Specification-Aware Machine Translation and Evaluation for Purpose Alignment

Yoko Kayano[1,2]   **Saku Sugawara[1,2]**
[1]The Graduate University for Advanced Studies (SOKENDAI)
[2]National Institute of Informatics
{yokokayano,saku}@nii.ac.jp

Abstract

In professional settings, translation is guided by communicative goals and client needs, often formalized as specifications. While existing evaluation frameworks acknowledge the importance of such specifications, these specifications are often treated only implicitly in machine translation (MT) research. Drawing on translation studies, we provide a theoretical rationale for why specifications matter in professional translation, as well as a practical guide to implementing specification-aware MT and evaluation. Building on this foundation, we apply our framework to the translation of investor relations texts from 33 publicly listed companies. In our experiment, we compare five translation types, including official human translations and prompt-based outputs from large language models (LLMs), using expert error analysis, user preference rankings, and an automatic metric. The results show that LLM translations guided by specifications consistently outperformed official human translations in human evaluations, highlighting a gap between perceived and expected quality. These findings demonstrate that integrating specifications into MT workflows, with human oversight, can improve translation quality in ways aligned with professional practice.

1 Introduction

High-quality translation in professional settings requires more than a literal rendering of the source text. It must also fulfill a communicative purpose, which depends on factors such as the intended function, target audience, and the broader context of the original text (Reiss and Vermeer, 1984; Nord, 2006). A single source text may yield different translations depending on these factors.

These contextual factors are typically documented as *translation specifications*. A specification is a predefined set of conditions that guide the translation process, including purpose, audience, tone, style, and content priorities (ISO17100:2015; JTF, 2018). They help translators make informed decisions and ensure that the translation meets user needs (Reiss and Vermeer, 1984; Nord, 2006). Without such guidance, translators may struggle to begin the process at all.

Specifications are also essential in translation evaluation. Frameworks such as ISO 5060 and the Multidimensional Quality Metrics (MQM) emphasize specification-based assessment (ISO5060:2024; Lommel et al., 2013). Lommel et al. (2013) state that "translation quality can only be assessed in terms of whether or not a translation meets specified requirements and meets its communicative purpose." When specifications are absent or vague, evaluations tend to focus on surface-level features such as lexical accuracy or fluency, rather than on whether the translation achieves its communicative purpose. This perspective is also central to functionalist theories in translation studies, which hold that quality should be judged by how well a translation fulfills its intended purpose in the target context, rather than by equivalence with the source text (Reiss and Vermeer, 1984).

MQM is widely adopted in machine translation (MT) research, including the Conference on Machine Translation (WMT), where it underpins human evaluation (Freitag et al., 2021, 2024; Zerva et al., 2024). Its detailed error typology has contributed to translation evaluation. However, specifications are often treated implicitly, and the idea of translation as a goal-oriented process is not fully integrated into MT research. As a result, MT outputs often fall short in real-world applications where purpose and audience matter. This gap is increasingly problematic as industry clients now expect translations to serve specific business objectives (Lommel et al., 2024a).[1]

In response, we propose a framework for specification-aware MT and evaluation. Figure 1 outlines this framework, contrasting traditional

---

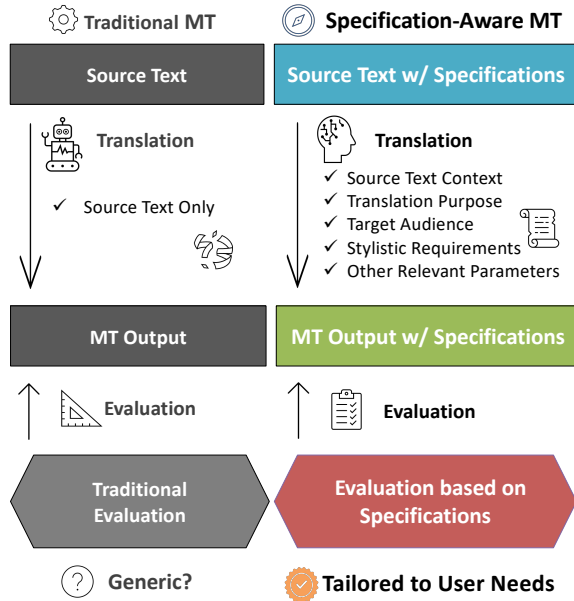[1]See Appendix A for further discussion.

Figure 1: Comparison of traditional MT pipelines, based solely on the source text, and specification-aware pipelines that incorporate contextual and purpose-specific information.

MT pipelines with our approach, which incorporates contextual information and specification-based evaluation. This approach reflects the functionalist view that translation should be guided by purpose, audience, and context. Incorporating specifications into MT workflows and evaluation is essential for improving translation quality in professional domains. To support this claim, we present the theoretical foundation for understanding the role of specifications in Section 3. We introduce a practical guide based on international standards such as ISO 5060, ISO 17100, MQM, and the JTF guidelines in Section 4 (ISO5060:2024; ISO17100:2015; MQM, 2025; JTF, 2018). We apply this framework to a case study in Section 5.

Our case study focuses on investor relations (IR) texts from 33 publicly listed Japanese companies, where alignment with specifications is crucial. We compare five English translations: the company's official version, a proprietary MT output, a basic prompt-based LLM output, a prompt-based LLM output using specifications, and a prompt-based LLM post-edit of the MT output.[2]

We evaluate the translations through expert er-

ror analysis, user rankings, and a reference-free automatic metric. Results show that prompt-based LLM outputs with specifications receive the highest ratings from both experts and users. Official translations score lower due to stylistic shortcomings, and conventional MT outputs also underperform. In contrast, the automatic metric favors MT output, highlighting a misalignment with human judgment. This suggests that specification-aware LLM outputs better fulfill communicative goals, even if not fully reflected in current automatic metrics.[3]

Our contributions are as follows:

- We provide a theoretical foundation for incorporating translation specifications into MT and its evaluation.

- We propose a practical guide for applying specifications in MT workflows and test it on IR texts.

- We demonstrate that specification-aware prompt-based LLMs outperform official human translations in human evaluations, supported by detailed analysis.

2 Related Work

2.1 Customizable Machine Translation

Recent studies have explored how MT can be adapted to specific contexts using external knowledge, prompts, and post-editing beyond the source text.[4] Though not framed as translation specifications, these efforts share similar goals with ours.

Fujita (2021) highlights the limits of text-to-text neural machine translations (NMT), emphasizing the need for style guides, terminology, and domain knowledge. He (2024) and Jiao et al. (2024) show that prompting GPT-4 with contextual cues and post-editing improves translation quality. Liu et al. (2025) further finds that detailed, domain-specific prompts enhance performance in specialized tasks.

For stylistic and functional control, Moslem et al. (2023), Wang et al. (2023), and Yamada (2023) show that incorporating tone, terminology, and information about the translation's purpose and audience leads to more targeted outputs. Raunak et al. (2023) also demonstrate that post-editing

---

[2]Throughout this paper, *LLM* refers to large language models guided by prompt-based customization. We use *post-edit* in a broad sense, including automated revision, and not limited to human editing as defined in ISO 18587 (ISO 18587:2017, 2017).

[3]All translation data and evaluation results are available at https://github.com/nii-cl/Specification_aware_MT.

[4]For a discussion of prior work on controllable MT in the Statistical Machine Translation (SMT) and NMT eras, see Appendix B.

GPT-4 output improves English–Chinese and English–German translation quality.

These studies share motivations with our work on specification-aware translation. Our study extends these efforts by using real corporate materials and evaluating translation outputs in a practical setting. We explore the potential of prompt-based LLMs to meet specific professional translation requirements, based on human evaluation.

2.2 Advances in Translation Evaluation

Recent work in translation evaluation moves beyond gold references and explores reference-free approaches (Blain et al., 2023; Freitag et al., 2023, 2024; Zerva et al., 2024). Evaluation criteria also expand to include contextual coherence and fine-grained error types. For document-level automatic evaluation, Vernikos et al. (2022) improve sentence-level metrics by incorporating context. Jiang et al. (2022) propose BlonDe, a metric that evaluates discourse coherence using span-level F1 scores. Meta-evaluation, such as that of Moghe et al. (2025), examine whether metrics can detect diverse error types and highlight their limitations.

MQM-based automatic evaluation has gained traction: GEMBA-MQM (Kocmi and Federmann, 2023a), AutoMQM (Fernandes et al., 2023), and xCOMET (Guerreiro et al., 2024) identify error spans and types without language-specific tuning. CATER (Iida and Mimura, 2024) offers reference-free, multi-dimensional evaluation with LLMs, while MQM-APE (Lu et al., 2025) adds automatic post-editing to LLM-based error annotation to focus on quality-improving edits.

Human evaluation also remains essential. Lommel et al. (2024b) present an MQM scoring framework with calibrated models. MQM-Chat (Li et al., 2025) adapts MQM for chatbot, and ESA (Kocmi et al., 2024) streamlines span-level annotations for non-expert assessments.

Building on these developments, our study incorporates ISO 5060 and MQM principles into a specification-aware evaluation framework. To better capture translation quality, we assess translations using expert annotators, end-user judgments, and automatic metrics, highlighting both linguistic quality and functional adequacy.[5]

2.3 Translation Theory and Machine Translation

Several studies explore interactions between translation studies and MT research. Tan et al. (2023) apply Skopos-based criteria to compare human and NMT outputs, showing that human translations perform better due to NMT's contextual and lexical limitations. Liu et al. (2024) recommend integrating Skopos theory into human evaluation, while Na et al. (2024) show that theory-informed prompts affect LLM outputs. Hiebl and Gromann (2023) call for a unified concept of translation quality to support collaboration between the fields.

The point raised by Hiebl and Gromann (2023) is important: clarifying how the two fields define and evaluate translation quality may help advance both. To this end, we combine theoretical insights from translation studies with empirical experiments based on real-world workflows, aiming to explore how MT can better address the practical needs of professional translation. This integration of theory and practice enables a more realistic understanding of MT's role in professional contexts.

3 Theoretical Background

While translation is often seen as producing an equivalent text in another language, the notion of *equivalence* has faced criticism in translation studies since the late 1970s. In response, functionalist approaches have gained prominence, viewing translation as a purpose-driven communicative act. Skopos theory, a widely cited framework, holds that translations should be guided by their purpose. Based on this view, we argue that translation specifications are essential for developing and evaluating MT systems that meet real-world goals.

We begin with equivalence theory, which frames translation as reproducing the meaning or value of the source text, a view reflected in early MT systems and many current automatic evaluation methods (Section 3.1). We then turn to Skopos theory, a functionalist perspective aligned with our emphasis on translation specifications (Section 3.2). Finally, we discuss how specifications matter not only for translation but also for evaluation (Section 3.3).

3.1 Equivalence Theory in Translation Studies and Machine Translation

Equivalence theory (Nida, 1964), which views translation as reproducing the source text's meaning and value, has long been central to translation

studies.[6] Dyvik (1992) explores this concept in MT, proposing a *situation schema*, an abstract representation that links source and target texts through shared meaning. He emphasizes the importance and difficulty of achieving equivalence, even with linguistic theories and technology.

In contrast, Hardmeier (2015) analyzes how statistical machine translation (SMT) operationalizes equivalence through techniques such as word alignment and domain modeling. While SMT reflects equivalence-based assumptions, he argues it oversimplifies translation complexity.

These studies illustrate how earlier MT systems, primarily rule-based and statistical, were shaped by equivalence-oriented thinking. Although neural networks and deep learning emerged in the mid-2010s (Bahdanau et al., 2015), earlier systems dominated the field and adhered to formal equivalence.

Even with advances in MT technology, equivalence continues to shape quality evaluation. Metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) assess similarity to references, often through n-gram overlap, reflecting a formal equivalence perspective. Recent model-based metrics like COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) seek semantic equivalence, yet rely on source-text alignment.

Since the late 1970s, however, equivalence theory has faced criticism. Snell-Hornby (1995) argues that it lacks precision and falsely implies symmetry between languages. She identifies the 1980s cultural turn as a shift from language-based approaches to views considering sociocultural context and the translator's role (Snell-Hornby, 2006).[7]

Although equivalence has become less central in translation studies, it is still prominent in MT practice and evaluation. While semantic equivalence remains foundational, it does not fully address the diverse purposes and communicative contexts of real-world translation. Functionalist approaches, such as Skopos theory, offer a useful complement. Emphasizing the intended function, Skopos theory provides a more practical framework for guiding both human and MT in applied settings.

---

[6]Equivalence theory includes various perspectives, including formal equivalence, which preserves structure, and dynamic equivalence, which aims for a similar reader response (Nida, 1964; Munday et al., 2022; Pym, 2023).

[7]This shift is reflected in major anthologies such as *The Translation Studies Reader* (Venuti, 2021), whose fourth edition retains only Nida (1964) for equivalence theory, omitting figures like Vinay and Darbelnet (1958) and Catford (1965).

3.2 Skopos Theory and the Functional Approach to Translation

Skopos theory defines translation not as a linguistic transfer but as an intentional activity to fulfill communicative goals (Reiss and Vermeer, 1984). Nord (2006) develops this perspective by introducing translation brief (or specification), a set of instructions outlining the purpose, audience, and conditions for the translation. She emphasizes that translation decisions are not solely determined by the source text, but by how well it serves its function.

Gouadec (2007) applies the functionalist approach to translation workflows by identifying three criteria: the client's objectives (e.g., increasing sales or enhancing brand image), the user's needs (e.g., clarity in technical documentation), and the relevant usage norms and standards. As Pym (2023) notes, this positions translators as "language technicians" who operate within a broader communication strategy, ensuring that the translation fulfills its specific role.

This functionalist perspective, once limited to human translation, may now extend to MT with the emergence of prompt-based LLMs. Earlier domain-specific MT systems required significant resources, including specialized datasets, expert tuning, and time-consuming model training (Saunders, 2022; Wang et al., 2023). In contrast, prompt-based LLMs allow users to specify translation requirements through prompting or fine-tuning, making it easier to adapt translations to their intended purpose (Section 2.1). Although empirical evidence is still emerging, customization is now easier, and the rise of LLMs marks a technological shift that aligns with the functionalist view of translation.

Our study investigates whether and how recent advances in MT, especially prompt-based LLMs, can support a functionalist approach by producing translations aligned with specifications. This perspective shifts the focus from linguistic equivalence to functional effectiveness and offers insights for improving MT design and evaluation.

3.3 Why Specifications Matter

Specifications may include parameters such as purpose, target audience, style, register, domain, timeline, cost, volume, reference materials (e.g., glossaries and style guides), file format, and quality evaluation methods (ISO17100:2015; JTF, 2018). In professional settings, such specifications guide translation decisions and ensure the translation

	Parameter	Description
1	Purpose of translation	Communicative goal (e.g., inform, persuade, comply, etc.)
2	Target audience	Intended readers and their language background or expectations
3	Style, register, and tone	Formality, style, and tone appropriate for the target context
4	Terminology and reference resources	Use of glossaries, style guides, and prior translations
5	Domain and legal requirements	Industry norms and compliance with relevant laws
6	Cultural adaptation	Adjustments for cultural norms or sensitivities
7	Length and formatting	Constraints on text length, layout, or structure
8	Localization needs	Regional or language variant customization

Table 1: Translation specification parameters. Items 1–3 are essential; others may vary by project.

meets client expectations. Even if undocumented, essential requirements are typically agreed upon in advance and vary by project. For example, legal translations emphasize consistency with terminology and style guides, while marketing texts prioritize creativity and persuasive language.

The importance is also reflected in how translation quality is evaluated. The MQM framework defines translation quality as follows:

> A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs (Melby, 2012a; Lommel et al., 2013).

As explained in the Multi-Range Theory (Lommel et al., 2024b), quality evaluation begins with an analysis of project specifications and user needs. Evaluators should *select* appropriate error categories and scoring models based on this analysis. These ideas are emphasized in the 2024 MQM anniversary paper (Lommel et al., 2024b).[8]

Specifications not only guide translators but also constrain the range of acceptable choices, helping to reduce subjectivity. As Gouadec (2007) argues, translators are language technicians whose "plurality is his enemy," highlighting the importance of clear instructions (Pym, 2023). This applies equally to evaluation: assessments grounded in specifications are less influenced by personal interpretation.

Research shows that providing clear criteria and context improves inter-annotator agreement (IAA) (Castilho, 2021; Popović, 2021). The official MQM website also notes that the framework supports standardized, objective evaluation by minimizing subjective judgment (MQM, 2025). For details on how translation specifications can be incorporated into MQM-based evaluation, see Appendix E.

4 A Practical Guide for Specification-Aware MT and Evaluation

We provide a brief overview of our practical framework for integrating translation specifications into both MT workflows and evaluation, where MT is performed using prompt-based LLMs.[9] A full version is available in Appendix F.

4.1 Specification-Aware Machine Translation with Prompt-Based LLMs

Step 1: Define Specifications Clarify translation requirements. These form the basis for both machine output and human review. Our specification parameters, listed in Table 1, are independently developed based on professional translation practice and informed by existing standards and research (ISO17100:2015; 11669:2024; Melby, 2012b).

The top three items are essential for all translation projects, regardless of domain or medium. The remaining items are project-dependent and may be included as needed. Additional parameters may be added depending on the context or client requirements. A brief explanation and examples for each item are provided in Appendix F.1.

Step 2: Design Instructions Specifications should be reflected in prompts or fine-tuning. It is important that the instructions also include source text information, target language, and relevant specification parameters, while preventing hallucination and over-generation.

Step 3: Generate and Review Use LLMs to generate the translation, followed by human review to ensure the output meets specifications. Reviewers make corrections and finalize the translation.

---

[8]The MQM framework draws on Garvin (1984)'s approach to quality. See Appendix D for further explanation.

[9]We base our translation and evaluation guidelines on a typical professional workflow and ISO 17100 for translation, and on ISO 5060, JTF guidelines, and MQM for evaluation (ISO17100:2015; ISO5060:2024; JTF, 2018; MQM, 2025).

4.2 Specification-Aware Evaluation

Translation evaluation is not always conducted alongside the translation itself. It may be required in various contexts, such as accepting or rejecting a translation, comparing outputs, selecting the best version, ensuring quality in professional workflows, evaluating MT results, or training and certifying translators. We outline a framework that incorporates both objective and subjective evaluations.

Step 1: Make Specifications Accessible Ensure all evaluators have access to the translation specifications. If not provided in advance, define them before evaluation begins.

Step 2: Define Error Categories Set error categories (e.g., Accuracy, Style, Terminology, etc.) aligned with the specifications. Use established frameworks such as MQM and ISO 5060 (MQM, 2025; ISO5060:2024).

Step 3: Weight and Score Errors Assign weights to error categories based on project priorities. Evaluate severity (e.g., minor, major) and calculate a total score using a weighted formula.

Step 4: Add Subjective Evaluation (Optional) In addition to error-based scoring, subjective evaluation helps assess whether a translation is appropriate, persuasive, and effective for its intended audience. Feedback from experts or users can offer insights into clarity, tone, and impact that error metrics alone may overlook.

The following case study demonstrates the application of this practical guide.

5 A Case Study in Japanese-to-English Translation of Investor Relations Materials

We present a case study to show how specification-aware translation can be applied using a prompt-based LLM. We compare it with human translation and non-prompt-based MT outputs, examining how each is evaluated through both human and automatic methods. This case study puts into practice the notions discussed earlier in Section 3 and assesses the effectiveness of our approach in a real-world Japanese-to-English translation task.

Figure 2 provides an overview of the case study. For LLM, we use ChatGPT via its public interface to simulate a scenario in which non-expert users, such as translators or corporate communications personnel, can control translation output through
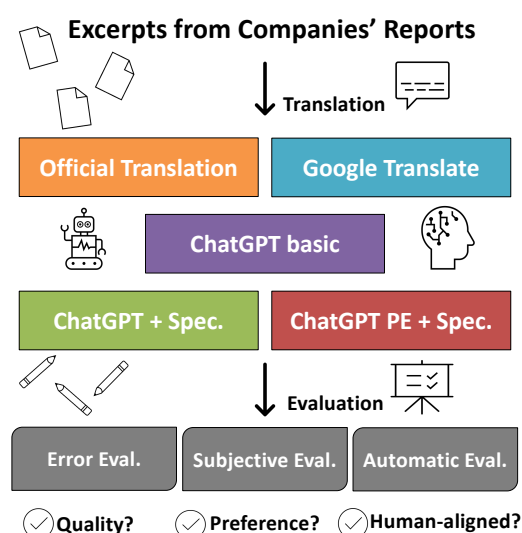

Figure 2: Overview of the case study: Five translation types and their evaluation via expert, user, and automatic methods.

prompting, without needing specialized tools or programming skills.

5.1 Experimental Setup

5.1.1 Integrated Reports

We use IR materials excerpted from integrated reports by publicly listed Japanese companies as the source text. The focus is on Japanese-to-English translation; the rationale for using this language pair is explained in Appendix G.

Integrated reports combine financial and non-financial information to communicate a company's value creation to investors and other stakeholders. Although not legally required in Japan, their publication has increased with growing interest in ESG investment. As of the end of 2023, 1,019 companies issue integrated reports, 70 percent of which also provide English versions (ESG/Integrated Reporting Research Laboratory, 2024). Among Prime Market companies, over half publish integrated reports. We choose integrated reports because they are more structured than websites and more interpretive than financial statements, posing challenges for both human translators and MT systems.

We focus on the corporate philosophy section, typically found at the beginning of these reports. According to the *Guidance for Collaborative Value Creation 2.0* (Ministry of Economy, Trade and Industry, 2022), such statements are central to investor communication and must clearly express a company's unique values. Translating them re-

Type	Translation Output
Source	"ワクワク"は，人を動かすエネルギー. それは人から人へと伝わり、世界をあかるく元気にする。
Official	"Waku waku" is what moves people to push what's possible. It's Japanese for the joy and excitement of discovering the unknown. And when passed from person to person, becomes a force that creates a brighter world, united in wonder.
Google	"Excitement" is the energy that moves people. It spreads from person to person, making the world brighter and more energetic.
GPT-b	"Excitement" is the energy that moves people. It spreads from person to person, bringing brightness and vitality to the world.
GPT+Sp	Excitement is the energy that moves people. It spreads from person to person, brightening and invigorating the world.
PE+Sp	"Excitement" is the spark that moves people, spreading from one person to another, brightening and energizing the world.

Table 2: Differences in translations (All Nippon Airways Co., Ltd.)

quires not only literal accuracy but also clarity, an appropriate corporate tone, and expressions that enhance appeal to stakeholders.

We select one company from each of the 33 industries defined by the Tokyo Stock Exchange (Japan Exchange Group, Inc., 2021), prioritizing those with higher market capitalization (27 first-ranked, five second-ranked, and one fourth-ranked).[10] The extracted sections range from 240 to 927 Japanese characters, with an average of 610.

We manually confirmed alignment with the Japanese source texts and asked companies how their English versions were produced. Of the 33 companies, 20 responded. Among these, 15 used only human translation, two combined human and MT, two declined to disclose their method, and one outsourced the work without providing details. None reported using MT alone, suggesting that human translation remains standard.

5.1.2 Five Translation Methods

To compare translation quality and effectiveness, we prepare five versions using different methods. The official translation consists of excerpts from English versions of integrated reports published by the companies.

We then create four MT-based versions:

- **Google Translate**: raw output from Google Translate

- **ChatGPT basic**: ChatGPT with a minimal prompt

- **ChatGPT + Spec**: ChatGPT with specifications

- **ChatGPT PE + Spec**: Google Translate post-edited by ChatGPT with specifications

To ensure consistency, we use the first output for all versions. All ChatGPT translations are generated using ChatGPT-4o.

---

[10]The full list appears in Appendix H.

For the specification-aware methods, we provide prompts that reflect key information such as source text context, intended purpose (e.g., appealing to global investors), target audience, and stylistic tone. The full prompt is shown in Appendix I.

Using these methods, we generate five translations for each of the 33 companies. Manual review indicates that all versions maintain overall meaning without serious accuracy errors. However, we observe a recurring issue in ChatGPT translations: kanji misinterpretation. For example, 文殊院旨意書 (Monjuin Shiigaki) is rendered incorrectly as *Monjuin Shiisho*, lacking accurate transliteration. This suggests that ChatGPT struggles with domain-specific terminology and proper nouns.

Table 2 shows translations of a corporate philosophy excerpt from All Nippon Airways Co., Ltd.'s integrated report. The official translation contains a grammatical error ("And when passed... becomes...") and phrases that may be unclear or unnatural ("to push what's possible"). It also gives an extended explanation of *waku waku*. The Google translation is grammatically correct but closely mirrors the source, resulting in a literal tone and basic vocabulary. The ChatGPT basic version improves fluency and uses slightly richer expressions ("vitality"), but its tone and structure remain similar to the Google version. The ChatGPT version with specifications uses more active verbs and parallel phrasing ("brightening and invigorating"), resulting in smoother rhythm and tone. The post-edited version with specifications introduces vocabulary like "spark" and "energizing," while preserving the original meaning and structure.

These examples show that each method yields distinct results and that adding specifications to ChatGPT prompts may encourage more purposeful and expressive language. Appendix J provides a comparative analysis of a longer excerpt from the same source, focusing on linguistic and stylistic differences.

5.2 Human Evaluation

After preparing translations for all 33 companies, we conduct two human evaluations of the five translation methods: expert error evaluation and subjective evaluation.

5.2.1 Error Evaluation

We conduct an error-based evaluation using the specification-aware framework introduced in Section 4.2 and detailed further in Appendix F.2. This evaluation focuses on three core categories: Accuracy, Linguistic Conventions, and Style. Other categories defined in the MQM framework, such as Design and Markup, are excluded as they are not applicable to the scope of this study.

All category definitions are based on the MQM standard and were provided to the evaluators to ensure consistency and shared understanding (MQM, 2025). See Appendix K for the detailed error categories used in the annotation.

Given the importance of stylistic quality in IR materials, we include four subtypes under Style: (1) Language register mismatch, (2) Awkward style, (3) Unidiomatic expressions, and (4) Inconsistent style. These errors do not hinder comprehension but result in unnatural English that may reduce clarity and impact. Subtypes help clarify scope, but annotators classify errors only at the main category level to reduce cognitive burden. Error categories are weighted based on JTF guidelines: Accuracy (0.7), Linguistic Conventions (0.8), and Style (1.5), averaging to 1.0 overall (JTF, 2018). We do not apply severity levels, as the texts do not involve high-stakes content such as financial figures.

Two professional evaluators, each either a professional translator or an expert in linguistics and culture, bilingual in Japanese and English, and with English as their first language, are recruited via Prolific.[11] They receive the Japanese source text, translation specifications, five anonymized English translations, an error typology table with definitions, and sample annotations. They identify errors, assign them to one of the three categories, mark their locations, and record error counts. Each evaluator is compensated £40 for approximately 270 minutes of work. Only two out of 24 recruited participants completed the task, highlighting the practical difficulty of securing qualified evaluators and the cognitive demands of error annotation, as noted in prior research (Kocmi et al., 2024; Zouhar

Type	Official	Google	GPT-b	GPT+Sp	GPT PE+Sp
Eval. 1	*2.60*	1.82	1.04	0.70	**0.38**
Eval. 2	*3.01*	2.29	1.28	1.29	**1.03**

Table 3: Weighted error scores averaged across 33 companies. Lower scores indicate fewer errors and higher translation quality.

et al., 2025).

Table 3 presents the evaluation results. ChatGPT PE + Spec receives the lowest error score (highest quality), followed by ChatGPT + Spec, ChatGPT basic, and Google Translate. The official translation ranks lowest, with particularly frequent Style errors, which will be discussed in Section 5.2.3. These findings suggest that LLM-based translations guided by specifications can outperform human translations in this context, challenging the assumption that human translations should serve as the default gold standard in MT evaluation.

We also assessed inter-annotator reliability by calculating the correlation between the error scores assigned by the two evaluators. Pearson's correlation is very high ($r = 0.985$ and $p = 0.0021$), while Spearman's rank correlation is also strong ($\rho = 0.90$ and $p = 0.037$), indicating statistically significant agreement. Nonetheless, we observe inconsistencies: the same expression was sometimes marked as an error in one translation but not in another by the same evaluator. This indicates the inherent difficulty of ensuring consistency in error-based evaluation, even among professionals. The low completion rate suggests that translation evaluation is time-consuming, cognitively demanding, and difficult to delegate, as it requires a high level of expertise. Our evaluation process incidentally reflected these challenges in practice.

5.2.2 Subjective Evaluation

To understand how translations are perceived by intended end users, we conduct a subjective evaluation alongside expert-based error analysis. As discussed in Section 4.2 and detailed in Appendix F.3, combining error-based and subjective evaluation is useful not only when qualified annotators are limited, but also when end-user perspectives take precedence. For texts like integrated reports, which aim to build trust and attract investment, reader impression may matter more than linguistic accuracy, making subjective feedback particularly valuable.

Subjective evaluation is generally divided into expert and end-user perspectives (JTF, 2018).
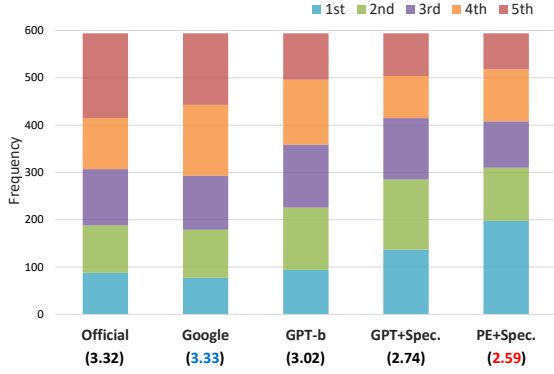
---

[11]https://www.prolific.com

Figure 3: Ranking counts from the subjective evaluation. The x-axis shows translation methods, and the y-axis shows frequency. Each bar is stacked by rank (1st to 5th). Numbers in parentheses indicate mean rankings; lower values reflect higher preference.

Since integrated reports target investors, we adopt an end-user perspective. We recruited eighteen native English speakers via Prolific. Participants are compensated £13.50 for approximately 90 minutes of work. Seventeen hold degrees in fields such as accounting or finance. One participant is a translator and linguistic expert who also took part in the error evaluation. Each evaluator receives the translation specifications and five English translations, without knowing the translation methods or having access to the Japanese source. They are asked to rank the translations based on overall appeal, defined as clarity, readability, word choice, and company presentation.

Figure 3 shows the distribution of rankings across translation types. ChatGPT PE + Spec is most often ranked 1st, whereas the official translation is most often ranked 5th. Google Translate is the least often ranked 1st, while the three ChatGPT-based translations are less often ranked last. Numbers in parentheses indicate mean rankings: ChatGPT PE + Spec has the best (lowest) average ranking, followed by ChatGPT + Spec, ChatGPT basic, and the official translation. Google Translate ranks lowest overall.

To assess significance, we conduct Wilcoxon signed-rank tests on all ten translation pairs, reporting the test statistic (W), p-values, and effect sizes (r) in Table 4. A p-value below 0.05 is considered significant; effect sizes are interpreted as small ($r = 0.1$), medium ($r = 0.3$), or large ($r = 0.5$). All comparisons are significant except Official vs. Google Translate and ChatGPT + Spec vs. Chat-GPT PE + Spec. Pairs with ChatGPT PE + Spec (vs.

Pair	W	Z	p	r
Off. vs Ggl	88018	0.081	0.9341	0.003
Off. vs GPT-b	74913	3.213	0.00113	0.132
Off. vs GPT+Sp	**63954**	**5.832**	$p < 0.00001$	**0.239**
Off. vs PE+Sp	**59011**	**7.013**	$p < 0.00001$	**0.288**
Ggl vs GPT-b	73843	3.469	0.00043	0.142
Ggl vs GPT+Sp	**62409**	**6.201**	$p < 0.00001$	**0.254**
Ggl vs PE+Sp	**57082.5**	**7.474**	$p < 0.00001$	**0.307**
GPT-b vs GPT+Sp	75367	3.104	0.00162	0.127
GPT-b vs PE+Sp	68326.5	4.787	$p < 0.00001$	0.196
GPT+Sp vs PE+Sp	81396.5	1.664	0.0903	0.068

Table 4: Wilcoxon signed-rank test for translation pairs. A significant difference is defined as $p < 0.05$. Effect size (r) is interpreted as small (0.1), medium (0.3), and large (0.5).

Official and Google Translate) show the strongest effects, with $p \approx 0$ and medium effect sizes.

These results show that ChatGPT PE + Spec is consistently preferred, in line with error-based evaluation findings. The low ranking of the official translation is notable, despite its presumed status as the gold standard. However, human translations often vary more than MT, depending on translator performance (Freitag et al., 2023; Ramos and Guzmán, 2024; Volz and von Thiessen, 2024). As a result, while the official translation was most frequently ranked in the lowest position, the number of times it was placed second, third, or fourth did not differ significantly. Moreover, it was ranked first more often than Google Translate and not far behind ChatGPT basic.

5.2.3 Qualitative Analysis

To gain insight into the stylistic and structural differences across translation types, we examine their sentence structure, focusing on relative clauses and clausal coordination. Our analysis shows that Google Translate and the official translations tend to use these forms more frequently, potentially reflecting source-language influence. Japanese allows for long, additive sentence constructions, which can lead to overuse of relative clauses or clausal coordination when translated too literally into English. Such structures may reduce readability, especially in English writing that values clarity and conciseness. Further details are provided in Appendix L.

We also analyze excerpts from the official translations and find recurring issues in grammar, style, and semantic clarity. For example, the expression "offering both a multitude of choices" contains a semantic mismatch between "both" and "multitude."

Type	Official	Google	GPT	GPT+Sp	GPT PE+Sp
Mean	*0.783*	**0.830**	0.822	0.821	0.810
SD	0.043	0.031	0.039	0.033	0.037

Table 5: Mean COMETKiwi scores and standard deviations for each translation type.

Other examples involved unidiomatic phrasing, sentence fragments, and inconsistent style. For detailed examples and qualitative error analysis, see Appendix M. These problems suggest that the low rating of the official translation may not stem from a lack of specifications but from variation in translator skill or mismatches with task requirements.

As mentioned in Section 5.2.2, human translations often vary in quality due to individual differences (Freitag et al., 2023; Ramos and Guzmán, 2024; Volz and von Thiessen, 2024). Combined with Japan's shortage of high-proficiency English translators (Appendix G), this may explain the observed results. By contrast, ChatGPT-based translations guided by specifications performed consistently well, suggesting their potential as a viable complement to traditional workflows.

5.3 Automatic Evaluation

We examine whether a reference-free automatic metric can capture differences in translation quality across specification and method types, compared to human judgment. To this end, we use COMETKiwi, a reference-free metric with the highest correlation to human evaluations in the WMT23 Metrics Shared Task (Rei et al., 2022; Freitag et al., 2023).[12] We adopt a reference-free approach because the official translations, typically used as references, are themselves part of the evaluation as one of the five translation types.

Table 5 shows scores from 0 to 1, with higher values indicating better quality. Low standard deviations suggest internal consistency, though overlapping ranges point to limited differences between types. Unlike the human rankings, COMETKiwi assigns the highest score to Google Translate and the lowest to the official translation. This divergence likely reflects differences in what COMET values, specifically literal fidelity and lexical similarity, as opposed to the more context-sensitive and stylistic qualities emphasized in our evaluation (Rei et al., 2022).

Although the official translation appears to preserve source-like structures such as relative clauses and clausal coordination (Section 5.2.3, Appendix L), its low score may be partly explained by a few explanatory additions not present in the source, intended to assist international readers. Such additions may reduce source alignment and result in lower automatic scores.

ChatGPT PE + Spec scores slightly below Google Translate, though the difference is small. This may reflect Google Translate's more literal style, while ChatGPT PE + Spec balances fidelity and fluency, resulting in higher subjective appeal despite a lower COMET score. ChatGPT translations, particularly those guided by specifications, prioritize clarity and appeal over strict lexical matching, which COMET may not fully capture.

Although COMET metrics are known to struggle with numbers and named entities (Amrhein and Sennrich, 2022), our manual check found no significant errors in these areas, suggesting they did not affect the results.

As MT evaluation increasingly considers contextual and communicative goals, it is vital to develop automatic metrics that better capture functional aspects of translation quality, such as how well a translation fulfills its purpose in context.

6 Conclusion

We demonstrate that translation specifications can improve MT quality and enable more targeted evaluation. We provide a theoretical rationale for the importance of specifications, drawing on Skopos theory to support a functionalist perspective. Based on this foundation, we outline a practical guide for specification-aware MT using LLMs, including prompt design, generation, and both error-based and subjective evaluation.

In our case study, LLM outputs guided by specifications received higher ratings than official translations, Google Translate, or unguided LLM outputs. Although COMET scores favored Google Translate, they diverged from human evaluations. These findings suggest that specifications help LLMs produce more contextually appropriate translations that better align with communicative goals. The gap between human and automatic evaluations highlights the limitations of current metrics in capturing functional adequacy. Through this work, we demonstrate the potential of specification-aware MT for professional, real-world use cases.

---

[12]We use the model `wmt22-cometkiwi-da`, also adopted as the WMT24 baseline for reference-free evaluation (Freitag et al., 2024).

Limitations

First, we use only a single LLM, ChatGPT. While its outputs are generally well-received, it occasionally introduces information not present in the source text. This highlights the importance of careful prompt design and human oversight, as is standard in professional translation workflows. Evaluating other LLMs remains an important area for future research to assess whether the findings generalize across models.

Second, our dataset consists of corporate philosophy statements from 33 Japanese companies, focusing solely on the Japanese-to-English language pair. While this allowed for a focused case study, broader validation will require larger datasets covering more diverse domains (e.g., legal and medical) and content types (e.g., marketing and technical manuals), as well as other language pairs.

Finally, our human evaluation process highlighted the difficulty of securing qualified annotators. Both the error analysis and the subjective evaluation were conducted through crowd-sourcing, and the compensation was set above the standard rates of that framework. For error analysis, which requires more specialized expertise, an alternative approach could have been to recruit evaluators through a more specialized platform and set the compensation accordingly. Such difficulties in recruiting and compensating qualified annotators emphasize the need to develop an automated and specification-based evaluation model. Future work could explore the *LLM as a Judge* (Zheng et al., 2023; Kocmi and Federmann, 2023b; Feng et al., 2025; Gunathilaka and de Silva, 2025), where an LLM evaluates outputs based on the same detailed specifications provided to human experts, potentially offering a scalable alternative to manual annotation.

Acknowledgments

References

ISO 11669:2024. ISO 11669:2024. Translation projects – General guidance. https://www.iso.org/standard/79089.html. Accessed: 2025-07-29.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *The Third International Conference on Learning Representations*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M Guerreiro, Diptesh Kanojia, José GC de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, et al. 2023. Findings of the wmt 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Michael Carl, Akiko Aizawa, and Masaru Yamada. 2016. English-to-Japanese translation vs. dictation vs. post-editing: Comparing translation modes in a multilingual setting. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC‘16)*, pages 4024–4031, Portorož, Slovenia. European Language Resources Association (ELRA).

Sheila Castilho. 2021. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of*

the Workshop on Human Evaluation of NLP Systems (HumEval), pages 34–45, Online. Association for Computational Linguistics.

J. C. Catford. 1965. *Translation Shifts*, pages 73–82. Oxford University Press, London. Originally published as part of the series "Language and Language Learning" (Vol. 8).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Directorate-General for Translation (DGT), European Commission. 2015. DGT translation quality guidelines. http://ec.europa.eu/translation/maltese/guidelines/documents/dgt_translation_quality_guidelines_en.pdf.

Helge Dyvik. 1992. Linguistics and machine translation. In *Proceedings of the 8th Nordic Conference of Computational Linguistics (NODALIDA 1991)*, pages 67–78, Bergen, Norway. Norwegian Computing Centre for the Humanities, Norway.

ESG/Integrated Reporting Research Laboratory. 2024. 「統合報告書発行状況調査2023」最終報告（final report on the survey of integrated report publication status 2023）.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107, Vienna, Austria. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Paul Fields, Daryl Hague, Geoffrey Koby, Arle Lommel, and Alan Melby. 2014. What is quality? a management discipline and the translation industry get acquainted. *Tradumàtica: tecnologies de la traducció*, page 404.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Atsushi Fujita. 2021. Attainable text-to-text machine translation vs. translation: Issues beyond linguistic processing. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 215–230, Virtual. Association for Machine Translation in the Americas.

David Garvin. 1984. What does "product quality"' really mean? *MIT Sloan Management Review*, 26:25–43.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Daniel Gouadec. 2007. *Translation as a Profession*, volume 73 of *Benjamins Translation Library*. John Benjamins Publishing.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

S. Gunathilaka and N. de Silva. 2025. Automatic analysis of app reviews using llms. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 828–839. SciTePress.

Christian Hardmeier. 2015. On statistical machine translation and translation theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 168–172, Lisbon, Portugal. Association for Computational Linguistics.

Douglas Harper. 2025. explore | etymology, origin and meaning of explore by etymonline. Accessed: 2025-04-30.

Sui He. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 316–326, Sheffield, UK. European Association for Machine Translation (EAMT).

Bettina Hiebl and Dagmar Gromann. 2023. Quality in human and machine translation: An interdisciplinary survey. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 375–384, Tampere, Finland. European Association for Machine Translation.

Kurando Iida and Kenjiro Mimura. 2024. Cater: Leveraging llm to pioneer a multidimensional, reference-independent paradigm in translation quality evaluation.

ISO 18587:2017. 2017. Translation services – post-editing of machine translation output – requirements (iso standard no. 18587:2017).

ISO17100:2015. Translation services – requirements for translation services (iso standard no. 17100:2015).

ISO5060:2024. Translation services – evaluation of translation output – general guidance (iso standard no. 5060:2024).

Japan Exchange Group, Inc. 2021. Topix sector indices / topix-17 series.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565,

Seattle, United States. Association for Computational Linguistics.

Hui Jiao, Bei Peng, Lu Zong, Xiaojun Zhang, and Xinwei Li. 2024. Gradable chatgpt translation evaluation.

JTF. 2018. JTF翻訳品質評価ガイドライン第1版 (Japan Translation Federation translation quality evaluation guidelines first edition).

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA. Association for Machine Translation in the Americas.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. 2025. MQM-chat: Multi-dimensional quality metrics for chat translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3283–3299, Abu Dhabi, UAE. Association for Computational Linguistics.

Qibang Liu, Wenzhe Wang, and Jeffrey Willard. 2025. Effects of prompt length on domain-specific tasks for large language models.

Ting Liu, Chi-kiu Lo, Elizabeth Marshman, and Rebecca Knowles. 2024. Evaluation briefs: Drawing on translation studies for human evaluation of MT. In

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 190–208, Chicago, USA. Association for Machine Translation in the Americas.

Arle Lommel, Donald A. DePalma, and Tahar Bouhafs. 2024a. Q3 2024 language services market sizing update. Retrieved from https://csa-research.com.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024b. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

A Melby. 2012a. Human and machine translation quality: Definable? achievable? desirable. In *LACUS Forum*, volume 39, pages 1–29.

Alan K. Melby. 2012b. Structured specifications and translation parameters (ttt.org specs). https://www.ttt.org/specs/#1b. Accessed: 2025-07-29.

Merriam-Webster Dictionary. 2025. Encounter. Accessed: 2025-04-30.

Ministry of Economy, Trade and Industry. 2022. Guidance for integrated corporate disclosure and company-investor dialogue for collaborative value creation 2.0.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

MQM. 2025. MQM (multidimensional quality metrics).

Jeremy Munday, Sara Ramos Pinto, and Jonathan Blakesley. 2022. *Introducing Translation Studies: Theories and Applications*, 5th edition. Routledge, London and New York.

Hongbin Na, Zimu Wang, Mieradilijiang Maimaiti, Tong Chen, Wei Wang, Tao Shen, and Ling Chen. 2024. Rethinking human-like translation strategy: Integrating drift-diffusion model with large language models for machine translation.

Kazuaki Nagata. 2025. Japanese companies rush to up english-language disclosures in 2025. *The Japan Times*.

Eugene A. Nida. 1964. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill, Leiden.

Iftitahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8568–8575.

Christiane Nord. 2006. Translating as a purposeful activity: A prospective approach. *TEFLIN Journal*, 17(2):131–143.

Haruka Ogawa. 2021. *Difficulty in English-Japanese translation: Cognitive effort and text/translator characteristics*. Ph.D. thesis, Kent State University.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson,

Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Anthony Pym. 2023. *Exploring translation theories*. Routledge.

Fernando Prieto Ramos and Diego Guzmán. 2024. The impact of specialised translator training and professional experience on legal translation quality assurance: an empirical study of revision performance. *The Interpreter and Translator Trainer*, 18(2):313–337. PMID: 38812808.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Katharina Reiss and Hans J. Vermeer. 1984. *Grundlegung einer allgemeinen Translationstheorie*. Niemeyer, Tübingen. Translated into Spanish by S. García Reina and C. Martín de León as *Fundamentos para una teoría funcional de la traducción*, Madrid: Akal, 1996; translated into English by C. Nord and M. Dudenhöfer as *Towards a General Theory of Translational Action*, London and New York: Routledge, 2013.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. Defining and detecting vulnerability in human evaluation guidelines: A preliminary study towards reliable NLG evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7965–7989, Mexico City, Mexico. Association for Computational Linguistics.

Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *J. Artif. Int. Res.*, 75.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Mary Snell-Hornby. 1995. *Translation Studies: An Integrated Approach*, 2nd edition. John Benjamins, Amsterdam.

Mary Snell-Hornby. 2006. *The Turns of Translation Studies: New Paradigms or Shifting Viewpoints?* John Benjamins, Amsterdam.

Ingemar Strandvik. 2017. *Evaluation of outsourced translations. State of Play in the European Commission's Directorate-General for Translation (DGT)*. Language Science Press.

Runjia Tan, Xiang Long, and Oluwatoba O. Bamigbade. 2023. Comparative research on machine translation and human translation of examples in dictionary from the perspective of skopos theory. In *Proceedings of the 3rd International Conference on Internet, Education and Information Technology (IEIT 2023)*, pages 342–353. Atlantis Press.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Takeshi Ueno. 2025. Superior information communication and documentation. *The Worldfolio*.

Lawrence Venuti, editor. 2021. *The Translation Studies Reader*, 4th edition. Routledge, London and New York.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi,

1124

United Arab Emirates (Hybrid). Association for Computational Linguistics.

J. P. Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: méthode de traduction*, 1972 edition. Didier, Paris. Translated by J. C. Sager and M. J. Hamel as *Comparative Stylistics of French and English*, Amsterdam and Philadelphia, PA: John Benjamins, 1995.

Stephanie Volz and Raphael von Thiessen. 2024. Machine translation – recommendations for public administration. White paper. Canton of Zurich, Division of Business and Economic Development.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023. Controlling styles in neural machine translation with activation prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620, Toronto, Canada. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.

Masaru Yamada. 2023. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. Ai-assisted human evaluation of machine translation. In *Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New Mexico. Association for Computational Linguistics.

A   Industry Trends from CSA Research

According to Lommel et al. (2024a) from CSA Research, an independent research firm specializing in the language services industry, language service providers (LSPs) that rely heavily on traditional human translation are experiencing declining performance. At the same time, while the growing demand for translation has outpaced the capacity of human translators, the increased use of MT alone has not led to sustainable growth (Lommel et al., 2024a).

The report encourages LSPs to shift their focus from merely producing high-quality translations to delivering greater value, such as providing cultural adaptation, adapting content for specific audiences, and training and customizing LLMs for domain-specific communication. It clearly states: "LSPs must focus on messaging that resonates with enterprise goals, and demonstrate that they use technology to achieve them (Lommel et al., 2024a)."

In response to these challenges, our study proposes a framework for specification-aware MT and evaluation.

B   Historical Context of Controllable MT

As noted in Section 2.1, the goal of tailoring translation output is not new. In the eras of SMT and NMT, significant research focused on incorporating external knowledge to control specific aspects of translation.

For example, a major line of work involved leveraging existing human translations to improve consistency. This began with the convergence of Translation Memories and SMT (Koehn and Senellart, 2010) and was later adapted to NMT, such as through neural fuzzy repair mechanisms (Bulte and Tezcan, 2019). Other approaches focused on controlling discrete linguistic features, including methods to enforce terminology constraints during NMT decoding (Dinu et al., 2019) and to manage stylistic aspects such as formality (Niu and Carpuat, 2020).

While these approaches provided powerful control over discrete phenomena, they often required specialized data preparation or model retraining.

Our work builds upon this tradition but explores how prompt-based LLMs can manage a broader set of communicative specifications in a more flexible manner.

C Reliability of Evaluation Methods

Recent research has highlighted the need for more robust and transparent evaluation methods across NLP tasks, including but not limited to MT. This includes a growing interest in developing evaluation frameworks that are comprehensive, detailed, and interpretable. For instance, Nimah et al. (2023) propose the *Metric Preference Checklist*, an analytical framework that evaluates automatic natural language generation (NLG) metrics from five distinct perspectives, providing a more multifaceted evaluation of their alignment with human judgments.

In contrast, Xiao et al. (2023) point out that most research focuses only on how well metrics correlate with human ratings, often overlooking the reliability and measurement error of the metrics themselves. They argue that concepts from Measurement Theory, used in educational and psychological testing, should be applied to NLG evaluation to better assess the reliability and validity of evaluation metrics.

Gehrmann et al. (2023) outline a long-term agenda for improving NLG evaluation, including robust human evaluation protocols and the development of metrics that go beyond surface-level overlap. Meanwhile, Ruan et al. (2024) emphasize the low reliability of human evaluation guidelines in NLG, showing that only 29.84 percent of 3,233 papers in major NLP conferences shared their guidelines, and 77.09 percent of those contained some kind of vulnerability. They propose principles for more reliable guideline design and introduce a method using LLMs to detect guideline flaws.

These concerns are relevant to evaluation in MT. If the evaluation procedures are unclear, the reliability of the results cannot be ensured. In our study, we develop guidelines for both translation and evaluation and describe the experimental procedure based on these guidelines. In addition, instead of simply reporting the results, we offer analysis and possible interpretations for each finding to improve clarity and transparency.

D Garvin's Approach to Understanding Quality

Since the MQM framework is grounded in Garvin (1984)'s approach to quality, we provide a brief overview of his perspectives.

Garvin (1984), a prominent scholar in quality management, introduces five approaches to understanding quality: transcendent, product-based, user-based, manufacturing-based, and value-based. Among these, the manufacturing-based approach defines quality as meeting pre-set specifications, and the user-based approach emphasizes satisfying user needs.

Fields et al. (2014) discuss the importance of incorporating Garvin (1984)'s approach into translation quality assessment. Although the definition of translation quality is debated and Fields et al. (2014) disagree on some points, they generally agree that the production-based approach (originally "manufacturing-based" in Garvin's words), evaluating translation according to specifications, is important.

E Incorporating Specifications in MT and its Evaluation: The Case of MQM

Current research using MQM in MT evaluation often focuses on detailed error assessment. For example, the MQM framework has been used in shared tasks at the Conference on Machine Translation (WMT), such as the General Translation Task, the Metrics Task, and the Quality Estimation Shared Task.

However, specifications are not explicitly addressed in these tasks. This is because the purpose of the General Translation Task is to evaluate general MT capabilities, which may not require specific requirements, such as the purpose of the translation or the target audience, to translate the source text. This kind of general MT capability can be useful when users simply want to gain a general understanding of foreign content. However, if the goal is to communicate clearly in the target language, the translation must convey the message in a way that reflects its purpose and suits the intended reader. This cannot be done without specifications.

Freitag et al. (2021) report that in MQM evaluations of English-to-German and Chinese-to-English translations, approximately 80 percent of errors fall into the accuracy and fluency categories, with accuracy-related mistranslations being the most common. While accuracy and fluency are

relatively straightforward to assess, other error categories may be more difficult to judge without specifications. For example, the definition of the *Style* error is simply "Translation has stylistic problems" (Freitag et al., 2021). While this is distinct from *Fluency-grammar* errors, it may be difficult for evaluators to identify a stylistic issue if the purpose of the translation is not known.

In specification-based evaluation, if the purpose of the translation is to convey the cultural otherness of the source text and the specified style is a literal translation that closely follows the original, the translation should adhere to that style. In this case, the translation is expected to preserve the expressions and cultural markers of the source to maintain the visibility of the translation. In other words, even a natural and fluent translation may be considered an error if it minimizes the sense of translation under such specifications.

As mentioned earlier, specifications can also serve as a guide for evaluators, helping to reduce subjectivity in the evaluation process. In this sense, the detailed error types in MQM could be applied more effectively when used in conjunction with detailed specifications.

Several recent MQM-based approaches, such as GEMBA-MQM (Kocmi and Federmann, 2023a), Auto-MQM (Fernandes et al., 2023), xCOMET (Guerreiro et al., 2024), and MQM-APE (Lu et al., 2025), have laid important groundwork for the automation of fine-grained evaluation. However, translation specifications are not the central focus of these approaches.

While human evaluation approaches like MQM-Chat adapt MQM error categories to specific settings (e.g., chatbot applications), they do so by simply modifying the original MQM typology (Li et al., 2025). Their approach could be further extended to integrate detailed specifications. For example, it could require translations to preserve source-text ambiguity as a stylistic feature or to handle internet slang accurately as part of terminology, in line with MQM's original design.

These considerations highlight that MQM, while widely used, can be made more effective when applied in combination with translation specifications.

F A Practical Guide for Specification-Aware Machine Translation and Evaluation

We present guidelines for specification-aware translation and evaluation, drawing on the translation process in a typical professional workflow and the "Production Process" outlined in ISO 17100 ─Requirements for Translation Services. This standard states that translation should be "in accordance with the purpose of the translation project, including the linguistic conventions of the target language and relevant project specifications" (ISO17100:2015).

Our evaluation method draws on ISO 5060, the JTF guidelines, and the MQM framework (ISO5060:2024; JTF, 2018; MQM, 2025).[13] Our approach combines these frameworks to make use of their different strengths in evaluation. MQM provides detailed error types for fine-grained analysis, while ISO 5060 and the JTF guidelines allow for weighted error categories based on specifications.[14] The JTF guidelines also emphasize the role of subjective evaluation (JTF, 2018). These evaluation guidelines are designed for use with both human and MT and assume that human evaluators will assess the output.

We first explain the method of specification-aware translation with prompt-based LLMs (Appendix F.1), followed by a description of the human error analysis procedure (Appendix F.2). It also addresses subjective evaluation, which provides a different perspective from error analysis for assessing whether the translation fulfills its intended purpose (Section F.3).

F.1 Specification-aware Machine Translations with Prompt-Based LLMs

Step 1: Define Translation Specifications Define detailed translation specifications. When working with clients or translation teams, all stakeholders should reach an agreement on the specifications. The following are examples corresponding to the parameters listed in Table 1, using a single translation project as context: the English translation of

---

[13]The European Commission's Directorate-General for Translation also provides quality evaluation guidelines based on a "fit for purpose" approach (Strandvik, 2017). Translations are assessed using error categories (e.g., Accuracy, Terminology, Linguistic Conventions, Style, and Formatting) with severity levels to calculate a quality score (Directorate-General for Translation (DGT), European Commission, 2015).

[14]The error classification used in ISO 5060 is based on the MQM framework (ISO5060:2024).

an integrated report published by a publicly listed Japanese company.

1. **Purpose of translation:** The translation's communicative goal (e.g., to inform, persuade, promote, or comply).

   Example: The goal is to attract foreign institutional investors. Therefore, the translation must emphasize growth potential and sustainability strategies in persuasive, investor-friendly language.

2. **Target audience:** The intended readers and their background knowledge, expectations, or needs.

   Example: The audience consists of non-Japanese institutional investors who may not be familiar with Japanese corporate structures. Key terms may require explanatory phrasing.

3. **Style, register and tone:** The desired level of formality, stylistic conventions, and voice appropriate to the target context.

   Example: A moderately formal and confident tone is preferred, neither overly technical nor overly casual, consistent with ESG reports by global competitors.

4. **Terminology and reference resources:** Required use of specific terms, glossaries, or past translations to ensure consistency.

   Example: The company previously translated 企業理念 as *Corporate Philosophy*, and this terminology should be maintained consistently across sections and future documents.

5. **Domain and legal requirements:** Industry-specific norms or legal constraints that affect wording or structure.

   Example: The report includes financial statements that must conform to IFRS terminology, and disclosures must reflect the Japan Financial Services Agency's guidelines on non-financial reporting.

6. **Cultural adaptation:** Modifications made to accommodate cultural expectations or sensitivities.

   Example: A phrase like *wa no seishin (*和の精神*)* may be unfamiliar to global readers and can be replaced with a culturally adapted equivalent such as *a spirit of harmony and mutual respect*.

7. **Length and formatting:** Constraints or allowances regarding text length, layout, or formatting elements.

   Example: Since the English translation must fit within the same layout as the Japanese version, sentence length and paragraph structure must be carefully managed to avoid overflow.

8. **Localization needs:** Adjustments for language variants, regional conventions, or localized preferences.

   Example: Dates, currencies, and units should follow international conventions (e.g., *FY2023* instead of *2023年度*, *million yen* instead of 百万円), and U.S. spelling is preferred for the target audience.

As these specifications demonstrate, creating appropriate translations requires numerous decisions. Defining these requirements in advance helps ensure that translations meet their intended purpose and are suitable for the target context.

Step 2: Define Roles Separate tasks performed by the machine, such as generating translations based on the given specifications, from those handled by humans, including supervision, quality assurance, verification of specification adherence, and the management of responsibilities such as meeting deadlines. Human reviewers oversee the overall process and focus on elements requiring expert judgment or domain-specific knowledge.

Step 3: Design Instructions Aligned with Specifications Create instructions for the LLM that reflect the defined specifications. Use the listed specifications to identify aspects that can be included in the model instructions. These instructions may include parameters such as:

- Source text information
- Target language
- Purpose of translation
- Target audience
- Style, register, and tone
- Length and formatting

These parameters may vary depending on the specific translation project. The instructions should also state that LLM must not add any information

that is not present in the original text. This is especially important in creative translation tasks, where the model may over-generate and introduce information not found in the source text.

Fine-tuning may be applied as needed, such as specifying terminology, aligning with existing translations, using consistent phrasing, maintaining preferred styles, or handling domain-specific vocabulary accurately.

Step 4: Generate Translation   Use the instructions to generate the initial translation.

Step 5: Review and Finalization   A human reviewer, or a team of reviewers, must carefully check whether the initial translation is accurate and ensure that it follows the defined specifications. Any errors should be corrected, and the translation should be finalized for delivery. This review process should be given sufficient time in the project schedule. It is standard practice to have reviewers check the translation to ensure quality and accuracy, even when the initial translation is done by professional human translators. The review helps identify issues that a single translator may miss. The same applies to LLMs.

F.2   Human Error Analysis Approach

Translation evaluation is not always conducted alongside the translation itself. There are various situations in which translation evaluation is required, such as deciding whether to accept or reject a translation, comparing multiple translation outputs, or selecting the most suitable version among candidates. Evaluation may also be needed for quality control in professional workflows, for assessing MT outputs, or for translator training and certification purposes.

Our error analysis methodology combines ISO 5060, the JTF guidelines, and the MQM framework to create a specification-aware evaluation system. This system is designed to score each translation based on how well it meets the predefined specifications. The more errors are found, the higher the score. Therefore, translations of higher quality should receive lower scores. By using these established frameworks, we develop a practical process that can be applied in various translation contexts.[15]

---

[15]Although full-text evaluation is ideal, practical constraints sometimes necessitate the use of samples. When translation samples are selected for evaluation, ISO 5060, MQM, and the JTF guidelines each offer instructions on how to carry out the sampling process (ISO5060:2024; MQM, 2025; JTF, 2018).

Step 1:   Ensure Specifications Are Accessible
Before evaluation begins, ensure that the translation specifications are available and accessible. If, for any reason, clear specifications are not defined at the time of translation, they should be established at this stage. The specifications should remain accessible throughout the evaluation process so that evaluators can refer to them consistently. Clear documentation helps ensure that all evaluators apply the same criteria and share a common understanding of the communicative goals of the translation.

Step 2: Define Error Categories Based on Specifications   Establish error categories aligned with the specifications to ensure consistency. ISO 5060 defines a translation error as a "failure to adhere to translation project specifications" (ISO5060:2024).

However, it is important to distinguish between specifications and error categories. Specifications describe the requirements agreed upon for a project. In contrast, error categories cover a broader range of issues, including problems that are commonly assumed but not always stated directly in the specifications. For example, accuracy-related errors or violations of general linguistic conventions, such as grammar or punctuation, are typically included in error taxonomies even when they are not listed in the specifications. As long as the selected error categories do not contradict the specifications, they can be applied to support consistent evaluation.

Reference standards like the JTF guidelines and MQM for error categorization (JTF, 2018; MQM, 2025). MQM provides fine-grained error categories that are especially valuable for detailed error analysis (MQM, 2025).

Error levels can be customized based on project priorities. For example, when translating a company's corporate philosophy for investor relations materials, where conveying nuance and tone is more important than achieving word-for-word accuracy, the following adjustments can be made:

- Accuracy: Prioritize mistranslation errors, while placing less emphasis on over-translation and under-translation.

- Style: Use more specific subcategories, such as:

  - Language register: Inappropriate level of formality
  - Awkward phrasing: Grammatically correct but stylistically poor

- Unidiomatic expressions: Unnatural to native speakers
- Inconsistent style: Stylistic variations across the document

This targeted approach allows evaluators to focus on errors that are most relevant to the specifications. Error categories should be adjusted to reflect the specific requirements of each project.

Step 3: Apply Weights to Error Categories ISO 5060 and the JTF guidelines apply weights to error categories based on project specifications. These weights can be set in advance by the client or project owner. Since different document types prioritize different aspects of translation quality, stakeholders should agree on appropriate weights (ISO5060:2024; JTF, 2018):

- 2.0 – Highly Important
- 1.5 – Somewhat Important
- 1.0 – Standard Importance (default)
- 0.5 – Less Important

To maintain balance, the average weight across all categories should be approximately 1.0. The weighting example below places greater emphasis on accuracy and consistent currency formatting, while giving less weight to stylistic elements:

Example: Financial Report—Revenue Forecast Section

- Accuracy: 1.0 (Standard): Basic factual correctness is required, with some flexibility in expressing forecasts.
- Style: 0.5 (Less Important): A professional tone is preferred, but it has little impact on understanding.
- Locale Convention: 1.5 (More Important): All monetary values must be shown in US dollars to ensure a consistent interpretation.

Step 4: Select Qualified Evaluators Evaluators should be professional translators or subject matter experts who are not only bilingual but also native speakers of the target language. Being bilingual alone is not sufficient for proper evaluation; evaluators must have a deep understanding of the linguistic nuances and cultural context of the target language.

Regarding the qualifications and competencies of evaluators, the ISO 5060 and MQM frameworks provide more rigorous requirements (ISO5060:2024; MQM, 2025). Practitioners should refer directly to these standards to ensure that evaluators meet the necessary professional criteria.

However, these strict requirements can be challenging in practice, as it is often hard to find and recruit qualified evaluators. When qualified evaluators are difficult to recruit, it may be useful to combine error-based evaluation conducted by available bilingual reviewers with subjective assessment by domain experts in the target language or end users. Domain experts may detect inconsistencies or errors by closely reading the content, while end users can provide direct feedback on whether the translation feels natural or conveys the intended message.

Step 5: Identify and Assess Errors Qualified evaluators identify errors, record them, and assess their severity. Severity indicates how serious an error is. It should always be judged based on whether the translation achieves its intended purpose and how much real-world impact the error may have (JTF, 2018). If an error has little or no practical impact, assigning a severity level may not be necessary.

Severity levels and their corresponding scores follow JTF (2018):

- Neutral (0): No penalty. These include stylistic preferences or repeated minor issues that do not affect comprehension.
- Minor (1): Errors that slightly affect readability but do not interfere with understanding.
- Major (10): Errors that significantly affect readability and comprehension.
- Critical (100): Errors that make the translation unusable and may cause harm, such as health risks, financial losses, or reputational damage. These must be corrected before publication.

Whether to count repeated errors multiple times should be decided through agreement among stakeholders (JTF, 2018).

Step 6: Calculate the Score Each identified error receives a score calculated as:

$$\text{Error Score} = \text{Category Weight} \times \text{Severity Score} \tag{1}$$

After assessing all errors, sum the scores to calculate the total error score. If severity scoring is not used, simply multiply the number of errors in each category by its assigned weight, and then sum the results.

A lower total score indicates fewer errors, and therefore a higher-quality translation. These scores make it possible to directly compare different translations.

For projects that require pass/fail decisions, evaluators can set a passing threshold based on acceptable error levels. Refer to the JTF guidelines or the MQM framework for recommended threshold values that fit different translation contexts (JTF, 2018; MQM, 2025).

As the evaluation is based on detailed specifications, it allows for a more objective assessment, reducing subjectivity and ensuring consistent criteria among evaluators.

F.3 Human Subjective Evaluation Approach

The JTF guidelines emphasize that while error-based evaluation methods provide a systematic approach, they represent only one aspect of translation quality (JTF, 2018). Error-based methods focus primarily on objectively identifiable errors and do not account for subjective quality factors that are essential in certain types of documents, such as advertisements, literary works, corporate vision statements, marketing slogans, and brand messages. For a more comprehensive assessment of quality, it is advisable to combine error-based evaluation with other approaches, particularly subjective evaluation by experts or end users, depending on the translation context (JTF, 2018).

Moreover, when qualified evaluators for error-based assessment are unavailable, subjective evaluation can complement error analysis performed by available bilingual reviewers. Subjective evaluation captures aspects such as clarity, persuasiveness, and appropriateness, which are essential for determining whether a translation effectively serves its intended purpose.

Subjective evaluation can be integrated with error-based assessment in the following ways:

A. Subjective Evaluation by Experts : Experts assess translations based on their specialized knowledge and professional judgment. Examples of such expert evaluations include:

- Legal professionals assess the accuracy and appropriateness of legal translations.

- Marketing specialists review the effectiveness and cultural relevance of promotional content.

- Technical staff on the client side evaluate clarity, precision, and technical correctness in the documentation.

B. Subjective Evaluation by End Users : End users evaluate translations based on their own perceptions and practical experience. For example, in the case of investor relations materials, investors may be asked:

- *Did you find the explanation clear, convincing, and appropriate?*

- *Were there any unnatural expressions or confusing elements in the translation?*

Their feedback is usually collected through surveys or questionnaires and provides valuable insight into how clear and usable the translation is.

Incorporating specifications into translation and evaluation enables both to go beyond basic accuracy and focus on communicative effectiveness. This ensures translations are not only correct but also appropriate for their intended audiences and contexts.

G Why Focus on Japanese-to-English Translation

Our experiment focuses on Japanese-to-English translation of integrated reports, which are typically published annually by companies. Here, we explain why this particular language pair and translation direction were chosen.

MT research often prioritizes universal approaches, aiming to develop models and evaluation methods that generalize across many language pairs. For example, Liu et al. (2024) note that this focus on generalization is evident in the field's pursuit of standardized methods. While such approaches help improve general performance across languages, they may overlook challenges specific to individual language pairs.

Indeed, prior research has reported substantial differences in MT performance between high-resource and low-resource languages (Team et al., 2022; Pang et al., 2025). These disparities have been attributed not only to the quantity of training data but also to inherent linguistic factors. For instance, Sarti et al. (2022) found that post-editing greatly improved English–Italian translations, but

had limited impact on English–Turkish and English–Japanese, likely due to word order and morphology differences. Similarly, Lee et al. (2022) report that mBART (Tang et al., 2021) performs well across domains but struggles with typologically distant languages, scoring below 3.0 BLEU (Papineni et al., 2002).

In response, our study takes a more targeted perspective, recognizing that translation difficulty varies widely depending on linguistic distance, grammatical structure, and cultural context. Japanese–English translation presents unique challenges due to fundamental differences in linguistic structure and writing systems. Unlike English–European language pairs, which often require minimal restructuring, Japanese–English translation typically involves major changes in sentence structure and word choice (Carl et al., 2016). Ogawa (2021) states that translating between Japanese and English takes significantly more processing time than English–German or English–French translation, leading to a higher cognitive load.

Due to these complexities, general approaches to translation models and evaluation, particularly those designed for multilingual settings, may not fully reflect the specific challenges of Japanese–English translation. Therefore, a specialized approach is necessary to understand not only how MT systems handle these linguistic difficulties, but also how their outputs can be appropriately evaluated.

We focus on the Japanese-to-English translation direction for two main reasons. First, ChatGPT and other LLM-based translation systems tend to perform better in English than in many other languages, due to the abundance of high-quality English training data (Chowdhery et al., 2022; Wendler et al., 2024). For example, OpenAI reports that GPT-3's training data is "primarily English (93 [percent] by word count)," and similar English-centric characteristics are reflected in the design and evaluation of GPT-4, as noted in its System Card (Brown et al., 2020; OpenAI et al., 2024). Working in this direction allows for both practical translation and evaluation while minimizing the influence of data-related limitations.

Second, the demand for Japanese-to-English translation greatly exceeds the supply of qualified human translators. Japan continues to face a shortage of professionals who can produce high-quality

English translations, especially in specialized fields such as investor relations and corporate communications (Nagata, 2025; Ueno, 2025). Given this shortage, MT guided by translation specifications represents a practical alternative that may help meet the demand for high-quality and cost-effective translations. In this study, we examine whether this approach can improve translation quality while also addressing the translator shortage and meeting international communication needs.

H   Company List

Table 6 lists the integrated reports used in this study. One company is selected from each of the 33 industry categories defined by the Tokyo Stock Exchange, prioritizing those ranked among the top four in market capitalization within each category during the data collection period (August 28–September 9, 2024).

The table includes the industry, company name, and publication year of the report used. Major companies with larger market capitalizations are chosen under the assumption that they are more likely to publish well-developed English versions of their integrated reports. As a result, many of the companies listed are well-known Japanese corporations.

I   Prompt Design for Specification-Aware ChatGPT Translations

Table 7 summarizes the prompts used for each ChatGPT translation method. For ChatGPT basic, ChatGPT receives only a minimal instruction: "Please translate the following Japanese text into English." For ChatGPT + Spec, we incorporate the content of the assumed specifications for the translation of integrated reports:

- **Source text context**: The official company name and a description of the integrated report as an IR document

- **Target language**: English

- **Intended purpose**: To enhance the company's appeal to a broad audience of investors

- **Target audience**: International investors

- **Style**: Clear and persuasive, suitable for a global investor audience

In the ChatGPT PE + Spec method, the model is instructed to improve the Google Translate out-

Industry	Company	Year
Transportation Equipment	Toyota Motor Corp.	'23
Banks	Mitsubishi UFJ Financial Group, Inc.	'23
Electric Appliances	Sony Group Corp.	'23
Retail Trade	Fast Retailing Co., Ltd.	'23
Services	Recruit Holdings Co., Ltd.	'23
Information & Communication	Nippon Telegraph and Telephone Corp.	'23
Chemicals	Shin-Etsu Chemical Co., Ltd.	'24
Wholesale Trade	Mitsubishi Corp.	'23
Pharmaceuticals	Chugai Pharmaceutical Co., Ltd.	'23
Other Products	ASICS Corp.	'23
Insurance	Tokio Marine Holdings, Inc.	'23
Foods	Japan Tobacco Inc.	'23
Precision Instruments	Terumo Corp.	'23
Fishery, Agriculture & Forestry	Nippon Suisan Kaisha, Ltd.	'23
Mining	Japan Petroleum Exploration Co., Ltd.	'23
Construction	Daiwa House Industry Co., Ltd.	'23
Textiles & Apparels	Goldwin Inc.	'23
Pulp & Paper	Oji Holdings Corp.	'24
Oil & Coal Products	ENEOS Holdings, Inc.	'23
Rubber Products	Bridgestone Corp.	'24
Glass & Ceramics Products	AGC Inc.	'24
Iron & Steel	Nippon Steel Corp.	'23
Nonferrous Metals	Sumitomo Electric Industries, Ltd.	'23
Metal Products	Sanwa Holdings Corp.	'23
Machinery	Mitsubishi Heavy Industries, Ltd.	'23
Electric Power & Gas	The Kansai Electric Power Co., Inc.	'23
Land Transportation	Central Japan Railway Co.	'23
Marine Transportation	Nippon Yusen Kabushiki Kaisha	'23
Air Transportation	All Nippon Airways Co., Ltd.	'23
Warehousing & Harbor Transportation Services	Mitsui-Soko Holdings Co., Ltd.	'23
Securities & Commodity Futures	Nomura Holdings, Inc.	'24
Other Financing Business	Japan Exchange Group, Inc.	'23
Real Estate	Mitsui Fudosan Co., Ltd.	'23

Table 6: Integrated reports of 33 listed companies used in the experiment.

put based on the same specifications, without access to the original Japanese text. This is a deliberate design choice; providing the source text risks the model disregarding the MT output and producing a new translation from scratch, a phenomenon observed in our initial pilot experiments. We therefore ensured that the task remained genuine post-editing, focused solely on enhancing the fluency and appeal of the existing translation. In both specification-aware methods, we instruct ChatGPT to avoid adding information not present in the source.

Since this is not a commissioned project, we define realistic specification parameters based on industry practices. Items like deadlines and formatting are excluded due to experimental constraints.

J Example Comparison of Five Translation Methods: All Nippon Airways Co., Ltd.

The five translations in Table 8 illustrate different approaches to conveying the original Japanese text in English. The source text is an excerpt from the corporate philosophy section of the integrated report published by All Nippon Airways Co., Ltd.

The official translation retains the Japanese term "waku waku," along with an explanatory note. It is unclear whether the original specifications required preserving the Japanese phrase. However, even if the intention was to reflect a sense of Japanese cultural identity, the primary objective should be to ensure that the translation appeals to international investors. One notable issue is that in the phrase "[a]nd when passed from person to person, becomes a force," the subject is missing, which could be considered an error. Additionally, the phrase "explore more incredible discoveries" may sound awkward, as "explore" typically refers to something unknown rather than something already discovered (Harper, 2025).

Google Translate output is simple and easy to understand, but expressions such as "make travel more fun" and "more exciting sky experiences" might be too casual and do not fit the context of

Translation Type	Prompt
ChatGPT basic	Please translate the following Japanese text into English.
ChatGPT + Spec	The following Japanese text is an excerpt from the integrated report of [company name], a key part of the company's investor relations materials. Please translate this text into English in a way that will be appealing to international investors. The purpose of this translation is to enhance the company's appeal to a wider audience of investors. Please do not add any additional information.
ChatGPT PE + Spec	The following text is a translation of an excerpt from the integrated report of [company name], a key part of the company's investor relations materials. The purpose of this translation is to enhance the company's appeal to a wider audience of investors. The initial translation was done using Google Translate. Please refine this translation to make it more engaging and appealing in English. Please do not add any additional information.

Table 7: Examples of ChatGPT prompts used in our study.

Type	Translation Output
Source	"ワクワク"は、人を動かすエネルギー. それは人から人へと伝わり、世界をあかるく元気にする。私たちは, 空の可能性を信じ, ヒト・モノ・コトをつなぐことで, これまでたくさんの笑顔と歓びに出会ってきた. そして今, 新しい未来に向けて, 次のワクワクを生み出す. 移動をもっと楽しいものにできないか, モノの運び方を新しくできないか, 心躍る日常を届けられないか, どこまでも地球にやさしくなれないか. これまで以上の空の体験とともに, 驚きと感動を創りたい.
Official	"Waku waku" is what moves people to push what's possible. It's Japanese for the joy and excitement of discovering the unknown. And when passed from person to person, becomes a force that creates a brighter world, united in wonder. The sky is full of possibilities, which has allowed us to bring together people, products and experiences for decades. Now, we're expanding our horizons and spreading "waku waku" across the world. So, we wonder, can we make travel more enjoyable? Can we renew the way we transport products? Can we bring excitement to life, while being kind to the earth? Together with an elevated experience in the sky, we can explore more incredible discoveries that will lead to a better tomorrow.
Google	"Excitement" is the energy that moves people. It spreads from person to person, making the world brighter and more energetic. We believe in the possibilities of the sky, and by connecting people, things, and events, we have encountered many smiles and joy. And now, we are creating the next excitement for a new future. Can we make travel more fun, can we innovate the way we transport things, can we deliver an exciting daily life, and can we be kind to the earth as much as possible? We want to create surprises and excitement along with more exciting sky experiences than ever before.
GPT	"Excitement" is the energy that moves people. It spreads from person to person, bringing brightness and vitality to the world. Believing in the possibilities of the sky, we have connected people, goods, and experiences, encountering countless smiles and joy along the way. Now, we are setting out to create the next source of excitement for a new future. Can we make travel more enjoyable? Can we reinvent the way goods are transported? Can we deliver everyday moments that make your heart dance? Can we be more environmentally friendly than ever before? Along with new and enhanced experiences in the sky, we aim to create wonder and inspiration.
GPT+Sp	Excitement is the energy that moves people. It spreads from person to person, brightening and invigorating the world. We believe in the limitless potential of the skies and, through connecting people, goods, and experiences, we have encountered countless smiles and moments of joy. Now, as we look toward a new future, we are ready to create the next wave of excitement. Can we make travel more enjoyable? Can we innovate the way things are transported? Can we bring a sense of wonder to everyday life? Can we go further in protecting our planet? We aim to create surprises and inspiration with an even greater sky experience.
GPT PE+Sp	"Excitement" is the spark that moves people, spreading from one person to another, brightening and energizing the world. Believing in the boundless potential of the skies, we bring people, goods, and experiences together, creating countless moments of joy and countless smiles. Now, we are embarking on a new journey to inspire even greater excitement for the future. Can we make travel more delightful, revolutionize the way we transport goods, infuse everyday life with excitement, and care for our planet in the best possible ways? We are committed to creating moments of surprise and delight, offering more thrilling experiences in the skies than ever before.

Table 8: Differences in translations: All Nippon Airways Co., Ltd.

investor relations materials. The phrase "encountered many smiles and joy" also sounds unnatural, as "encounter" is typically used with concrete entities or situations, such as difficulties or opposition, rather than with abstract concepts like joy or smiles (Merriam-Webster Dictionary, 2025).

ChatGPT basic captures the emotional aspects of the original text, particularly with phrases like "moments that make your heart dance." However, the use of "encountering countless smiles and joy," similar to Goolge Translate, does not sound natural in English.

ChatGPT + Spec employs expressions such as "limitless potential of the skies," which are effective in creating a positive and aspirational tone. However, the use of "things" in "the way things are transported" sounds casual, and "even greater sky experience" would sound more natural if "experiences" were used in the plural form.

ChatGPT PE + Spec employs strong and active expressions. The word "spark" in "[e]xcitement is the spark that moves people" creates a vivid and powerful impression, while the phrase "embarking on a new journey to inspire even greater excitement" conveys a positive and future-oriented feeling.

These observations suggest that each translation method tends to produce distinct results, and that including specifications in ChatGPT prompts may encourage the use of more purposeful and engaging language.

K Error Typology for Human Error Evaluations

We use the following categories and subtypes for error annotation: Accuracy (subtypes: mistranslation, addition, omission), Linguistic Conventions (grammar, spelling, unintelligible, textual conventions), and Style (language register, awkward style, unidiomatic style, inconsistent style). Definitions and examples of each error type are available on the official MQM website (`https://themqm.org/downloads/`). Evaluators are instructed to refer to this table during error analysis. However, only the main categories are used for error counting; subtypes are provided to help annotators better understand and identify specific issues.

L Syntactic Pattern Analysis of Translation Outputs

To investigate stylistic tendencies, we analyzed syntactic patterns across the five translation types, focusing on relative clauses and clausal coordination. Specifically, we counted the number of relative pronouns (*which*, *who*, *that*) and clausal instances of *and* (i.e., those connecting two clauses with subject-verb structures). We excluded uses of *that* as complementizers, demonstratives, or in cleft constructions, and excluded *and* used at the phrase or word level.

These structures are common in Japanese texts and may reflect a literal transfer of source syntax. Their frequent use can lead to complex, additive structures that may reduce readability in English.

Table 9 reports total word counts, raw counts, and normalized frequencies per 1,000 words. The results show that Google Translate and the official translations use relatively more relative clauses and clause-level coordination, suggesting less restructuring. In contrast, ChatGPT outputs display simpler sentence structures regardless of prompt specificity. These patterns indicate that prompt-based LLMs tend to favor fluency and conciseness.

M Examples and Error Analysis from Official Translations

To explore why the official translations received low ratings, we examined problematic excerpts from companies' corporate philosophies. Table 10 presents these examples.

To begin with, Excerpt (1) contains the phrase "offering both a multitude of choices," in which the use of *both* appears semantically inappropriate. The word *both* typically introduces two parallel elements, but "a multitude of choices" is a singular, collective concept, resulting in a semantic mismatch. In this context, *both* is presumably intended to refer to the two entities mentioned earlier, *individuals* and *businesses*. However, its placement creates a structurally awkward and confusing expression.

Excerpt (2) opens with the sentence "[w]hat we do isn't a job," which appears to aim for an inspirational tone but lacks a clear referent or elaboration. As a result, its meaning may be ambiguous to readers who are not familiar with the intended message behind the expression. A clearer alternative might be "[w]hat we do is more than a job," which conveys the intended message more directly.

Excerpt (3) contains the phrase "in this era of search," which is not a commonly used expression in English discourse and may not be immediately clear. Additionally, the combination of the degree

Type	Official	Google	ChatGPT	ChatGPT+Sp	ChatGPT PE+Sp
Word Count	8,776	**8,894**	8,655	8,351	8,216
Clausal *and*s	58	**66**	54	53	57
Rel. Pronouns	72	**83**	61	59	61
*and*s / 1000w	6.61	**7.42**	6.24	6.35	6.94
RelP / 1000w	8.20	**9.34**	7.05	7.07	7.42

Table 9: Syntactic feature counts per translation type. RelP = relative pronouns (*which*, *who*, and *that*). Frequencies normalized per 1,000 words.

No.	Excerpt
(1)	Opportunities for Life. Faster, simpler and closer to you. Since our foundation, we have connected individuals and businesses, offering both a multitude of choices. (Recruit Holdings Co., Ltd.)
(2)	What we do isn't a job. We enjoy exploring what is possible for our future. We question the status quo, fail well and overcome with resilience. We are a force for change. (Recruit Holdings Co., Ltd.)
(3)	In this era of search, where information has become available anytime anywhere, we need to focus more on proposing the optimal choice. We seek to provide 'Opportunities for Life,' much faster, surprisingly simpler and closer than ever before. (Recruit Holdings Co., Ltd.)
(4)	Today, what we mean by Our Hopes for the Future, a world where we are our truest selves, respecting, and inspiring each other. Living together in harmony with our planet—in harmony with People and Nature. (Daiwa House Industry Co., Ltd.)
(5)	Create a virtuous cycle between Society and Earth by fully utilizing less of her limited resources. Make the world a richer, better place by bringing out the best out in people and the potential of buildings. (Daiwa House Industry Co., Ltd.)
(6)	For the sake of the Earth, which future generations of children have entrusted in our care. Together with you. (Bridgestone Corp.)
(7)	The single continuous curve represents the dynamism and our commitment for continuous innovation and delivering value to people and society. (Nippon Telegraph and Telephone Corporation)

Table 10: Excerpts from the official translation cited in the qualitative analysis.

adverb *surprisingly* with the comparative adjective *simpler* creates a stylistic inconsistency.

Excerpt (4) is grammatically incomplete. The subject and predicate do not form a complete clause, making the intended meaning difficult to determine.

Excerpt (5) uses the phrase "fully utilizing less," which appears to aim for a concise message about efficiency, likely meaning "to make the most of fewer resources." However, the expression is semantically ambiguous. The adverb *fully* suggests maximization, while *less* implies minimization, cre-

ating a tension that may confuse readers rather than clarify the company's commitment to sustainability. In addition, the phrase "bringing out the best out in people" is grammatically incorrect. The structure redundantly includes both "out" before and after "the best," where only one instance is appropriate. A corrected version would be "bringing out the best in people," which is idiomatic and clear.

Excerpt (6) contains grammatical issues. The phrase "entrusted in our care" is unidiomatic. In standard English, the verb *entrust* typically appears in the form "entrust someone with something" or entrust something to someone." In addition, the sentence lacks a main clause and does not constitute a complete grammatical unit.

Finally, Excerpt (7) contains two issues. First, "commitment for" is a grammatical error. The standard preposition in this context is "commitment to." Second, the coordination of "continuous innovation," a noun phrase, and "delivering value," a gerund phrase, is unbalanced and stylistically awkward. For clarity and parallel structure, both elements should be in the same grammatical form, such as "continuous innovation and value creation," or "innovating continuously and delivering value."

As these examples show, the official translation includes not only grammatical inaccuracies but also semantic and stylistic inconsistencies, which may have contributed to its lower rating in the evaluation.

The characteristics of the source texts themselves may help explain the relatively low ratings of the official translations. Integrated reports sometimes contain expressions that are abstract, culture-specific, or metaphorical in Japanese, which can result in awkward or even ungrammatical output if translated too literally. In some cases, the official translations appear to reflect such overly direct translations, suggesting that the translator may have prioritized fidelity to the source text's wording or syntax at the expense of naturalness and clarity in English. While this approach may have been inten-

tional, for example to preserve a uniquely Japanese tone, it can hinder readability and reduce the overall appeal of the translation. This is reflected in the lower rankings observed in the subjective evaluation.

However, it is important to note that professional translations are typically produced based on specifications or an internal guideline (ISO17100:2015). In practice, translators are expected to follow such instructions from the outset; without them, it would be difficult to even begin the task. Therefore, the types of problems identified in the official translations, such as grammatical errors or awkward phrasing, are unlikely to stem from missing or unclear specifications. Rather, these issues may be related to language proficiency or a mismatch between the translator's background and the specific requirements of the task.

Specifications can support translation decisions, but achieving linguistic accuracy and fluency may still require a high level of language proficiency. As noted earlier (Section 5.2.2), human translation quality tends to vary, as translators differ in background and ability (Freitag et al., 2023; Ramos and Guzmán, 2024; Volz and von Thiessen, 2024).

Furthermore, as discussed in Appendix G, Japan continues to face a shortage of translators capable of producing high-quality English translations. This shortage may have influenced the present results.

In this context, the potential of MT systems that use specifications, such as ChatGPT with customized prompts or post-edited outputs, deserves more attention. Both ChatGPT + Spec and ChatGPT PE + Spec are favorably evaluated in our study, not only in subjective rankings but also in error-based analysis, suggesting that specification-aware MT may offer a useful complement to traditional workflows.