differentiation. Conversely, a translation containing formal departures from the source may sometimes represent a deliberate attempt to reproduce an equivalent literary effect.

The theoretical framework adopted here consequently avoids treating fluency as a sufficient proxy for translation quality. Instead, it combines the multidimensional description offered by MQM with Levý's decision-oriented understanding of translation. This provides the conceptual basis for examining whether LLMs merely generate plausible target-language formulations or demonstrate a more systematic capacity to manage the competing demands posed by highly experimental literary texts.

3. Methodology

3.1. This Corpus and Translation Data

The study adopts a comparative case-study design centred on James Joyce's "Oxen of the Sun", Episode XIV of *Ulysses*. The episode was selected because its extreme stylistic heterogeneity provides a particularly demanding environment for examining AI-mediated translation. Joyce's prose moves through historically differentiated linguistic registers and incorporates multilingual elements, rhetorical experimentation, lexical ambiguity, and forms of wordplay. These characteristics make the episode particularly suitable for investigating whether LLM-generated translation can preserve not only propositional meaning but also the linguistic and cultural distinctions encoded in literary form.

The analysis compares three Italian versions of the selected passage: Enrico Terrinoni's professional translation, a translation generated by ChatGPT, and a translation produced by Google Translate. The human translation functions as the professional reference point, while the two AI-mediated versions allow comparison between an LLM and a conventional neural machine translation system. The purpose is not to establish a simple hierarchy between human and machine translation, but to identify differences in the types and patterns of translational intervention produced by each system.

The comparative analysis combines quantitative MQM-based assessment with qualitative close reading. Each translation was examined for issues affecting accuracy,

fluency, style, pastiche, register, terminology, and formal experimentation. Particular attention was given to cases in which a linguistically marked feature of the source was preserved, transformed, neutralised, or replaced in the target text.

3.2. Adaptation of MQM for Literary Translation

The quantitative component is based on the Multidimensional Quality Metrics (MQM) framework, originally developed as a structured framework for describing translation-quality issues through defined error categories and severity levels (Lommel et al., 2014). For the purposes of this study, MQM was adapted by the author specifically for the evaluation of literary translation. This adaptation was necessary because the standard framework, while suitable for systematic quality assessment, does not by itself fully capture the significance of literary phenomena such as sustained pastiche, diachronic stylistic differentiation, literary register, tone, and formal experimentation.

The adapted framework retains the principal MQM dimensions of Accuracy, Fluency, Style, Terminology/Register, and Locale Convention, while assigning particular analytical importance to stylistic and formal features relevant to Joyce's text. Within the Style category, for example, errors may concern the reproduction or consistency of pastiche, archaism, register, or tone. Accuracy includes conventional issues such as mistranslation, omission, and addition, while Fluency captures problems affecting grammaticality, readability, and target-language coherence.

Each identified issue is assigned a severity level: Minor, representing a limited deviation with little effect on overall interpretation; Major, representing a substantial problem affecting meaning or literary quality; and Critical, representing a severe distortion, omission, or unintelligible rendering. The weighted score consequently allows different types of problems to be distinguished rather than treating all deviations as equivalent.

The purpose of this adaptation is not to reduce literary quality to an aggregate numerical value. Rather, the MQM results provide a systematic empirical basis for identifying recurring patterns, which are subsequently interpreted through qualitative analysis and Levý's conception of translation as a decision-making process (Levý, 1967, 2011). This combination makes it possible to examine not only where translation problems occur, but also what broader patterns of linguistic normalisation, stylistic intervention, and strategic choice they reveal.

Because the present study adapts MQM specifically to the literary characteristics of "Oxen of the Sun", the resulting totals are corpus-specific comparative indicators rather than universal measures of literary value. The same categories and severity distinctions were applied to all three translations, so the comparison is based on a common analytical procedure. The large numerical differences should therefore be interpreted primarily as the cumulative effect of repeated issues across the same passage, especially in Style/Pastiche and Terminology/Register, rather than as evidence that one translation is simply a fixed percentage better than another. The weighted penalties and Quality Scores reported below are consequently treated as descriptive outcomes of this study's coding procedure, and not as standardized scores that can be directly generalized beyond the corpus.

3.3. Analytical Procedure

The analysis proceeded in two complementary stages. First, each translation was examined systematically according to the categories and severity levels established in the adapted MQM framework. Instances of Accuracy, Fluency, Style/Pastiche, and Terminology/Register issues were identified and classified as Minor, Major, or Critical. The resulting counts were then weighted according to the severity assigned to each issue, allowing the three translations to be compared on a common analytical scale.

The quantitative assessment was subsequently complemented by qualitative close reading. Particular attention was paid to passages in which the three translations adopted substantially different solutions to the same source-text feature. These instances were examined in relation to four principal dimensions: semantic transfer; diachronic style and register; multilingualism; and wordplay and formal experimentation. The purpose of the qualitative stage was to determine whether the numerical distribution of MQM issues corresponded to recurring patterns of translational behaviour.

The magnitude of the differences is also related to the unit of analysis. The study counts individual instances of Accuracy, Fluency, Style/Pastiche, and Terminology/Register issues, and then considers their severity. Consequently, a translation that repeatedly neutralises stylistically marked features can accumulate a substantially larger penalty even when many individual target sentences remain grammatical and intelligible. This is particularly relevant to Joyce's episode, where stylistic differentiation is distributed across the passage rather than concentrated in a small number of isolated expressions.

A further analytical distinction was made between translation error and translational intervention. A departure from the source text was not automatically classified as an error when it could be interpreted as a deliberate strategy for reproducing a literary effect in Italian. Conversely, a formally close rendering could be classified as problematic when its literal correspondence resulted in semantic distortion, stylistic inconsistency, or the loss of a functionally significant feature. This distinction follows the study's application of Levý's conception of translation as a process of selecting among alternative solutions (Levý, 1967, 2011).

The final stage integrated the quantitative and qualitative findings. The MQM results were used to establish the distribution and severity of identified problems, while close reading was used to interpret the direction and consistency of translational intervention. This procedure makes it possible to distinguish isolated errors from recurring patterns such as normalisation, insufficient intervention, or stylistic flattening. It also provides the basis for examining whether the stronger linguistic fluency of LLM-generated translation corresponds to a comparable level of strategic control over linguistically, historically, and culturally marked literary features.

3.4. Translational Intervention and Decision-Making

The final analytical stage connects the MQM-based results with Levý's conception of translation as a decision-making process (Levý, 1967). The purpose of this connection is not to infer the internal reasoning of the systems under examination, but to interpret the observable features of their outputs as the consequences of different patterns of translational intervention. In other words, the study does not claim direct access to the decision-making processes of an LLM; it analyses the choices embodied in the resulting target text.

For each translation, particular attention was therefore paid to what was preserved, transformed, omitted, added, or normalised in relation to the source text. These interventions were considered especially significant when they affected stylistically marked elements. For instance, the replacement of an archaic expression with a contemporary equivalent may preserve its basic propositional meaning while simultaneously reducing the historical character of the passage. Similarly, the omission of a multilingual expression or the simplification of a wordplay may produce a semantically intelligible target text while altering the formal and stylistic function of the original.

The analysis consequently distinguishes between linguistic performance and strategic translational intervention. Linguistic performance refers to the ability to produce a grammatically acceptable, semantically coherent, and readable target text.

Strategic intervention, by contrast, concerns the way in which a translator responds to competing demands when no single solution can simultaneously preserve all dimensions of the source text. In such cases, translation necessarily involves prioritisation: a translator may favour semantic transparency, formal similarity, stylistic equivalence, readability, or the preservation of historical and cultural associations.

This distinction is particularly relevant when comparing the three translation outputs. The professional translation provides evidence of human literary decision-making, whereas the LLM and NMT outputs provide evidence of computationally generated translational choices. The comparison does not assume that human decisions are inherently superior or that machine decisions are inherently defective. Instead, it examines whether systematic differences emerge in the way the three approaches manage the tensions between semantic accuracy, fluency, register, and literary style.

The analysis therefore treats patterns of intervention, rather than individual errors alone, as the principal unit for interpreting translational decision-making. Repeated normalisation, systematic omission of marked expressions, or recurrent substitution of stylistically neutral forms may indicate a broader tendency in the treatment of literary language. Conversely, consistent preservation or deliberate recreation of stylistic distinctions may indicate a greater degree of strategic control.

By combining the adapted MQM framework with Levý's decision-oriented perspective, the methodology thus moves from the identification of translation problems towards an analysis of how different translation agents respond to the choices imposed by literary form. This provides the basis for the subsequent Results and Discussion sections, where the quantitative distribution of MQM issues is examined alongside representative examples of translational intervention.

4. Results

The analysis reveals substantial differences among the three translations across the dimensions assessed through the adapted MQM framework. The quantitative results indicate that the professional human translation consistently presents fewer and less severe quality issues than the two AI-generated versions, while the comparison between ChatGPT and Google Translate shows a marked difference in both error frequency and type. These differences are particularly visible when accuracy and fluency are considered alongside stylistic and register-related dimensions.

The distribution of MQM issues also indicates that the performance of the two AI systems cannot be adequately described through a single measure of translation quality. ChatGPT generally produces more fluent and grammatically coherent Italian than Google Translate, but it also presents a substantial number of Style/Pastiche and Terminology/Register issues. Google Translate, by contrast, shows higher frequencies of problems across both semantic and formal dimensions. Terrinoni's translation presents comparatively limited deviations, including a small number of minor issues, while maintaining a considerably more consistent treatment of the stylistic organisation of the source passage.

These quantitative differences are complemented by the qualitative analysis. The most relevant contrasts emerge in the treatment of semantic transfer and fluency, diachronic style and register, multilingualism, and wordplay and formal experimentation. Across these dimensions, the AI-generated translations display recurring patterns of intervention that differ from those observed in the professional translation.

The following subsections examine these patterns in detail. The analysis first considers fluency and semantic transfer, then focuses on diachronic style, register and multilingualism, followed by wordplay and formal experimentation. The final

subsection examines the broader patterns of translational intervention and stylistic flattening emerging from the comparison.

The magnitude of the numerical gap requires careful interpretation. The figures do not represent independent observations or a statistically normalised scale; they are aggregate results produced by applying the same study-specific coding procedure to three translations of the same literary passage. The difference between Terrinoni (18), ChatGPT (215), and Google Translate (485) therefore reflects the accumulation of identified issues and their assigned severity across the corpus. In particular, the much larger contribution of Style/Pastiche and Terminology/Register issues in the two AI-generated versions is important because these categories capture recurring losses of stylistic differentiation rather than only isolated lexical mistakes. The numerical distance should thus be read as evidence of different distributions and concentrations of problems within this case study, not as a claim that the three systems occupy fixed positions on a universally valid translation-quality scale.

4.1. Quantitative Distribution of MQM Errors

The quantitative analysis provides an initial comparison of the three Italian translations according to the MQM framework adapted for literary translation. Table 1 presents the distribution of identified issues across the main evaluation categories and severity levels, together with the resulting weighted penalties and overall MQM Quality Scores. The data show a substantial difference between the professional human translation and the two AI-generated versions. Terrinoni's translation records the lowest number of issues and the lowest weighted penalty, whereas ChatGPT and Google Translate present progressively higher frequencies of errors and greater concentrations of Major and Critical issues.

Table 1. Distribution and weighted severity of MQM issues across the three Italian translations of “Oxen of the Sun”

MQM category	Terrinoni	ChatGPT	Google Translate
Accuracy – Minor	4	12	28
Accuracy – Major	1	8	19
Accuracy – Critical	0	3	9
Fluency – Minor	3	15	32
Fluency – Major	0	6	14
Style/Pastiche – Minor	2	18	25
Style/Pastiche – Major	1	14	22
Style/Pastiche – Critical	0	5	11
Terminology/Register – Minor	1	9	16
Terminology/Register – Major	0	4	8
Total weighted penalties	18	215	485

The distribution of Accuracy issues shows the clearest difference in semantic transfer. Terrinoni's version contains only five identified accuracy issues, with no Critical errors, whereas ChatGPT records substantially more Major and Critical issues. Google Translate presents the highest frequency across all three severity levels, indicating a greater incidence of deviations affecting the transfer of source-text meaning.

A similar pattern emerges in Fluency. Terrinoni presents only a small number of Minor issues and no Major issues. ChatGPT produces considerably more Fluency issues, but the majority are Minor, while Google Translate records the highest number of both Minor and Major problems. This distinction is relevant because the comparatively strong

fluency of ChatGPT does not eliminate difficulties in other dimensions of translation quality.

The differences are particularly pronounced in Style/Pastiche, which is central to the present corpus. Terrinoni records only three stylistic issues, whereas ChatGPT records 37 and Google Translate 58. The concentration of Major and Critical issues in the two AI-generated translations indicates that the difficulties observed in the AI outputs extend beyond isolated lexical choices and affect the reproduction of the stylistic organisation of the episode.

The Terminology/Register category shows the same general tendency. Terrinoni records only one Minor issue, while ChatGPT and Google Translate present substantially higher numbers of both Minor and Major issues. This pattern is particularly relevant to “Oxen of the Sun”, where shifts in historical and social register form an essential component of Joyce’s formal design.

Taken together, the weighted results show a substantial quantitative difference among the three translations. The total weighted penalty is lowest for Terrinoni (18), followed by ChatGPT (215), and highest for Google Translate (485). The corresponding study-specific MQM Quality Scores are 96.4%, 57.0%, and 3.0%, respectively. These figures provide a quantitative overview of the differences identified in the analysis; the following sections examine how these differences are manifested in the translation of semantic, stylistic, and formal features.

4.2. Fluency and Semantic Transfer

The comparison of fluency and semantic transfer reveals a clear distinction between the three translations. The quantitative results show that Terrinoni’s version presents only a limited number of issues in these dimensions, while ChatGPT and Google Translate display progressively higher frequencies of both Accuracy and Fluency problems. The qualitative analysis further indicates that fluency and semantic transfer do not necessarily develop in parallel in the AI-generated translations.

Terrinoni’s translation maintains a close relationship between semantic interpretation and target-language formulation. Its limited number of Accuracy issues suggests that the propositional content of the source is generally preserved, while the absence of Major Fluency issues indicates a high degree of grammatical and syntactic control. Importantly, this does not mean that the translation remains structurally identical to the English source. Departures from the source syntax frequently function as adaptations to Italian rather than as evidence of inaccurate transfer.

ChatGPT produces a substantially more fluent target text than Google Translate. Its output generally maintains grammatical coherence and produces readable Italian, which is reflected in the lower number of Major Fluency issues recorded in comparison with Google Translate. However, the higher number of Accuracy issues indicates that this fluency does not always correspond to equivalent semantic precision. In several cases, complex or marked source constructions are rendered through more conventional formulations, resulting in a target text that is locally coherent but does not fully preserve the semantic or rhetorical relationships of the original.

Google Translate presents the greatest difficulty at both levels. Its higher number of Accuracy issues is accompanied by a considerably larger number of Fluency problems, indicating that difficulties in interpreting Joyce’s complex syntax can also affect the grammatical and semantic stability of the target text. Literal or closely aligned renderings may preserve individual lexical elements while failing to reconstruct the relationships between them, particularly when meaning depends on unusual syntax or multiple simultaneous interpretations.

The results therefore indicate that fluency cannot be treated as a sufficient proxy for semantic adequacy. ChatGPT’s comparatively fluent output represents a significant

linguistic strength, but the MQM distribution demonstrates that fluent target-language production can coexist with substantial semantic and stylistic intervention. In “Oxen of the Sun”, this distinction is particularly important because meaning is frequently constructed through the interaction of syntax, register, historical language, and formal experimentation. The following analysis consequently turns from semantic and fluency-related performance to the more specific problem of preserving Joyce's diachronic stylistic organisation.

4.3. Diachronic Style, Register, and Multilingualism

The preservation of diachronic style represents one of the most demanding aspects of translating “Oxen of the Sun”. Joyce structures the episode through a succession of historically and stylistically differentiated forms of English, making linguistic variation itself part of the organisation and meaning of the text. The translator must therefore reproduce not only the propositional content of the source, but also the perception of historical distance and stylistic change created through the sequence of registers.

The MQM results show a substantial difference among the three translations in the Style/Pastiche and Terminology/Register categories. Terrinoni records only a small number of issues in both dimensions, whereas ChatGPT and Google Translate present considerably higher frequencies. The difference is particularly pronounced for Style/Pastiche, where the AI-generated versions accumulate both Major and Critical issues. This indicates that the difficulty of the AI systems is not restricted to individual archaic expressions, but extends to the maintenance of stylistic differentiation across the passage.

Terrinoni's translation demonstrates a comparatively consistent management of register. Rather than reducing the passage to a uniform contemporary Italian register, it employs marked lexical and syntactic choices to construct target-language equivalents for the historical progression of Joyce's prose. The resulting variation operates across the text rather than being confined to isolated lexical items.

ChatGPT displays a stronger tendency towards diachronic neutralisation. Although individual archaic or stylistically marked expressions may be recognised and reproduced, the distinctions between successive stylistic phases are less consistently maintained. Marked formulations are frequently rendered through relatively standard contemporary Italian, producing a target text that is generally readable but less differentiated. The issue is therefore not necessarily incorrect translation at the sentence level, but the cumulative reduction of stylistic contrasts that are structurally significant in the source.

Google Translate exhibits a more pronounced difficulty in maintaining these distinctions. Historically marked or rhetorically complex formulations are more frequently reproduced through literal equivalents or inappropriate contemporary formulations. Register-related problems also interact with Accuracy and Fluency issues, making it difficult for the translation to establish a coherent diachronic progression.

Multilingualism introduces an additional difficulty. Linguistic variation in Joyce cannot always be treated as supplementary information that can simply be translated into a standard target-language equivalent. The translator must determine whether a foreign or culturally marked element should be retained, adapted, explained, or replaced by a functionally comparable form. The three translations differ in the extent to which they maintain this linguistic differentiation.

Overall, the results indicate that the central challenge lies in constructing systematic relationships among stylistic choices, rather than reproducing isolated archaic or multilingual expressions. This distinction is particularly important for an episode whose literary organisation depends on the cumulative interaction of historical register, linguistic variation, and formal experimentation.

4.4. Wordplay and Formal Experimentation

Wordplay and formal experimentation constitute another major area of difference among the three translations. In “Oxen of the Sun”, Joyce frequently makes linguistic form inseparable from meaning: lexical ambiguity, sound patterning, syntactic distortion, rhetorical parallelism, and historically marked vocabulary contribute simultaneously to semantic and stylistic effects. The translation of such features therefore requires more than the reproduction of individual lexical meanings. It requires decisions concerning which formal relationships can be preserved, transformed, or compensated for in Italian.

The MQM results indicate a substantial difference between the three versions in the treatment of Style/Pastiche. Terrinoni records only three issues in this category, including one Major issue, whereas ChatGPT records substantially more Minor, Major, and Critical issues. Google Translate presents the highest number across all three severity levels. The quantitative distribution therefore suggests that the difficulties observed in the AI-generated translations extend beyond isolated lexical inaccuracies and affect the reproduction of formal characteristics that contribute to the literary identity of the passage.

In Terrinoni’s translation, departures from the surface structure of the English text can function as deliberate adaptations. Where a literal reproduction would fail to recreate a rhetorical or stylistic effect in Italian, the translator may modify lexical or syntactic structures in order to construct a functionally comparable effect. Such intervention is particularly relevant to wordplay, where formal correspondence and semantic correspondence may be impossible to maintain simultaneously.

ChatGPT is capable of reproducing individual figurative, rhetorical, or formally marked expressions, but these solutions appear less consistently integrated into the wider formal organisation of the passage. Ambiguous expressions may be rendered through more explicit alternatives, while syntactically unusual constructions may be regularised into conventional Italian structures. The resulting text can remain fluent and intelligible, but some of the formal tension and ambiguity of the source is reduced.

Google Translate shows the greatest difficulty in this respect. Its tendency towards close lexical and syntactic transfer can preserve individual elements of the source while failing to reconstruct the relationships between them. When Joyce’s meaning depends simultaneously on sound, ambiguity, syntax, and register, such literal correspondence may result in a target expression that is formally recognisable but functionally ineffective.

The comparison therefore shows that the central issue is not simply whether a system can translate a particular instance of wordplay, but whether it can manage competing formal and semantic priorities consistently across the text. In this respect, the results reinforce the distinction between linguistic reproduction and strategic translational intervention. The preservation of literary form may require a departure from the source wording, while excessive adherence to surface form may itself become a source of semantic or stylistic loss.

4.5. Patterns of Translational Intervention and Stylistic Flattening

The comparison of the three translations indicates that the differences identified through the MQM analysis are not limited to isolated errors. Rather, they form recurring patterns of translational intervention that affect the way stylistic, historical, and formally marked features are represented in the target text. The most significant contrast concerns the relationship between intervention and stylistic differentiation.

Terrinoni’s translation displays a relatively controlled and systematic pattern of intervention. Departures from the linguistic form of the source are frequently accompanied by an effort to preserve distinctions of register, historical style, and

rhetoical effect. The translation therefore does not seek uniform formal correspondence; instead, it modifies the target-language expression when necessary to maintain the functional differentiation of the source passage.

ChatGPT displays a different pattern. Its output frequently resolves marked or unusual formulations into more conventional contemporary Italian. This tendency produces a generally readable target text, but it reduces the contrast between stylistically differentiated portions of Joyce's prose. The phenomenon can be described as stylistic flattening: individual choices may be grammatically and semantically acceptable, yet their cumulative effect is to make the target text less heterogeneous than the source.

Google Translate exhibits a third pattern, characterised more frequently by insufficient adaptation to the target language. Its close adherence to lexical and syntactic features of the source does not necessarily preserve literary function. Where Joyce's formulations require substantial restructuring, limited intervention can produce awkward, semantically unstable, or stylistically inappropriate renderings. In this case, the problem is not excessive intervention but the absence of sufficiently adaptive intervention.

The comparison therefore reveals three broadly distinguishable patterns: controlled intervention in the professional translation, normalising intervention in ChatGPT, and insufficient or literal intervention in Google Translate. These patterns help explain why the differences identified in the MQM scores cannot be reduced to a simple hierarchy of sentence-level accuracy.

Most importantly, stylistic flattening emerges as a cumulative phenomenon rather than as a single identifiable error. It results from the repeated reduction of lexical, syntactic, historical, and rhetorical variation across the text. The quantitative MQM results make this pattern visible through the concentration of Style/Pastiche and Terminology/Register issues, while the qualitative comparison shows how individual translational choices contribute to the broader transformation of Joyce's stylistic architecture.

4.6. Comparative Summary of Findings

The quantitative and qualitative analyses converge on a consistent pattern across the three translations. Terrinoni's version records the lowest frequency and severity of MQM issues and, at the qualitative level, demonstrates the most consistent management of semantic, stylistic, historical, and formal features. The translation does not simply reproduce the linguistic surface of the source text; rather, it frequently employs adaptive solutions that maintain distinctions that are central to Joyce's literary design.

ChatGPT occupies an intermediate position. Its output demonstrates substantially greater fluency and grammatical coherence than Google Translate, together with a stronger capacity to transfer the general semantic content of the source. At the same time, the analysis identifies recurring problems in the treatment of stylistically marked language, historical register, multilingual elements, and formal experimentation. The principal pattern is not systematic semantic failure, but normalisation and stylistic flattening, whereby marked features are frequently rendered through more conventional contemporary Italian.

Google Translate presents the highest concentration of MQM issues across the principal categories. Its difficulties affect both semantic transfer and target-language fluency, while its treatment of stylistic and formal features is particularly unstable. The results indicate that close lexical correspondence does not necessarily provide greater literary adequacy when the source text depends on syntactic complexity, ambiguity, historical register, or formal experimentation.

Taken together, the findings establish a distinction between linguistic performance and strategic translational management. The three translations differ not only in how

many problems they contain, but in the ways they respond to competing semantic, stylistic, and formal demands. This distinction provides the empirical basis for the Discussion that follows, where the results are considered in relation to translational decision-making, Levý's theoretical framework, and the broader implications of AI-mediated language and cultural representation.

5. Discussion: LLMs as Translational Decision-Makers

5.1. Linguistic Competence versus Strategic Competence

The results indicate that the distinction between linguistic competence and strategic competence is essential for interpreting LLM performance in literary translation. The three translations do not differ only in their number of errors; they reveal different relationships between linguistic form, semantic interpretation, and the management of stylistic variation. In particular, the relatively high fluency of the ChatGPT translation does not correspond to an equivalent capacity to preserve the stylistic differentiation of Joyce's text.

At the level of linguistic competence, LLMs demonstrate clear strengths. ChatGPT is generally capable of producing grammatically well-formed Italian, maintaining local coherence, and resolving many syntactic relationships in a way that results in a readable target text. This confirms observations in recent research that LLMs can exploit contextual information effectively in translation tasks, including at document level (Wang et al., 2023). Such capabilities represent a substantial development compared with earlier generations of machine translation.

However, the results suggest that linguistic competence alone is insufficient for the translation of highly experimental literary prose. In "Oxen of the Sun", the translator must continuously determine whether an unusual linguistic feature is accidental, semantically relevant, stylistically marked, historically significant, or formally motivated. The problem is therefore not simply understanding the linguistic input, but determining what should be done with that understanding in the target text.

This distinction corresponds closely to Levý's conception of translation as a decision process. Translation involves selecting among alternative solutions, with individual choices contributing to the construction of the target text as a whole (Levý, 1967, 2011). From this perspective, strategic competence concerns the ability to recognise competing translational possibilities and select among them according to the function of the passage.

The present results suggest that this capacity is less consistently demonstrated by the LLM-generated translations. ChatGPT frequently produces locally plausible solutions, but these choices can contribute cumulatively to stylistic normalisation. Google Translate exhibits even more pronounced difficulties, with problems affecting both semantic interpretation and linguistic realisation. Terrinoni's translation, by contrast, shows greater consistency between individual interventions and the broader stylistic organisation of the passage.

The distinction is therefore not between a system that "understands" Joyce and one that does not. Rather, the evidence points towards different levels of control over translational choice. An LLM may possess extensive linguistic knowledge and generate highly fluent language while still lacking a sufficiently stable strategy for deciding which features of an experimental literary text require preservation, transformation, compensation, or deliberate sacrifice. This distinction is central to understanding the limitations and possibilities of LLMs as literary translators.

5.2. Translational Intervention and Decision-Making

The differences identified in the Results can also be interpreted in terms of the degree and direction of translational intervention. The comparison suggests that the three translations do not simply occupy different positions on a scale from “accurate” to “inaccurate”; rather, they employ different strategies for responding to the formal and stylistic complexity of the source text. This distinction is particularly important in literary translation, where a departure from the linguistic form of the source may constitute an intentional solution rather than an error.

Terrinoni's translation displays the most sustained pattern of adaptive intervention. Changes in syntax, lexical selection, register, and rhetorical form appear to be coordinated with the broader stylistic development of the episode. Such interventions can be understood through Levý's model of translation as decision-making, in which the translator chooses among alternative possibilities according to the anticipated function of the target text (Levý, 1967). The relevant criterion is therefore not proximity to the source at the level of individual words, but the relationship between individual decisions and the overall organisation of the translation.

ChatGPT demonstrates a more uneven pattern. Its interventions are often effective at the local level, particularly when restructuring complex English syntax into idiomatic Italian. However, the cumulative effect may be one of normalisation. Stylistically marked expressions are sometimes replaced by more conventional alternatives, and historically differentiated passages may converge towards a relatively uniform contemporary register. The issue is consequently not an inability to intervene, but the absence of a consistently identifiable strategy governing when and how much to intervene.

Google Translate occupies a different position. Its output frequently remains closer to the surface structure of the source, even where such closeness produces awkward or semantically unstable Italian. In this case, insufficient intervention becomes itself a source of translational difficulty. The system does not consistently adapt constructions whose literary or grammatical function depends on substantial restructuring in the target language.

The empirical findings also clarify what is meant here by epistemic hierarchies. The study does not claim that the MQM scores directly measure epistemic hierarchy, nor that a translation output by itself establishes a broad social hierarchy of knowledge. Rather, the relevant empirical mechanism is selective visibility. When historically, socially, or stylistically marked forms are repeatedly replaced by standard contemporary formulations, some forms of linguistic knowledge remain visible in the target text while others become less visible. The recurring pattern of normalisation identified in the ChatGPT output therefore provides a concrete textual mechanism through which AI-mediated translation may participate in asymmetrical representation. In this sense, the connection between the quantitative findings and epistemic hierarchy is indirect but empirically grounded: MQM identifies the repeated Style/Pastiche and Terminology/Register deviations; close reading identifies normalisation and stylistic flattening as recurring interventions; and the epistemic argument concerns the unequal visibility produced by those interventions.

This comparison suggests that translational intervention should not be measured simply in quantitative terms. A high degree of intervention is not necessarily evidence of superior translation, just as limited intervention does not necessarily indicate fidelity. What matters is whether interventions are functionally motivated and consistently integrated into the target text. This is also compatible with research on translator style, which demonstrates that recurring patterns of linguistic choice can distinguish translations beyond individual instances of error (Baker, 2000).

The findings therefore support a conception of LLM translation in which the crucial issue is not merely the generation of alternative linguistic formulations, but the capacity to select among them according to a stable hierarchy of textual priorities. In highly experimental literature, this strategic dimension becomes particularly visible because decisions concerning register, syntax, lexical choice, and formal patterning can affect the cultural and historical identity of the target text.

5.3. Reconsidering Translation Quality through Levý's Framework

The findings also invite a reconsideration of how translation quality is conceptualised in the context of LLM-generated literary translation. Conventional evaluation models tend to focus on the identification of deviations from the source, particularly in relation to accuracy, fluency, terminology, and register. These dimensions remain important, but the present analysis suggests that they do not fully explain the qualitative differences observed between the three translations. In particular, they do not necessarily distinguish between a deviation resulting from an unsuccessful translation decision and one resulting from a deliberate attempt to recreate a literary effect.

Levý's theoretical framework provides a useful basis for addressing this distinction. His conception of translation as a decision process emphasises that the translator operates within a set of alternatives and constraints, selecting one solution from among several possible options (Levý, 1967). Translation quality can therefore be considered not only in terms of the distance between source and target texts, but also in terms of the appropriateness and consistency of the decisions through which that distance is produced.

This perspective is particularly relevant to the adapted MQM framework employed in the present study. MQM was originally developed as a structured framework for describing translation-quality issues through defined categories and severity levels (Lommel et al., 2014). In the present research, it has been adapted by the author specifically for literary translation, giving greater analytical weight to stylistic phenomena including pastiche, register, tone, and formal experimentation. The purpose of this adaptation is not to redefine MQM as a theory of literary value, but to make its descriptive categories more responsive to the characteristics of an experimental literary text.

Viewed through Levý's framework, the quantitative MQM results acquire a further interpretative dimension. A concentration of Style/Pastiche issues, for example, may indicate more than a series of isolated stylistic deviations: it may reveal a recurring difficulty in selecting appropriate solutions when several plausible alternatives are available. Similarly, Accuracy issues involving ambiguity or wordplay may reflect not simply deficient comprehension, but the selection of a solution that privileges one dimension of the source at the expense of another.

This interpretation is also consistent with recent interdisciplinary discussion of AI and language-related educational contexts. Imran et al. (2026), in a systematic review of AI in language education, identify generative language support as an important function of AI while also noting concerns related to reliability, equity, and human-guided use. Threadgill et al. (2026) similarly emphasise the importance of critical reflection and ethical awareness in AI literacy. These studies do not evaluate literary translation and are therefore not treated as evidence for the present translation results; they provide a broader context for the article's argument that AI-mediated language practices should be examined not only for output quality but also for how users and systems mediate linguistic choices. Recent work on machine translation ethics also highlights bias and cultural sensitivity as relevant dimensions of AI-mediated translation (Yao et al., 2025).

The comparison therefore suggests that literary translation quality is better understood as a relationship between error, intervention, and decision-making. MQM

provides the means of identifying and weighting observable problems, while Levý's framework helps explain the strategic dimension of the choices underlying them. This combination is particularly valuable for evaluating LLMs because it avoids equating linguistic fluency with translational competence. An LLM may generate a highly acceptable sentence while nevertheless making a strategically inappropriate decision at the level of the text as a whole.

From the perspective of AI, language, and epistemic hierarchies, this distinction is also significant. When a translation system repeatedly selects standardised or less culturally marked formulations, the resulting effect is not limited to stylistic simplification. It can also alter which linguistic and historical forms remain visible in the target text. The present findings therefore suggest that translation quality assessment can contribute to the study of AI-mediated language by making such patterns of selection, transformation, and normalisation empirically observable.

5.4. Implications for Literary Translation and Translation Studies

The findings have implications beyond the specific comparison of the three translations. First, they suggest that the evaluation of LLMs in literary translation should distinguish more carefully between linguistic performance and strategic translational competence. Current systems can generate fluent, coherent, and contextually appropriate language, but the present analysis indicates that these capabilities do not necessarily entail sustained control over stylistic and formal choices. This distinction becomes particularly significant when AI is used to mediate between languages and cultural contexts.

The epistemic-hierarchy interpretation should nevertheless remain proportionate to the evidence. Because the study examines one episode, one language pair, and three translation outputs, the findings demonstrate a textual pattern of differential visibility rather than a general causal claim about AI systems or knowledge hierarchies. Broader claims would require larger multilingual corpora, repeated generations, and comparative analysis across systems and cultural contexts.

Second, the results contribute to broader discussions of AI, language, and epistemic hierarchies. Translation is not a neutral transfer of linguistic information: choices concerning what is preserved, transformed, or normalised affect the visibility of particular linguistic and cultural forms. In the present case, the recurrent normalisation of historically and stylistically marked language in the LLM-generated translation demonstrates how AI-mediated language production may reduce forms of linguistic difference even when the resulting text remains fluent and intelligible. The issue is therefore not simply whether an AI system produces an accurate translation, but which forms of expression become more or less visible through its output. From a global perspective, examining these processes is particularly important because AI-mediated translation may reproduce or reduce linguistic and cultural asymmetries across languages, influencing which forms of knowledge and expression remain visible in transnational contexts.

The quantitative results should consequently be understood as corpus-bound evidence supporting the qualitative interpretation. The large differences in reported penalties do not by themselves establish a universal ranking of translation systems; their significance lies in the concentration and recurrence of particular issue types and in the corresponding patterns of translational intervention.

Third, these findings are relevant to multilingual and internationally oriented educational environments, where AI translation tools may increasingly mediate communication across linguistic boundaries. Although the present study does not examine students or educational practices directly, it identifies a mechanism that warrants further investigation in such contexts: AI-mediated language can facilitate

access while simultaneously reshaping the linguistic and cultural forms through which knowledge is encountered. This tension is particularly important when translated materials involve historically, culturally, or linguistically marked forms of expression.

Finally, the study supports the value of combining quantitative and qualitative approaches to AI translation assessment. The adapted MQM framework makes differences between the translations measurable, while close reading is necessary to determine the significance of those differences. Together, these approaches provide a basis for examining not only whether AI-generated translations are fluent or accurate, but also how they mediate linguistic and cultural difference. In this sense, literary translation provides a focused case through which broader questions about AI, language, and knowledge mediation can be investigated.

6. Conclusions

6.1. Main Findings and Contributions

The present study examined how LLM-mediated translation negotiates linguistic, stylistic, and culturally marked forms in James Joyce's "Oxen of the Sun" by comparing a professional human translation with outputs generated by ChatGPT and Google Translate. By combining qualitative close reading with an MQM framework adapted by the author specifically for literary translation, the study examined not only semantic accuracy and fluency but also the preservation of diachronic style, register, multilingualism, wordplay, and formal experimentation.

The results reveal a clear distinction between linguistic competence and strategic translational competence. ChatGPT demonstrates substantial strengths in grammatical fluency, local coherence, and general semantic transfer, but its output also displays recurrent patterns of normalisation and stylistic flattening. Google Translate presents more extensive difficulties across both semantic and stylistic dimensions. Terrinoni's translation, by contrast, demonstrates a more consistent relationship between individual translational interventions and the broader stylistic organisation of the source text.

A further methodological limitation concerns the reporting of the weighting procedure used to obtain the aggregate penalties. The present manuscript reports the resulting weighted penalties but does not provide a numerical weighting formula. Accordingly, these values are interpreted as the outcomes of the study's coding procedure rather than as a standard MQM scale. Future versions of the study should report the exact weighting scheme and coding protocol in full so that the numerical results can be reproduced independently.

The study makes two principal contributions. Methodologically, it demonstrates the potential of an MQM framework adapted specifically for literary translation to identify and quantify problems that conventional accuracy-oriented evaluation may overlook. Theoretically, it applies Levý's conception of translation as a decision process to the evaluation of LLM-generated literary translation, shifting attention from linguistic output alone towards the strategic selection of competing translational possibilities.

More broadly, the findings contribute to discussions of AI, language, and epistemic hierarchies by showing how AI-mediated translation may preserve, transform, or normalise linguistically and historically situated forms of expression. Literary translation thus provides a particularly sensitive case for examining how AI systems mediate linguistic and cultural difference – an issue with wider relevance to multilingual and internationally oriented educational contexts.

6.2. Limitations and Future Research

This study has several limitations that should be considered when interpreting its findings. First, the analysis is based on a single episode of a single literary work, namely

“Oxen of the Sun” from *Ulysses*. Although this episode represents an exceptionally demanding case of stylistic and formal experimentation, its characteristics cannot be assumed to be representative of literary translation as a whole. The findings should therefore be understood as those of a focused case study rather than as evidence of the general performance of LLMs in literary translation.

Second, the comparison involves a limited number of translation outputs. LLM-generated translation can vary according to model version, prompting, contextual information, and other generation conditions. Different prompts or configurations could therefore produce different translational decisions. Future research should incorporate multiple LLMs, repeated generations, and controlled prompting conditions in order to establish the stability and reproducibility of the patterns identified here.

A further limitation concerns the adaptation of MQM itself. The framework developed for this study introduces analytical categories intended to capture literary phenomena such as pastiche, diachronic style, register, and stylistic flattening. While this adaptation makes MQM more responsive to literary texts, its application necessarily involves interpretative judgement. Future research should therefore test the framework across larger datasets and multiple evaluators, with particular attention to inter-rater reliability and the operational distinction between stylistic deviation, legitimate translational intervention, and genuine error.

Finally, future research could extend the analysis to other episodes of *Ulysses*, other experimental literary works, and additional language pairs and LLM architectures. Such comparative research could investigate whether the patterns identified here – particularly normalisation and stylistic flattening – remain stable across linguistic and cultural contexts. It could also examine how AI-mediated translation functions in multilingual and internationally oriented educational environments, where questions of linguistic accessibility, cultural representation, and the mediation of knowledge become especially significant.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable. This study did not involve human participants or animals.

Conflicts of Interest: The authors declares no conflicts of interest.

References

- Baker, M. (2018). In other words: A coursebook on translation (3rd ed.). Routledge.
- Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target*, 12(2), 241–266.
- Bakhtin, M. M. (2010). The dialogic imagination: Four essays (M. Holquist, Ed.; C. Emerson & M. Holquist, Trans.). University of Texas Press.
- Boase-Beier, J. (2004). Translation and style: A brief introduction. *Language and Literature*, 13(1), 9–11.
- Castilho, S., Gaspari, F., Moorkens, J., Popović, M., & Toral, A. (2019). Editors’ foreword to the special issue on human factors in neural machine translation. *Machine Translation*, 33(1–2), 1–7.
- Duscovich, D. (2026). Intercultural competence and translation quality: Assessing AI and human literary translations. *Dialogica*, 2(1), 51–75.
- Elkins, K. (2024). In search of a translator: Using AI to evaluate what’s lost in translation. *Frontiers in Computer Science*, 6, 1444021.
- Gifford, D., & Seidman, R. J. (1988). *Ulysses annotated: Notes for James Joyce’s Ulysses*. University of California Press.
- Guerberof-Arenas, A., & Toral, A. (2020). The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2), 255–282.
- Guerberof-Arenas, A., & Toral Ruiz, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2), 184–212.
- House, J. (2015). *Translation quality assessment: Past and present*. Routledge.

- Imran, M., Almusharraf, N., & Shehzad, K. (2026). Artificial intelligence in school language education: A systematic review of trends, challenges, and future directions. *Journal of Interdisciplinary Studies in Education*, 15(6), 47–78. <https://doi.org/10.32674/w506dk3>
- Jadjuli, A.-fahad E., Lukman, R. G., Asibih, K. J., Godinez, V. L., Muñez, J. J., Maruji, A. L., & Cuaresma, J.-A. G. (2026). Preparing students for a global workforce: Artificial intelligence literacy, employability readiness, and learning experiences in technical higher education. *Journal of Interdisciplinary Studies in Education*, 15(7). <https://doi.org/10.32674/0yhz2m52>
- Joyce, J. (2021). *Ulissee: Testo inglese a fronte* (E. Terrinoni, Ed. & Trans.). Bompian
- Karpinska, M., & Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation (WMT)* (pp. 419–451). Association for Computational Linguistics.
- Levý, J. (1967). Translation as a decision process. In *To honor Roman Jakobson: Essays on the occasion of his 70th birthday*, 11 October 1966 (Vol. 2, pp. 1171–1182). De Gruyter Mouton.
- Levý, J. (2011). *The art of translation* (P. J. Corness, Trans.; Z. Jettmarová, Ed.). John Benjamins Publishing Company.
- Lommel, A. R., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12, 455–463.
- 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
- Orăsan, C. (2023). ChatGPT for translators: A survey. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications* (pp. 61–63). Association for Computational Linguistics
- Threadgill, B., Paudel, S., Bista, K., Gaulee, U., Prime, G., Anderson, P. C., Gole, R., Sharma, S., Mandal, U., Rajput, M., Tiwari, K., & Pandey, V. (2026). Understanding AI literacy among college students: Voices from a Historically Black College and University. *Journal of Interdisciplinary Studies in Education*, 15(2), 169–194. <https://doi.org/10.32674/w76bnk04>
- Toral, A., & Way, A. (2018). What level of quality can neural machine translation attain on literary text? In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment: From principles to practice* (pp. 263–287). Springer.
- Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 16646–16661). Association for Computational Linguistics.
- Wang, S., Wong, D. F., Yao, J., & Chao, L. S. (2024). What is the best way for ChatGPT to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 14025–14043). Association for Computational Linguistics.
- Yao, J., Kasuma, S. A. A., & Al-Hashimy, H. N. H. (2025). Exploring ethical considerations in machine translation: Addressing bias and cultural sensitivity. *Journal of Interdisciplinary Studies in Education*, 14(4), 99–116. <https://doi.org/10.32674/g97fy88>

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). The publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.