Universiteit Leiden

Master Computer Science

A Two-Track Evaluation of LLM and NMT Translation for Regulated Pension Communication

Name:	Nicholas Radunovic
Student ID:	s2072742
Date:	29/05/2026
Specialisation:	Data Science
1st supervisor:	Dr. P.W.H. van der Putten
2nd supervisor:	Dr. F.M. Plaza del Arco
3rd supervisor:	Dr. A. Karakanta

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Pension communication often requires recipients to understand decisions and required actions, yet such messages do not always elicit a response. For recipients whose primary language does not match the language of the communication, language barriers may further reduce comprehension and engagement, making clear and reliable translation of client-facing pension communication important. Against this background, we evaluate off-the-shelf machine translation for Dutch pension-domain client texts into seven target languages (Turkish, Spanish, English, Polish, Ukrainian, Arabic, German) in a two-track pilot that combines human and automated evaluation. Track A is reference-free: native speakers evaluate anonymized translations of six domain-representative excerpts translated by GPT-4 and Azure Translator using a rubric tailored to enterprise use, covering meaning preservation, register, and readability. This track is supplemented by an LLM-as-a-judge analysis using the same rubric. Track B is controlled and reference-based: nine systems (three commercial NMT engines and six LLMs) are benchmarked on a fixed five-paragraph document against professional references using overlap- and edit-based metrics (BLEU, chrF++, TER), learned semantic metrics (BERTScore, COMET), bootstrap uncertainty, and paragraph-level fidelity probes targeting omission- and formatting-related risk. Native-speaker judgments consistently prefer GPT-4 over Azure across all languages, with the largest gains in register. The reference-based results show a tight top cluster, strong cross-language heterogeneity, and occasional disagreement between character-overlap and semantic metrics in a single-reference setting. Paragraph-level diagnostics highlight omission- and formatting-related risks, especially for Arabic and Ukrainian. Finally, an LLM-as-a-judge tracks the direction of human preferences but exhibits score compression, supporting its use only as a coarse monitoring signal. Together, the findings support deployment workflows that combine rubric-aligned acceptability checks, controlled audits with uncertainty, and explicit terminology and structure governance for regulated multilingual communication.

1 Introduction

Client-facing communication in regulated domains must be both accurate and understandable. In the pension domain, messages often describe eligibility conditions, benefit changes, and required actions; small translation errors, such as a mistranslated instruction or an omitted condition, can materially change a reader's understanding of their rights and obligations.

Communication effectiveness in the pension domain cannot be assumed. Pension funds routinely inform clients about entitlements and decisions that require action, yet recipients often do not respond to such messages [1–3]. Research on Dutch pension communication also highlights that mandated disclosure and legal precision can conflict with comprehensibility, and that revisions can change information-finding and understanding outcomes [4]. To support informed decision-making and timely action, information and calls to action must be understandable and well received. For some populations, including migrant workers, language barriers may contribute to low engagement, underscoring the importance of clear, accessible multilingual communication [5].

Organizations increasingly adopt machine translation (MT) to deliver such communication at scale, including in financial-services and institutional settings where quality requirements emphasize continuity of institutional discourse and terminology [6–8]. Over the past decade, neural MT (NMT) has become the dominant commercial approach, improving fluency and meaning preservation relative to earlier paradigms [9–11]. More recently, large language models (LLMs) have entered the MT landscape as general-purpose systems capable of translation via prompting and/or fine-tuning [12–14].

The current MT ecosystem creates both opportunity and context-dependent risk. In client-facing content delivery, risk is not merely the possibility of translation error, but the possibility that such errors have harmful consequences for particular users and use cases; at the same time, these risks may need to be weighed against the higher-level risk of not translating information at all [15]. Commercial NMT services offer strong general-purpose translation, but regulated enterprise deployments often require additional guarantees such as stable institutional register, consistent domain terminology, and predictable formatting [8, 16, 17]. LLMs are attractive because their behavior can be guided at inference time via prompts and structured constraints (e.g., preferred terminology and protected entities) [12, 18, 19], yet empirical work shows that instruction semantics can be partially ignored in-context and that translation performance varies strongly by language-resource conditions [13]. Moreover, both NMT and LLM-based MT can exhibit catastrophic pathologies such as hallucinations and omissions, which undermine user trust and are especially salient in low-resource and out-of-English directions [20–22]. These trade-offs are amplified in domain-specific settings, where in-domain parallel data and expert validation are scarce and out-of-domain performance can degrade sharply, motivating domain adaptation and controlled evaluation [22, 23].

A second challenge is evaluation. Trusted reference translations are frequently unavailable because professional references are costly and content changes rapidly, so reference-free quality estimation (QE) and alternative evaluation protocols are widely studied [24, 25]. Operational risk is also driven by *tail behavior*: a single mistranslated paragraph or instruction can invalidate an otherwise acceptable document, even when document-level averages look strong. This motivates error-based evaluation frameworks and document-aware analyses that surface concentrated failures [20, 26]. We therefore treat translation quality as *enterprise acceptability*: faithful meaning transfer plus stable institutional register, readability for non-expert recipients, and consistent handling of protected entities and pension-domain terminology [26].

In this paper, we present a two-track pilot study of Dutch pension-domain client communication translated into seven target languages: Turkish, Spanish, English, Polish, Ukrainian, Arabic, and German. These languages were selected in consultation with an external stakeholder in the Dutch pension sector, based on prior, unpublished internal analyses of multilingual communication needs in the Netherlands. The set was intended to include both widely used international languages and languages commonly spoken by recipient groups that had shown lower response rates to Dutch-language pension communications containing calls to action. The first track is reference-free and evaluates enterprise acceptability directly through native-speaker ratings using a rubric tailored to regulated client communication, covering meaning correctness, register/style, and readability, supplemented by qualitative error notes and an exploratory LLM-as-a-judge signal. The second track is controlled and reference-based: it compares system outputs against professional translations of one fixed multi-paragraph document and adds uncertainty estimates and paragraph-level diagnostics to identify concentrated failures that document-level averages may hide. Both tracks include a

comparison between GPT-4 and Azure Translator (see Section 3; results in Section 4), thereby linking the two tracks and allowing us to examine where human judgments and automatic metrics converge or diverge [27, 28].

This pilot study addresses four applied research questions:

1. Under a rubric tailored to regulated pension communication, how do native speakers compare GPT-4 and Azure Translator across seven target languages?
2. In a controlled reference-based setting, how do system rankings vary across languages, metric families, and paragraph-level diagnostics?
3. In a cross-track comparison using the shared GPT-4 versus Azure Translator contrast, to what extent do reference-based metrics align with human acceptability judgments?
4. To what extent can LLM-as-a-judge grading serve as a proxy for human evaluation in pension-domain translation settings?

This paper makes three contributions:

- It presents an applied multilingual evaluation of Dutch pension-domain client communication across seven target languages.
- It introduces a two-track evaluation design that combines rubric-based human assessment with a controlled reference-based benchmark and paragraph-level diagnostics for localized failures.
- It examines alignment in two complementary ways: between automatic reference-based metrics and human acceptability judgments through the shared GPT-4 versus Azure Translator comparison, and between LLM-as-a-judge rubric scores and human evaluation as an exploratory proxy signal.

Taken together, these contributions provide a reproducible pilot framework for evaluating translation quality in regulated multilingual communication. To support transparency and reuse, the study resources, including the analysis code, prompts, scoring scripts, table and figure generation code, and publicly shareable text materials, are available in an archived public repository release [29].

The remainder of this paper is structured as follows. Section 2 situates pension-domain client communication within regulated, action-oriented translation requirements and reviews related work on domain adaptation, terminology/register control, and MT evaluation. Section 3 describes the two-track study design, the data and target languages, the compared systems, and the reference-free and reference-based evaluation protocols. Section 4 presents the empirical findings from both tracks, including cross-language variation, metric sensitivity, paragraph-level diagnostics, and cross-track triangulation. Section 5 interprets these results in terms of enterprise acceptability, operational risk, and alignment between automatic metrics, LLM-as-a-judge scores, and human judgments. Section 6 discusses the main limitations of the pilot study. Section 7 outlines concrete extensions toward deployment-grade evaluation and monitoring. Finally, Section 8 summarizes the main takeaways and implications for regulated multilingual pension communication.

2 Background and related work

This paper builds on two strands of prior work: (i) domain- and risk-sensitive machine translation (MT) for regulated client communication and (ii) the growing use of large language models (LLMs) both as translation systems and as evaluators. We focus on what prior work implies for regulated, client-facing acceptability, with particular emphasis on meaning-critical errors, institutional register and terminology control, and readability, and we discuss how these requirements shape practical evaluation.

2.1 Pension communication as a regulated, action-oriented genre

Pension-domain communication (letters, emails, webpages) typically combines specialized terminology with an institutional register and action-guiding instructions. In this genre, localized meaning errors, such as omitted conditions, polarity reversals, altered dates/numerals, or misrendered entities (e.g., benefit types,

forms, authentication services), can change a reader's understanding of rights and obligations and the actions they take [30, 31]. Importantly, challenges in pension communication are not limited to translation. Prior work suggests that recipients may fail to act on pension information due to limited comprehension, low engagement, or difficulty converting information into decisions [1–3]. In multilingual settings, language barriers can further reduce comprehension and engagement for some populations, including migrant workers [5]. Together, these findings motivate treating readability and stable institutional voice as first-class requirements in addition to semantic correctness.

2.2 Domain shift, adaptation assumptions, and enterprise constraints

Neural MT (NMT) dominates production systems and achieves strong fluency on many general-domain benchmarks, yet it can degrade under domain shift due to mismatched terminology, genre conventions, and distributional differences between training and deployment text [22, 23]. In regulated enterprise communication, the quality bar extends beyond surface fluency to faithful transfer of conditions and numerals, consistent handling of institution- and product-specific entities, and an appropriate formal register [30, 31]. While domain adaptation techniques are well studied [23], many approaches implicitly presume access to in-domain parallel data and/or substantial monolingual resources for fine-tuning, back-translation, or terminology learning—resources that may be limited, costly, or sensitive in regulated business contexts.

A common operational reality is that organizations start with *off-the-shelf* systems (commercial NMT services or general-purpose LLMs) applied to domain-representative text. Prior applied studies in other safety-relevant settings (e.g., healthcare translation) emphasize that fluent output can still contain salient omissions or distortions, motivating continued human oversight and domain-aware evaluation design [32]. These findings support evaluating systems on domain-representative inputs and analyzing the error types that drive downstream decisions and operational risk (e.g., conditions, instructions, dates/numerals, protected entities, and regulated phrasing).

These constraints are uneven across languages. In MT, "low-resource" typically refers less to the number of speakers than to the availability and quality of training and evaluation resources (parallel data, in-domain corpora, and qualified human evaluators) [33]. As a result, cross-language heterogeneity is expected: systems may degrade differently under domain shift, and reliable evaluation may be harder to scale for some language–domain combinations [21, 33].

2.3 From NMT to LLM-based translation

The MT landscape is changing as LLMs are increasingly deployed as general-purpose translation engines or as components within MT pipelines [34, 35]. Comparative studies report that prompting can elicit competitive translation quality from GPT-family models, while also revealing sensitivity to prompt formulation and robustness concerns [12, 36]. Human evaluation based on Multidimensional Quality Metrics (MQM) further suggests that strong LLMs can approach the performance of junior professional translators in some directions and domains, but still exhibit characteristic weaknesses such as lexical inconsistency and occasional meaning errors [37]. At the same time, the MT landscape is becoming more diverse: alongside general-purpose LLMs, newer translation-oriented LLMs such as TranslateGemma illustrate the emergence of models explicitly optimized for multilingual translation rather than broad instruction following [38].

For enterprise use, a central consideration is *inference-time controllability*: LLM outputs can often be steered toward preferred register, terminology, and formatting through prompts and structured constraints. At the same time, prior work highlights accompanying risks, including over-paraphrase, terminology drift, omissions, and hallucination-like additions, especially in lower-resource settings and meaning-sensitive translation tasks [20, 21]. This motivates evaluation that assesses not only average quality but also faithfulness and institutional appropriateness under controllability pressures.

2.4 Terminology governance and register control

A recurring challenge in specialized communication is enforcing consistent translations of domain terms (product names, legal notions, fixed phrases) and maintaining the intended level of formality. For NMT, lexically constrained decoding can guarantee the appearance of required terms [39, 40]. Related work uses

controllable generation signals (e.g., side constraints) to steer style dimensions such as formality or politeness [41]. For LLM-based MT, analogous control is often pursued through prompting and structured inputs (e.g., glossaries or protected term lists), including proposals to leverage LLMs for domain terminology integration [42]. Register preservation remains important for trust and clarity in client-facing communication; recent work explicitly targets formality improvements using LLMs [43].

These threads jointly suggest that evaluation instruments should be sensitive to terminology drift and register shifts even when translations are otherwise fluent.

2.5 Evidence sources for MT quality

MT quality can be supported by three evidence streams: (i) reference-based metrics when trusted references exist, (ii) human evaluation and qualitative error analysis, and (iii) reference-free methods such as QE and LLM-as-a-judge.

2.5.1 Reference-based evaluation

When high-quality references are available, reference-based evaluation remains the most common experimental paradigm in MT. Classic overlap and edit metrics such as BLEU [44], chrF [45], and TER [46] are inexpensive and reproducible, but can under-reward valid paraphrases and are sensitive to reference choice and domain style [47, 48]. This limitation can be salient when comparing systems with different paraphrastic behavior (e.g., LLMs versus more literal NMT outputs) or in domains where institutional phrasing is constrained.

Learned metrics aim to better track semantic equivalence and human ratings, including embedding-based similarity (BERTScore [49]) and supervised evaluators such as BLEURT [50] and COMET [51]. Shared-task meta-evaluations indicate that neural metrics often improve correlation with expert human judgments across settings [52]. At the same time, learned metrics raise interpretability and domain-transfer questions: a high score does not, by itself, explain *which* content is wrong or *why* a segment is unacceptable in a risk-sensitive context. Applied work therefore commonly pairs corpus-level scores with segment-level inspection and qualitative evidence [53].

When systems are close in score, MT studies often use resampling-based uncertainty estimates (e.g., paired bootstrap resampling) to assess whether apparent gaps are robust under a fixed test set [54].

2.5.2 Human evaluation frameworks and qualitative error analysis

Human evaluation remains the most direct way to assess whether translations satisfy domain- and user-specific requirements. Large-scale MT evaluations have used direct assessment and related segment-level rating protocols [55], while structured frameworks aim to improve interpretability and reliability. MQM provides a multidimensional taxonomy and severity framework that supports systematic annotation and aggregation [56, 57]. Recent work highlights that expert, context-aware evaluation can yield rankings and diagnostics that differ from crowd-based judgments, underscoring the importance of protocol choices [26].

In domain-specific applied settings, evaluation often faces practical constraints: limited budgets for professional references, limited access to trained MQM annotators, and the need to cover multiple target languages. Under such constraints, smaller-scale native-speaker panels combined with rubric-style ratings and qualitative comments can still provide actionable evidence about high-impact failure modes (e.g., incorrect conditions, number/date errors, register instability, and terminology drift) [32].

2.5.3 Reference-free assessment

In many enterprise deployments, trusted reference translations are unavailable for substantial portions of the content stream, motivating reference-free approaches. Traditional MT quality estimation (QE) predicts translation quality without a reference and spans multiple granularities [24]. More recent work also emphasizes explainable QE and related diagnostic approaches that localize and characterize translation errors [58]. For instance, xCOMET combines sentence-level evaluation with MQM-style error span detection, while adjacent work such as xTower explores free-text explanations of translation errors to support interpretation and correction [59, 60]. More recently, LLMs have been proposed as reference-free evaluators through

prompting. GEMBA reports that GPT-family models can produce system-level MT quality scores that correlate with human judgments under certain conditions, while also highlighting prompt sensitivity and model capability thresholds [61]. In broader NLG evaluation, frameworks such as G-Eval emphasize rubric-style prompting and structured scoring [62]. Other work suggests that providing references can materially improve the reliability of LLM-based MT evaluation [63], and rubric- or MQM-inspired prompting can increase interpretability while still requiring meta-evaluation for bias and calibration [64].

In regulated settings, a single meaning-critical error can dominate operational risk, motivating segment-level diagnostics alongside document-level aggregates [53, 65].

2.6 Positioning and gap

Taken together, prior work implies three practical points for machine translation in regulated domains. First, reference-based automatic metrics are useful for comparing systems, but do not fully capture whether a translation is acceptable for client-facing use [47, 52]. Second, structured human evaluation and qualitative error analysis remain necessary for identifying failures such as terminology drift and inappropriate register [26, 56]. Third, reference-free methods, including QE and LLM-as-a-judge, are becoming more feasible, but their outputs still require validation and careful use [61, 62]. What remains scarce is applied evidence that examines these evaluation approaches together in multilingual, regulated client communication. This gap is especially relevant in enterprise settings, where in-domain adaptation data are often limited, professional references are costly or unavailable, and translation quality must be assessed under governance constraints. This paper addresses that gap through a two-track evaluation of Dutch pension-domain client communication that combines human acceptability judgments, reference-based benchmarking, and exploratory reference-free assessment.

3 Methods

This section describes the two-track design used to assess translation quality for Dutch pension-domain client communication. We first outline the study design, then describe the source materials, target languages, and systems under comparison, followed by the procedures for translation generation and the two evaluation tracks.

3.1 Study design

We conducted a pilot study comparing translations for Dutch pension-domain client communications produced by a commercial neural machine translation (NMT) service and by large language model (LLM) approaches. The evaluation was organized into two complementary tracks.

Track A (reference-free) evaluated translation quality without reference translations. Its quality assessment relied on a structured human rubric and, in parallel, an LLM-as-a-judge proxy. This track focused on whether translations preserved meaning, maintained an appropriate institutional register, and remained readable for recipients.

Track B (reference-based) evaluated system outputs against professional reference translations for one fixed Dutch source document. It used automatic metrics, uncertainty estimates, and paragraph-level diagnostics. Because overlap-based reference metrics may under-reward acceptable paraphrases, we interpreted the results from Track B as complementary to the rubric-based judgments from Track A rather than as a substitute for them.

3.2 Source materials and target languages

Track A source excerpts. For Track A, we compiled six Dutch pension-domain excerpts: four from client letters and two from informational webpages. Sample construction was carried out in collaboration with an external stakeholder in the pension sector. The stakeholder provided a pool of 20 client letters in digital form, from which we selected four for inclusion, primarily on the basis of length. To broaden document-type coverage, we also considered a pool of 10 informational webpages from the same stakeholder and selected two excerpts, again using length as the main criterion.

Across both document types, we targeted excerpts of approximately 200–500 words. This range was intended to keep the evaluation manageable for human raters while preserving enough context to reflect realistic communication. In consultation with the stakeholder, we limited the final Track A set to six excerpts because a larger set would have imposed substantially more manual work on the native-speaker evaluators. The resulting four-versus-two distribution was therefore not based on stratified sampling or a predefined proportional design, but on practical choices appropriate to this pilot study.

Before translation, all source documents were anonymized by replacing personal names with placeholders (e.g., `xxxx. xxxxxx`). The selected excerpts covered recurring communicative functions such as informing and instructing, and included translation-sensitive terminology and institutional phrasing.

Track B source document. Track B used one Dutch client-facing pension letter selected from the six Track A documents. The letter informed the recipient about a change in premium-free pension accrual following updated information from the Uitvoeringsinstituut Werknemersverzekeringen (UWV) on disability status and WAO/WIA benefits. It combined explanatory and instructive language and was representative of pension-domain client communication because of its institutional phrasing and pension-specific terminology.

The document contains 197 words across five short paragraphs, with paragraph lengths of 28, 50, 31, 52, and 36 words. It was selected because it was short enough to make professional reference translation into seven target languages feasible within the project's resource constraints, while still remaining representative of pension-domain client communication. For each target language, we commissioned one professional reference translation from an external commercial translation provider (Translayte).[1] The translations were delivered as plain text with the original five-paragraph structure preserved. These paragraph boundaries were used as the atomic evaluation units in Track B. The professional reference translations are not reproduced in the main text, but are included in the GitHub source materials released with this study [29]. The full Dutch source text is provided in Appendix A.

Target languages. We evaluated seven target languages: Turkish, Spanish, English, Polish, Ukrainian, Arabic, and German. The language set was defined in consultation with the same external stakeholder in the Dutch pension sector in order to reflect multilingual communication needs relevant to the Dutch context. It combined widely used international languages (English, Spanish, and German) with languages commonly spoken by recipient groups that earlier internal analyses had associated with lower response rates to Dutch-language pension communications containing calls to action (Turkish, Polish, Ukrainian, and Arabic). The language set also spanned multiple language families and writing systems, including Latin, Cyrillic, and Arabic scripts.

3.3 Systems and translation generation

Systems compared. The system comparison differed by track. Track A compared one LLM approach (GPT-4) with one commercial NMT baseline (Azure Translator). Track B evaluated nine systems: three NMT systems (DeepL, Azure Translator, and Google Translate) and six LLM-based systems (Gemini 3 Pro, GPT-4, GPT-5.2, Llama 3.1 8B, Llama 3.1 405B, and Llama 3.3 70B). Table 1 reports the exact system metadata used in the experiments, including provider, access platform, route, model identifier where available, and version information where exposed by the platform.

The three NMT systems were included as widely used proprietary translation services and served as strong non-LLM baselines. Publicly available information places Google Translate, Azure Translator, and DeepL's classic model within the modern neural MT tradition, although their exact production architectures are not fully disclosed [28, 66–69].

The six LLMs were selected to capture variation along several dimensions that may affect translation quality: provider diversity (Google, OpenAI, and Meta), openness (proprietary and open-weight models), generation (older and more recently released systems), and scale within the same model family. In particular, including both Llama 3.1 8B and Llama 3.1 405B enabled a within-family comparison of model size, while the inclusion of Llama 3.3 70B, GPT-4, and GPT-5.2 allowed for a limited comparison across older and newer model generations. The system set was not intended to be exhaustive, but to provide a representative pilot sample of current NMT and LLM approaches relevant to enterprise translation in this domain.

[1] `https://translayte.com/`

Table 1: System metadata for all evaluated translation systems, including provider, access platform, route, model identifier, and version information where exposed by the platform.

Type	System	Provider	Access platform	Route	Model name	Model version
NMT	Azure Translator	Microsoft	Azure	API	—	v3.0
	DeepL Translator	DeepL	DeepL	Web interface	—	Not exposed
	Google Translate	Google	Google	Web interface	—	Not exposed
LLM	Gemini 3 Pro	Google	Google	Web interface	—	Not exposed
	GPT-4	OpenAI	Microsoft Foundry	API	gpt-4	turbo-2024-04-09
	GPT-5.2	OpenAI	Microsoft Foundry	API	gpt-5.2	2025-12-11
	Llama 3.1 8B	Meta	Microsoft Foundry	API	Meta-Llama-3.1-8B-Instruct	6
	Llama 3.1 405B	Meta	Microsoft Foundry	API	Meta-Llama-3.1-405B-Instruct	1
	Llama 3.3 70B	Meta	Microsoft Foundry	API	Llama-3.3-70B-Instruct	9

Note. GPT and Llama models were accessed through Microsoft Foundry, but the model provider column refers to the underlying model provider rather than the hosting platform. For DeepL, Google Translate, and Gemini 3 Pro, the exact model version / snapshot is not explicitly exposed in the web interface.

Table 2: Generation settings used for the models accessed via API. For `azure-translator`, temperature, top-p, seed, and reasoning effort do not apply because it is not an LLM API. Reasoning effort does only apply for 'gpt-5.2' because that's the only reasoning model included.

Model	API path	Temp.	Top-p	Seed	Reasoning effort	Token limit
azure-translator	Microsoft Translator Text API	—	—	—	—	—
gpt-4	Chat Completions API	0	1	0	—	2048
gpt-5.2	Chat Completions API	0	1	0	none	2048
Llama-3.3-70B-Instruct	Chat Completions API	0	1	0	—	2048
Meta-Llama-3.1-8B-Instruct	Chat Completions API	0	1	0	—	2048
Meta-Llama-3.1-405B-Instruct	Chat Completions API	0	1	0	—	2048

Access routes and generation settings. All LLM systems except Gemini were accessed via Microsoft Foundry using API-based inference; the exact Foundry model identifiers and exposed versions are listed in Table 1. Gemini 3 Pro was accessed through the Google Gemini web interface. DeepL and Google Translate were likewise accessed through their public web interfaces, while Azure Translator was accessed via the Azure Translator Text Translation REST API. For interface-based systems, the Dutch source text was pasted into the interface, and the resulting translation was copied and stored as plain text.

For each combination of source text, target language, and system, we generated one translation output and stored it as a plain-text file. Table 2 reports the generation settings for models accessed via API. For API-accessed LLMs, we aligned configurable settings where possible to improve comparability: temperature was set to 0, top-p to 1, seed to 0, and the token limit to 2048. Azure Translator did not expose comparable LLM-style sampling controls, and for interface-based systems (Gemini 3 Pro, DeepL, and Google Translate), such parameters were not exposed or user-configurable.

Repeated generations were not performed, so run-to-run variance was not estimated directly. Residual stochasticity therefore remained a potential confound, especially for systems whose decoding settings were not fully controllable. No human post-editing or content correction was applied after generation. Outputs were decoded and saved as UTF-8 plain-text files.

To support consistent scoring and paragraph alignment, we applied only minimal formatting normalization. When a model added Markdown markers (e.g., ** around paragraph titles), these markers were removed so that the stored output consisted of plain text only. In Track B, where paragraph-level alignment was required, we additionally restored explicit paragraph separation when necessary: if paragraph content appeared on separate lines but without an empty line between paragraphs, a single empty line was manually inserted to match the paragraph layout of the Dutch source document. No punctuation was added, removed, or corrected, and no lexical content was changed.

Track-specific generation. In Track A, each of the six excerpts was translated into each target language by GPT-4 and Azure Translator. The LLM translations were generated using a structured prompt designed to preserve meaning, register/style, and readability (Appendix B, Figure B.1).

Table 3: Translation task matrix (evaluation track × target language × system). Column bands group systems into commercial NMT services vs. LLM-based translators. A checkmark indicates that a system was evaluated for that language within the specified track.

Evaluation track	Target language	NMT			LLM					
		Azure Transl.	DeepL	Google Transl.	GPT-4	Gemini 3 Pro	GPT-5.2	Llama 3.3 70B	Llama 3.1 405B	Llama 3.1 8B
Reference-free (rubric-based)	Turkish	✓			✓					
	Spanish	✓			✓					
	English	✓			✓					
	Polish	✓			✓					
	Ukrainian	✓			✓					
	Arabic	✓			✓					
	German	✓			✓					
Reference-based (metric-based)	Turkish	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Spanish	✓	✓	✓	✓	✓	✓	✓	✓	✓
	English	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Polish	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Ukrainian	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Arabic	✓	✓	✓	✓	✓	✓	✓	✓	✓
	German	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note. Reference-free: Azure Translator vs. GPT-4, scored with native-speaker rubric ratings and GPT-4-as-a-judge using the same rubric. Reference-based: all systems compared against professional references using BLEU, chrF++, TER, BERTScore, and COMET; 95% bootstrap confidence intervals are computed by segment resampling.

In Track B, each of the nine systems translated the fixed five-paragraph Dutch document into each target language. These outputs were then used for reference-based scoring and diagnostics. Table 3 summarizes the full evaluation matrix (track × language × system).

3.4 Track A: Reference-free evaluation

Evaluation materials and anonymized presentation. Human evaluation was administered through a spreadsheet workbook. The first sheet contained a welcome message, written instructions, and the scoring rubric. Separate language-specific sheets were provided for each target language. On each language sheet, evaluators saw the Dutch source text together with two anonymized target-language outputs in a fixed side-by-side layout. The outputs were labeled *Vertaling 1* and *Vertaling 2*; system names were hidden in the workbook. Next to each translation were fields for three numeric scores and an optional free-text comment.

This presentation format was chosen to make the task manageable for participants. Presenting the Dutch source and both candidate translations in a single view reduced navigation within the workbook and supported efficient completion. Evaluators completed the scoring individually and remotely, using the digital workbook and returning the completed file once finished.

At the same time, presenting both candidate translations on the same sheet may have allowed evaluators to use one version to disambiguate or recover from problems in the other. The Track A results should therefore be interpreted in light of this presentation format.

Rubric and scoring criteria. Table 4 presents the evaluation rubric used in Track A, translated from the Dutch evaluation workbook. The rubric operationalized three criteria through plain-language guiding questions and anchored score bands so that participation did not require formal training in translation studies. Evaluators scored each anonymized translation on a 1–10 scale for (i) *correctness*, (ii) *register*, and (iii) *readability*. For each criterion, the rubric provided verbal anchors for score ranges 1–3, 4–6, 7–9, and 10. The instruction sheet asked evaluators to compare each translation with the Dutch source text, enter the three scores in the adjacent columns, and optionally add a brief qualitative comment. We report *Overall* as the mean of the three criterion scores.

Table 4: Track A human-evaluation rubric, translated and adapted from the evaluator workbook.

Criterion	Guiding questions shown to evaluators	Score 1–3	Score 4–6	Score 7–9	Score 10
Correctness	Are the important information and action points preserved? Is the translated content correct and complete? Is terminology translated consistently, including pension-specific and fund-specific terms?	The translation is so inaccurate that the original message is no longer recognizable.	There are substantial translation errors that may distort the content or meaning.	The translation is generally correct, with only minor errors that do not affect overall comprehensibility. Some subtleties of the source text may be lost.	The translation is accurate and preserves the full meaning of the original. There are no errors in terminology, grammar, or punctuation.
Register	Is the intended level of formality preserved? Are cultural and contextual nuances handled appropriately? Is the degree of urgency the same as in the source text?	The translation uses a completely different tone and style, causing the original feel and communicative purpose to be lost.	There are noticeable deviations in tone and style, which may make the text less appropriate for the intended audience.	Tone and style are generally preserved, though some subtle differences remain. The text still reads fluently and appropriately.	The translation preserves not only the meaning but also the tone and style of the original. It sounds natural and appropriate in the target language.
Readability	Is the same level of readability as in the original preserved? Does the translated text have the same flow as the original? Does the translation contain grammatical or other linguistic errors?	The translation is incoherent and difficult to understand.	The text is difficult to read because of confusing sentence structure or unnatural word choice.	The translation is generally readable, though some parts could be structured better for a smoother reading experience.	The translation is fluent and easy to read. It reads as if written by a native speaker.

Evaluators and recruitment. For each target language, we recruited 1–3 evaluators who were native speakers of the target language and also proficient in Dutch. Dutch proficiency was required because the evaluation workbook displayed the Dutch source text alongside the target-language translations, and the task required direct assessment of meaning preservation relative to the source. All participating evaluators were recruited from the same external stakeholder in the Dutch pension sector that also helped define the language set for this pilot study.

Because the evaluators worked for the stakeholder organization from which the source materials and language priorities were derived, they were familiar with this type of client-facing pension communication. The study therefore relied on native-speaker judgments from domain-familiar readers who could compare the target text directly with the Dutch source.

The number of evaluators per language was not fully balanced. Due to recruitment constraints, it was not feasible to secure the same number of qualified native-speaker evaluators for every language. For this pilot study, we therefore prioritized obtaining at least one evaluator per language over enforcing a fixed per-language quota. Evaluators worked independently and received the same workbook, rubric, and written instructions, but no formal calibration session was conducted. Participation was voluntary with informed consent, and results are reported only in anonymized form. Evaluator counts per language are reported in Table 5.

Rating protocol. For each (excerpt, language), evaluators were shown the Dutch source text together with two anonymized candidate translations in the same worksheet view. They assigned 1–10 scores for meaning correctness and completeness, register and style, and readability and flow to each translation separately. They could also provide short qualitative comments. Evaluators were not asked to compare the two candidate translations directly; system comparison was introduced only in the analysis stage, where scores for the two systems were compared within the same excerpt-language item. Because Track A used a consistent side-by-side presentation format rather than a counterbalanced design, resulting score differences should be interpreted within that evaluation setup.

Table 5: Number of native-speaker human evaluators per target language used for translation quality assessment.

Target language	# Evaluators
Turkish	3
Spanish	1
English	1
Polish	2
Ukrainian	2
Arabic	2
German	2
Total	**13**

Aggregation and uncertainty. Because the number of evaluators varied by language, we first computed excerpt-level mean rubric scores by averaging across evaluators for each (language, system, excerpt). All subsequent summaries were then computed from these excerpt-level means so that excerpts contributed equally within each language and languages with more evaluators did not receive greater weight.

For each (language, system), we report the mean and standard deviation across the six excerpts for each criterion and for Overall. To compare GPT-4 with Azure within a language, we computed paired excerpt-level deltas (GPT-4 minus Azure) and reported mean deltas. Uncertainty was estimated via nonparametric bootstrap resampling over the six excerpts (95% CIs). Statistical significance was assessed using an exact sign-flip test on paired excerpt deltas (two-sided; $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$).

Macro-averaging across languages. We report macro-averaged paired deltas as the unweighted mean of the per-language mean deltas, so that each language contributes equally. Macro-level 95% confidence intervals were estimated by bootstrap resampling over languages.

LLM-as-a-judge. In parallel with human evaluation, we used a GPT-4-based judge that received the Dutch source text and one target-language translation with system identity withheld. The judge applied the same rubric and returned three 1–10 criterion scores; we additionally report Overall as their mean. Human ratings were first aggregated by averaging evaluators for each (language, system, excerpt) item. Judge scores were then compared with these excerpt-level human means across all items using Spearman rank correlation, mean absolute error (MAE), root mean squared error (RMSE), and a least-squares calibration fit of judge scores on human scores. To assess whether the judge preserved the same system-level contrast as the human evaluators, we also compared judge- and human-derived paired excerpt-level deltas (GPT-4 minus Azure) within each language. The full judge prompt is provided in Appendix B (Figure B.2).

3.5 Track B: Reference-based evaluation

Professional references and alignment. Each target language was associated with a single professional reference translation of the Track B letter. Because these references preserved the original five-paragraph structure, they enabled both strict paragraph-level alignment and document-level scoring. All system outputs and references therefore contained exactly five paragraphs, and we enforced strict one-to-one paragraph alignment with no merging or splitting.

Because Track B was based on a single short document, the resulting automatic evaluation should be interpreted as complementary to the broader reference-free evidence from Track A rather than as a standalone basis for general conclusions.

We did not perform additional sentence-level alignment. Although paragraph boundaries were stable, sentence segmentation varied across target-language references and system outputs, so reliable sentence-level alignment would have required extra manual or language-specific preprocessing beyond the scope of this pilot study.

Reference-based metrics. We evaluated reference-based translation quality with five metrics. For lexical-overlap and edit-based scoring, we computed BLEU, chrF++, and TER using `sacreBLEU` v2.6.0. To make

Table 6: Reference-based metrics and exact scoring configurations used in Track B.

Metric	Implementation	Signature / model configuration
BLEU	`sacreBLEU` v2.6.0	`nrefs:1|case:mixed|eff:yes|tok:intl|smooth:exp|version:2.6.0`
chrF++	`sacreBLEU` v2.6.0	`nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.6.0`
TER	`sacreBLEU` v2.6.0	`nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.6.0`
BERTScore	`bert-score`	F1; `xlm-roberta-large`; no baseline rescaling
COMET	`COMET`	`Unbabel/wmt22-comet-da`; checkpoint `2760a223ac957f30acfb18c8aa649b01cf1d75f2`

the evaluation setup fully reproducible, the exact metric signatures and model identifiers are reported in Table 6. We further computed BERTScore (F1) using `xlm-roberta-large` without baseline rescaling, and COMET using `Unbabel/wmt22-comet-da`. For reporting, BERTScore and COMET were multiplied by 100 and are therefore shown on a 0–100 scale.

Aggregation and macro-averaging. For each (language, system), we computed both paragraph-level scores and document-level scores. For BLEU, chrF++, and TER, document scores were computed corpus-style over the five aligned paragraph units; paragraph scores were retained for dispersion and worst-segment analyses. For BERTScore and COMET, scores were computed separately for each aligned paragraph, and the document score was defined as the mean of the five paragraph scores. Overall performance is summarized using macro-averaged document-level scores across the seven languages, calculated as an unweighted mean over languages.

Bootstrap confidence intervals. Given the five aligned paragraphs per language, uncertainty was estimated by nonparametric bootstrap resampling over paragraphs for each (language, system, metric). We used 3125 bootstrap samples with fixed seed 42. For BLEU, chrF++, and TER, the corpus-level score was recomputed for each bootstrap resample. For BERTScore and COMET, the mean of the resampled paragraph-level scores was recomputed.

Operational probes and within-document stability. To surface omission and formatting risks relevant to regulated client communication, we computed lightweight paragraph-level probes in addition to the main reference-based metrics. The probes target three operational risks: length ratio approximates compression or verbosity relative to the reference, acronym-preservation F1 checks whether protected uppercase acronyms are retained exactly, and punctuation preservation checks whether punctuation and formatting counts remain close to the reference. All probes were computed on each aligned system–reference paragraph pair.

Length ratio was defined as the normalized character length of the system paragraph divided by the normalized character length of the reference paragraph:

$$\mathrm{LenRatio}(s, r) = \frac{|s|_{\mathrm{char}}}{|r|_{\mathrm{char}}},$$

where s is the normalized system paragraph and r is the normalized reference paragraph. Values below 1 indicate shorter outputs than the reference and may signal compression or omission, while values above 1 indicate longer outputs and may signal expansion or verbosity.

Acronym preservation was computed as exact-match set-based F1 over Latin-script uppercase acronym tokens extracted from the system and reference paragraphs. Acronyms were identified using the regular expression

$$\texttt{\textbackslash b[A-Z]\{2,\}(?:/[A-Z]\{2,\})*\textbackslash b},$$

which captures forms such as `UWV`, `WAO`, and `WAO/WIA`. Precision and recall were computed over the system and reference acronym sets, and F1 was set to 1.0 when both sets were empty. Because this extraction pattern only targets uppercase Latin-script forms, the acronym probe should be interpreted as a lightweight diagnostic and may be less reliable for non-Latin or mixed-script realizations.

Punctuation preservation was computed from counts over a list of punctuation and formatting characters, including quotes, brackets, percentage signs, slashes, colons, semicolons, commas, periods, exclamation and

question marks, ellipses, and hyphen/dash variants. For each aligned paragraph pair, we computed the total absolute count difference across punctuation types and converted it into a bounded preservation score:

$$\text{PunctPres}(s, r) = \max\left(0, \min\left(1, 1 - \frac{\text{AbsDiff}(s, r)}{\text{RefPunctTotal}(r) + 1}\right)\right).$$

A score of 1.0 therefore indicates identical punctuation counts, while lower values indicate larger punctuation-count mismatches relative to the reference. We computed the absolute punctuation-count difference as a supplementary diagnostic and reported within-document variability as the standard deviation of paragraph-level metric scores for each (language, system).

3.6 Artifact availability and reproducibility

To support reproducibility, we released the materials used in this study in a public repository [29]. The repository contains the analysis code, prompts, scoring scripts, table and figure generation code, and text materials. For exact replication, we recommend using the archived release associated with this study.

4 Results

We report results from two complementary tracks. **Track A (reference-free)** measures enterprise acceptability directly via native-speaker rubric ratings of meaning correctness, register and style, and readability, with GPT-4 also used as an auxiliary *LLM judge*. **Track B (reference-based)** benchmarks nine systems against professional references on a controlled five-paragraph document using standard automatic metrics plus lightweight fidelity probes. We then triangulate the two tracks through their shared GPT-4 versus Azure comparison and close by summarizing the evidence for the four research questions. Throughout, we distinguish GPT-4 used as a translation system ("GPT-4 translator") from GPT-4 used as an evaluator ("GPT-4 judge").

4.1 Track A: Reference-free

Track A covers 6 excerpts × 7 languages × 2 systems ($n = 84$ language–excerpt–system items). We first report the GPT-4 vs. Azure contrast under *native-speaker evaluation*, then show how these human-rated effects vary by language. We subsequently assess whether the GPT-4 judge provides a usable proxy signal and summarize the qualitative error patterns that explain the rubric scores.

4.1.1 Main effect and robustness across languages

Table 7a summarizes native-speaker rubric means aggregated over all 84 items. Under human evaluation, GPT-4 translator scores higher than Azure on all rubric dimensions, with the largest separation on *Register and style*, indicating more consistent preservation of institutional register and letter-like style.

To quantify whether this advantage is robust across target languages, Table 8 reports macro-averaged paired deltas based on native-speaker ratings (unweighted across languages). GPT-4 yields consistent gains: Meaning correctness +1.90 (95% CI [1.37, 2.42]), Register and style +2.28 ([1.12, 3.62]), Readability +1.77 ([1.24, 2.40]), and Overall +1.98 ([1.31, 2.64]). All four macro deltas are statistically different from zero under a sign-flip test (Table 8), indicating that the advantage is not driven by a single language.

4.1.2 Per-language variation and dispersion

Figure 1a visualizes per-language mean rubric scores and dispersion (SD over the six excerpts) under native-speaker evaluation, and Table 9 reports within-language paired deltas (GPT-4 minus Azure) with excerpt-level bootstrap CIs. Judge-based results are reported separately in Section 4.1.3.

GPT-4 improves Overall in *all seven* languages, but effect sizes vary. Spanish shows the largest register/style gain (+5.83) and a large Overall gain (+3.11), indicating substantial improvements in institutional register and pragmatic fit beyond literal adequacy. Turkish shows a similarly large Overall gain (+3.07), with the largest readability gain (+3.39). German also shows a substantial Overall gain (+2.44), while Arabic

Table 7: Track A (reference-free): rubric means aggregated over all language×excerpt×system items ($n = 84$: 7 languages × 6 excerpts × 2 systems), reported separately for native-speaker evaluation (a) and GPT-4 judge evaluation (b). Scores are on a 1–10 scale for Correctness, Register and style, and Readability; Overall is the mean of the three. Aggregation uses excerpt-level means per (language, system) so that languages with more evaluators do not receive greater weight.

(a) Native-speaker scores

System	Corr.	Register	Readability	Overall
Azure Translator	6.0	5.7	6.1	5.9
GPT-4 (translator)	7.9	8.0	7.9	7.9

(b) GPT-4 judge scores

System	Corr.	Register	Readability	Overall
Azure Translator	6.5	5.9	6.8	6.4
GPT-4 (translator)	8.5	8.0	8.4	8.3

Table 8: Track A (reference-free): macro-averaged paired deltas (GPT-4 minus Azure Translator) across target languages ($N_{\text{lang}} = 7$). Within each language, deltas are computed on excerpt-level mean scores from native-speakers (paired by excerpt; 6 excerpts per language). Macro deltas are the unweighted mean over languages. 95% CIs are bootstrap over languages; p values are from an exact sign-flip test over language-level deltas. Positive values favor GPT-4.

Metric	Mean Δ	95% CI	p (sign-flip)	Cohen's d_z	N_{lang}
Correctness	1.90	[1.37, 2.42]	0.0156	2.42	7
Register and style	2.28	[1.12, 3.62]	0.0312	1.21	7
Readability	1.77	[1.24, 2.40]	0.0156	2.08	7
Overall	1.98	[1.31, 2.64]	0.0156	2.05	7

(+1.83), Polish (+1.67), and Ukrainian (+1.31) show smaller but consistently positive gains. English shows the smallest gap (Overall +0.44) and is not significant at $p < 0.10$ (Table 9), consistent with both systems performing relatively strongly on this direction for the sampled excerpts.

Across languages, the same broad pattern visible in the macro analysis remains: GPT-4's advantage is positive in every language on the aggregate Overall measure, with the strongest macro-level separation in register/style. At the same time, the language-level results show that the dominant criterion differs by language: register/style drives the especially large Spanish gain, readability is strongest in Turkish, and correctness is particularly large in German. The native-speaker results also show that excerpt-level dispersion varies by language and criterion, but without changing the overall pattern that GPT-4 is preferred to Azure in each language on Overall.

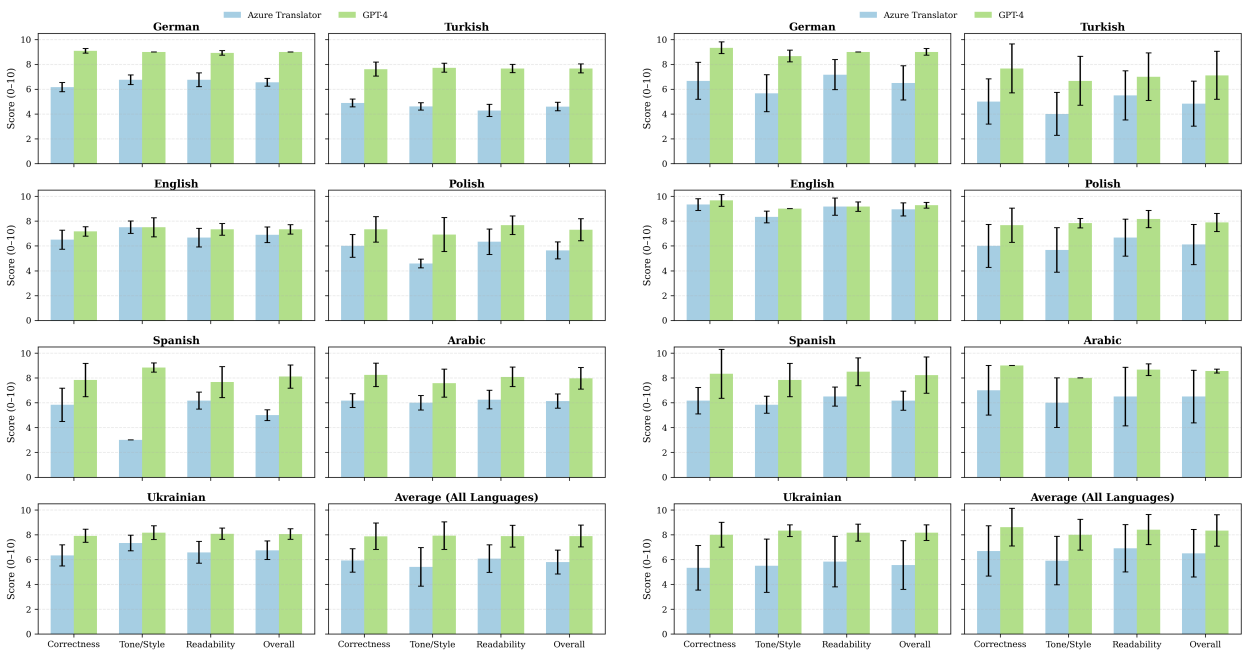
4.1.3 LLM-as-a-judge: agreement with native speakers

Table 7b summarizes GPT-4 judge rubric means aggregated over all 84 items, and Figure 1b shows the corresponding per-language means and dispersion.

Across all $n = 84$ items, agreement between judge scores and native-speaker means is moderate (Overall Spearman $\rho = 0.53$; MAE = 1.35; Table 10). When computed separately by translation system, Overall

Table 9: Track A (reference-free): native-speaker paired deltas by language (GPT-4 minus Azure Translator). Deltas are computed on excerpt-level mean rubric scores (paired by excerpt; 6 excerpts per language). Brackets show 95% bootstrap CIs over excerpts within language. Significance markers from an exact sign-flip test on paired excerpt deltas: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$. Positive values favor GPT-4.

Language	Correctness Δ	Register and style Δ	Readability Δ	Overall Δ
Arabic	2.08** [1.25, 3.00]	1.58* [0.67, 2.42]	1.83** [1.25, 2.42]	1.83** [1.14, 2.56]
English	0.67 [-0.17, 1.33]	0.00 [-0.67, 0.67]	0.67 [-0.17, 1.33]	0.44 [-0.11, 1.06]
German	2.92** [2.75, 3.00]	2.25** [2.00, 2.58]	2.17** [1.67, 2.67]	2.44** [2.22, 2.72]
Polish	1.33** [0.67, 2.00]	2.33* [1.17, 3.50]	1.33* [0.58, 2.08]	1.67** [0.94, 2.44]
Spanish	2.00* [1.00, 2.83]	5.83** [5.50, 6.00]	1.50 [0.17, 2.67]	3.11** [2.44, 3.72]
Turkish	2.72** [2.17, 3.28]	3.11** [2.67, 3.44]	3.39** [2.89, 3.83]	3.07** [2.61, 3.44]
Ukrainian	1.58** [1.00, 2.08]	0.83* [0.42, 1.17]	1.50* [0.75, 2.17]	1.31** [0.75, 1.75]



(a) Native-speaker evaluation

(b) GPT-4 judge evaluation

Figure 1: Track A (reference-free): rubric results by language (0–10), reported separately for native-speaker evaluation (a) and GPT-4 judge evaluation (b). Within each plot, bars show mean scores for Azure Translator and GPT-4 translator for Correctness, Register and Style, Readability, and Overall (mean of the three). Error bars indicate ±1 SD across the six excerpts within each language, highlighting within-language variability across text types.

Table 10: Track A (reference-free): agreement between GPT-4 judge scores and native-speaker excerpt-level means across all language×excerpt×system items ($n = 84$). We report Spearman ρ and absolute error metrics (MAE, RMSE). The linear fit $\hat{y} = a + bx$ models judge score (y) as a function of human score (x); $b < 1$ indicates score compression, and $a \neq 0$ indicates systematic offset.

Metric	n	Spearman ρ	MAE	RMSE	Slope b	Intercept a
Correctness	84	0.48	1.57	1.95	0.69	2.70
Register and style	84	0.57	1.42	1.94	0.52	3.40
Readability	84	0.40	1.36	1.82	0.57	3.57
Overall	84	0.53	1.35	1.73	0.71	2.45

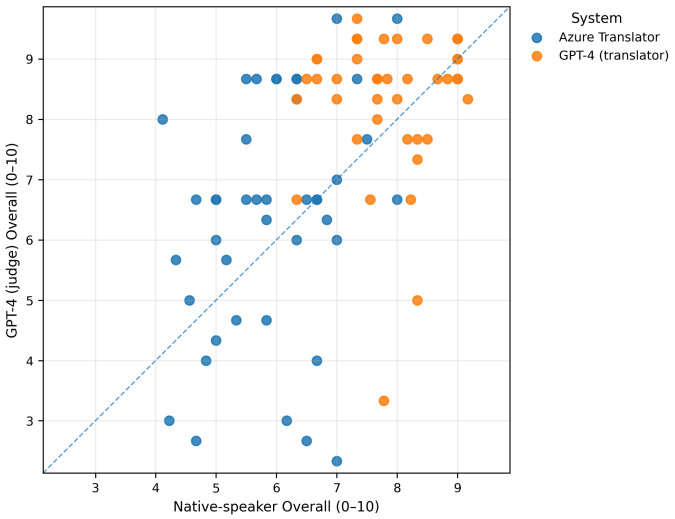


Figure 2: Track A (reference-free): GPT-4 judge vs. native-speaker Overall scores ($n = 84$; one point per language×excerpt×system). The dashed line indicates perfect agreement ($y = x$). Points above the line indicate cases where the judge scored the translation higher than the native-speaker mean; points below the line indicate lower judge scores. Deviations from the line reflect judge bias, including offset and score compression. Point color indicates the translation system.

agreement is higher for Azure Translator ($\rho = 0.31$, $n = 42$) than for GPT-4 translator outputs ($\rho = 0.09$, $n = 42$), indicating that the pooled correlation is partly driven by between-system score separation, not fine-grained within-system calibration. Linear fits show **score compression** ($b < 1$) for all four reported dimensions, meaning that the judge assigns scores over a narrower range than the native-speaker means. Positive intercepts further indicate an upward shift on several dimensions, especially for lower- and mid-scoring items (Table 10).

Figure 2 visualizes this calibration pattern for Overall. The judge often assigns relatively high scores to translations that native speakers rated more moderately, especially among lower-scoring Azure outputs, while some high-scoring GPT-4 translator outputs fall closer to or below the perfect-agreement line. This pattern is consistent with score compression and an upward shift in the judge scale. Together with the judge-based system summaries, this indicates that the GPT-4 judge broadly preserves the direction of the GPT-4 versus Azure contrast, but on a shifted and compressed scale and with weak within-system rank agreement, especially for GPT-4 translator outputs. The judge should therefore be interpreted as a coarse ranking or monitoring signal, not as a source of calibrated absolute acceptability scores.

4.1.4 Qualitative error taxonomy (enterprise drivers)

Native-speaker notes indicate that rubric penalties are driven by recurring, user-facing issues that matter in regulated pension communication: (i) register control (consistent formality, genre conventions, pragmatic softening) and (ii) terminology/entity governance (stable handling of pension concepts, protected abbrevi-

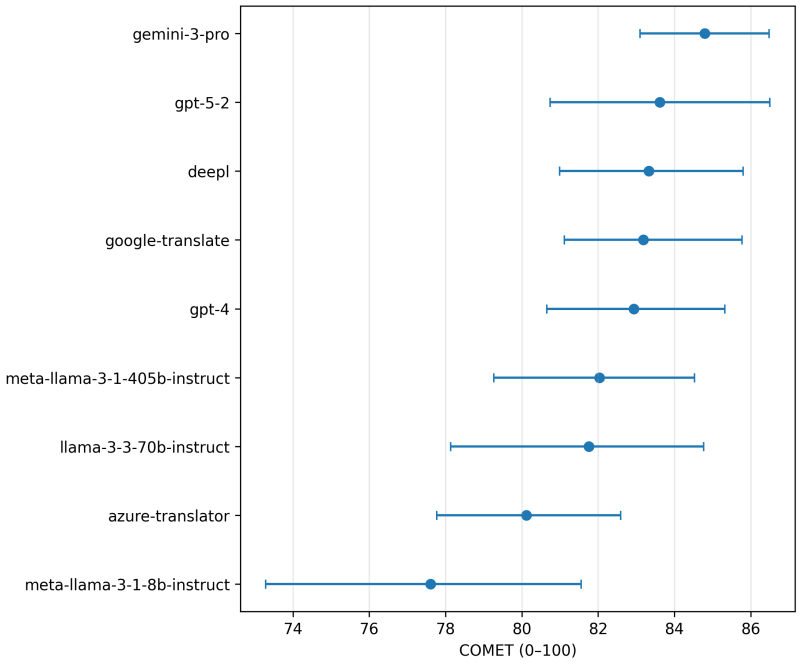


Figure 3: Track B (reference-based): macro COMET (0–100) by system. Points are macro-averaged over the seven target languages (unweighted). Error bars show 95% bootstrap confidence intervals obtained by resampling target languages with replacement, reflecting uncertainty due to language-to-language variation on the controlled document.

ations, and institution names). Table 11 summarizes *language-specific* issues with concrete Dutch triggers from Documents 1–6 and evaluator-cited examples of what went wrong (and, where explicitly provided by evaluators, what wording would be preferable). We preserve Dutch excerpts in the table, where they map directly to the cited issues.

4.2 Track B: Reference-based

Track B compares nine systems against professional reference translations for one controlled Dutch source document (five aligned paragraphs) across seven target languages. We report (i) a macro system overview, (ii) language effects and ranking sensitivity, and (iii) diagnostics that expose tail risk and localized failures.

4.2.1 System ranking overview (macro metrics and uncertainty)

Table 12 reports macro-averaged corpus-level metrics (0–100 for chrF++/BLEU/BERTScore/COMET). The top cluster under chrF++ is `gemini-3-pro` (56.27) and `gpt-5-2` (55.98), followed closely by `deepl` (54.67). Under COMET, `gemini-3-pro` ranks first (84.80), with `gpt-5-2` and `deepl` close behind (83.62 and 83.33). Figure 3 visualizes macro COMET with 95% bootstrap CIs over languages. The leading systems form a tight top cluster, but they are difficult to separate once language-level variation is accounted for. Because Track B contains only one short source document and seven language-level scores, a larger and more diverse test set would be needed to determine whether these small rank differences are stable.

4.2.2 Language effects and per-language winners

Performance varies substantially by target language (Table 13). Arabic and Ukrainian are hardest on average in this controlled setting (lowest Avg z-score), while German and Spanish are easiest (highest Avg z-score). Figure 4 shows the system-by-language chrF++ breakdown, illustrating cross-language heterogeneity (deviations from the macro ordering within a language column).

Table 11: Track A (reference-free): language-specific qualitative issues from native-speaker notes, shown separately for Azure Translator and GPT-4 (translator). "Dutch trigger" quotes are verbatim snippets from the Dutch sources (Documents 1–6) anchoring where an issue was observed. In "Evaluator notes + examples", **[quote]** marks evaluator-cited strings/tokens and **[pattern]** marks described issues without exact quoting. Rubric tags: (C)=Correctness, (T)=Tone/register and style, (R)=Readability. Evaluators scored anonymized outputs with system identity hidden.

Lang.	Dutch trigger (Doc.)	System	Evaluator notes + examples
DE	*"Beste meneer . . . "* (Doc. 3); *"Bij deze brief zit een folder."* (Doc. 1)	**Azure**	**[quote]** Mixed address forms (*Du/Sie*) within one letter. (T) **[pattern]** Inconsistent institutional phrasing (e.g., "folder/leaflet" variants). (T,R)
		GPT	**[pattern]** Generally accurate/usable; critique concerned genre norms (concise Dutch admin prose can feel "too short/blunt" in German official letters). (T,R)
TR	*". . . over uw arbeidsongeschiktheid."* (Doc. 2); *"bijzonder partnerpensioen"* (Doc. 5)	**Azure**	**[pattern]** Substantially weaker overall; frequent literal choices and terminology instability reduce usability. (C,R,T)
		GPT	**[quote]** Unnatural salutation *"Değerli Bay/Bayan"*; suggested *"Sayın"*. (T) **[quote]** *arbeidsongeschiktheid* rendered literally; suggested *"malulen emeklilik"*. (C) **[quote]** *bijzonder partnerpensioen* rendered with *"özel"* judged confusing/misleading. (C,T)
ES	*"U logt in met DigiD . . . "* (Doc. 1); *". . . (UWV) . . . arbeidsongeschiktheid."* (Doc. 2); URLs/structured strings (Doc. 1/4)	**Azure**	**[quote]** Mixed address strategy (*usted/tú*) within a document. (T) **[quote]** Unstable handling of abbreviations/entities (e.g., `UWV`). (C,R)
		GPT	**[pattern]** Strong meaning and tone overall; occasional over-literal handling of structured strings (e.g., translating URL slugs). (R) **[pattern]** Within-letter term drift for relationship/pension concepts. (C,T,R)
EN	*". . . over uw arbeidsongeschiktheid."* (Doc. 2); *"partnerpensioen"/"bijzonder partnerpensioen"* (Doc. 5/6); *"U komt te overlijden."* (Doc. 6)	**Azure**	**[pattern]** Understandable but often "very literal" and "choppy" in flow. (R)
		GPT	**[quote]** *arbeidsongeschikt* → *"disabled"* flagged; preferred *"inability/incapacity for work"*. (C) **[quote]** Framing shift: *"survivor's pension"* vs. *partner pension*. (T) **[quote]** Blunt *"when you die"* penalized; suggested *"when you pass away"*. (T,R)
PL	*"Beste mevrouw . . . "* (Doc. 5); *"Dat betekent dat u geld mist."* (Doc. 1); `PROEFDRUK` (Doc. 3/5)	**Azure**	**[quote]** Switch into informal address (*ty*) in an official letter. (T) **[pattern]** Literal-but-unidiomatic constructions; layout-string handling reduces acceptability. (R,T)
		GPT	**[quote]** Mixed formal strategies (e.g., *Pan/Pani* vs. *Państwo*). (T) **[pattern]** Inflection/agreement issues reduce professionalism and sometimes clarity. (R,C) **[quote]** Template artifact leakage/mishandling (e.g., `PROEFDRUK`). (R,T)
UK	*"Krijgt u pensioen van ons?"* (Doc. 4); *". . . vanaf 1 februari 2023."* (Doc. 3); *"U komt te overlijden."* (Doc. 6)	**Azure**	**[pattern]** Tense mistakes affecting timelines/obligations. (C) **[pattern]** Inconsistent benefit-term choices; blunt mortality phrasing reduces acceptability. (T,C)
		GPT	**[pattern]** Unnatural sentence formation/word order (incl. if-clauses as questions). (R) **[quote]** Formal-letter convention: capitalization of polite *Bu/Bau* expected. (T) **[pattern]** Sensitive-event phrasing criticized as overly grim; euphemistic register preferred. (T)
AR	*"U komt te overlijden . . . "* (Doc. 6); numbers/identifiers/phone fields (Doc. 3/5); administrative punctuation patterns (Doc. 3)	**Azure**	**[pattern]** Structural/grammatmatical issues (tense/word order; occasional repetition/looping) harm comprehensibility. (R,C) **[pattern]** Formatting/structure issues make output hard to follow. (R)
		GPT	**[pattern]** Higher acceptability, but overly direct mortality phrasing penalized vs. Dutch softer register. (T) **[pattern]** RTL presentation risks noted (numbers/identifiers ordering), affecting usability independent of adequacy. (R)

Table 12: Track B (reference-based): macro-averaged, corpus-level metrics across seven target languages for one controlled five-paragraph document (9 systems). chrF++/BLEU/BERTScore/COMET are scaled to 0–100 (higher is better); TER is 0–100 (lower is better; ↓). Operational probes (length ratio, acronym preservation F1, punctuation preservation) are macro-averaged over languages; length ratio closer to 1 indicates fewer compression/verbosity deviations relative to the reference. Best values per column are bolded (for TER, the minimum is bolded).

System	chrF++	BLEU	TER ↓	BERTScore F1	COMET	Len. ratio	Acr. F1	Punct. pres.
`gemini-3-pro`	**56.27**	36.25	52.31	94.90	**84.80**	**0.896**	0.893	**0.662**
`gpt-5-2`	55.98	**36.47**	**51.69**	**95.08**	83.62	0.867	0.893	0.607
`deepl`	54.67	34.79	53.17	94.66	83.33	0.871	0.893	0.634
`gpt-4`	51.18	29.39	57.83	94.31	82.93	0.887	0.846	0.581
`google-translate`	49.65	27.75	60.43	93.72	83.18	0.882	0.893	0.619
`meta-llama-3-1-405b-instruct`	49.56	27.16	58.67	94.35	82.03	0.846	0.893	0.510
`azure-translator`	48.83	26.86	60.31	94.00	80.12	0.866	0.869	0.533
`llama-3-3-70b-instruct`	48.28	26.54	61.33	93.96	81.75	0.865	0.869	0.569
`meta-llama-3-1-8b-instruct`	42.60	20.48	69.31	93.15	77.60	0.868	**0.897**	0.502

Table 13: Track B (reference-based): mean document-level metric scores by target language, averaged across the nine systems, with an aggregate difficulty index. Avg z-score is the mean of per-metric z-scores across languages (TER inverted so higher is better for all metrics). Languages are ordered from hardest (lowest Avg z-score) to easiest (highest Avg z-score) for this controlled five-paragraph document.

Rank	Language	BLEU	chrF++	TER↓	BERTScore	COMET	Avg z-score
1	Arabic	14.59	36.72	75.21	91.91	76.90	-1.39
2	Ukrainian	14.61	38.35	71.76	92.66	80.18	-1.00
3	Turkish	23.71	44.12	71.58	94.01	81.78	-0.46
4	Polish	28.58	51.30	60.90	93.50	85.61	0.09
5	English	40.10	57.88	47.20	95.18	81.20	0.52
6	German	36.82	61.71	46.24	95.81	86.63	0.99
7	Spanish	48.23	65.39	35.48	96.60	82.76	1.25

Table 14 reports per-language winners under chrF++ vs. COMET. In four of seven languages, the best system is the same under both metrics; in Turkish, Spanish, and Ukrainian, the winner differs, reflecting metric-driven rank sensitivity in a single-reference setting. The Ukrainian COMET winner is a notable anomaly: `meta-llama-3-1-8b-instruct` scores highest by COMET (81.91) despite being the weakest system on the main macro reference-based metrics in Table 12. Given the five-paragraph sample, we interpret this as a COMET-specific point-estimate anomaly rather than as evidence that the 8B model is the strongest Ukrainian system.

4.2.3 Metric sensitivity: chrF++ vs. COMET

Figure 5 plots corpus chrF++ vs. COMET for all language×system points. Outliers indicate cases where character overlap and learned semantic evaluation disagree (e.g., acceptable paraphrase with low overlap,

Table 14: Track B (reference-based): per-language winner by corpus chrF++ (character overlap) and by COMET (learned semantic evaluation). Differences indicate metric-driven rank sensitivity in a single-reference setting; "Different?" marks languages where the top system depends on metric choice.

Language	Best by chrF++	chrF++	Best by COMET	COMET	Different?
Turkish	`deepl`	50.56	`google-translate`	85.76	Yes
Spanish	`deepl`	72.17	`gpt-5-2`	85.11	Yes
English	`gemini-3-pro`	61.73	`gemini-3-pro`	81.96	No
Polish	`google-translate`	63.62	`google-translate`	89.75	No
Ukrainian	`gpt-5-2`	44.47	`meta-llama-3-1-8b-instruct`	81.91	Yes
Arabic	`gemini-3-pro`	48.66	`gemini-3-pro`	84.51	No
German	`gpt-5-2`	70.24	`gpt-5-2`	88.78	No

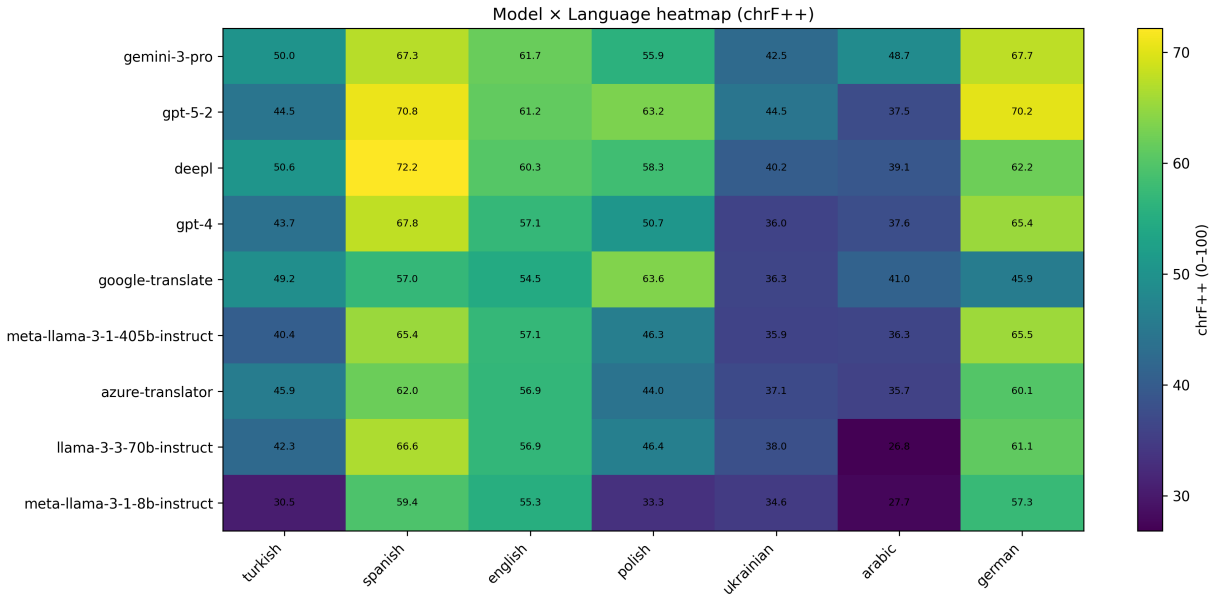


Figure 4: Track B (reference-based): corpus-level chrF++ (0–100; higher is better) by system (rows) and target language (columns). Rows are ordered by macro-average chrF++ across languages. Cell values are annotated with exact scores; within-column deviations from the row ordering visualize language-specific rank changes (cross-language heterogeneity).

or overlap-heavy output with weaker semantic adequacy). Table 15 quantifies how often large character–semantic disagreements occur at the paragraph level. Disagreement is highest in Spanish (15.6%), and elevated in Arabic and Turkish (8.9%), suggesting that metric choice can materially affect conclusions for these directions.

4.2.4 Risk diagnostics: tail behavior, probes, and worst segments

Corpus-level averages can mask concentrated failures. Figure 6 shows the distribution of paragraph-level chrF++ scores pooled across languages, highlighting cross-paragraph variability within systems.

Operational probes (Table 16; Figure 8) surface practical risks beyond overlap and semantic scoring. At the language level, the strongest length-compression signals appear in Ukrainian (mean length ratio 0.732; length absolute deviation 0.268) and Arabic (0.779; 0.221). Arabic also has the weakest punctuation preservation (0.396), while Ukrainian has unusually low acronym preservation (0.407), consistent with the Latin uppercase acronym extraction used by the probe.

Table 15: Track B (reference-based): rate of large character–semantic metric disagreement at paragraph level, defined as $|z_{\text{lex}} - z_{\text{sem}}| \geq 1$. Left: disagreement rate aggregated per system over 35 paragraph pairs per system (5 paragraphs × 7 languages). Right: disagreement rate aggregated per language over 45 paragraph pairs per language (9 systems × 5 paragraphs).

System	Disagree. rate
gpt-4	11.4%
gpt-5-2	11.4%
meta-llama-3-1-8b-instruct	8.6%
deepl	5.7%
llama-3-3-70b-instruct	5.7%
meta-llama-3-1-405b-instruct	5.7%
azure-translator	2.9%
gemini-3-pro	2.9%
google-translate	2.9%

Language	Disagree. rate
Spanish	15.6%
Arabic	8.9%
Turkish	8.9%
English	4.4%
German	2.2%
Polish	2.2%
Ukrainian	2.2%

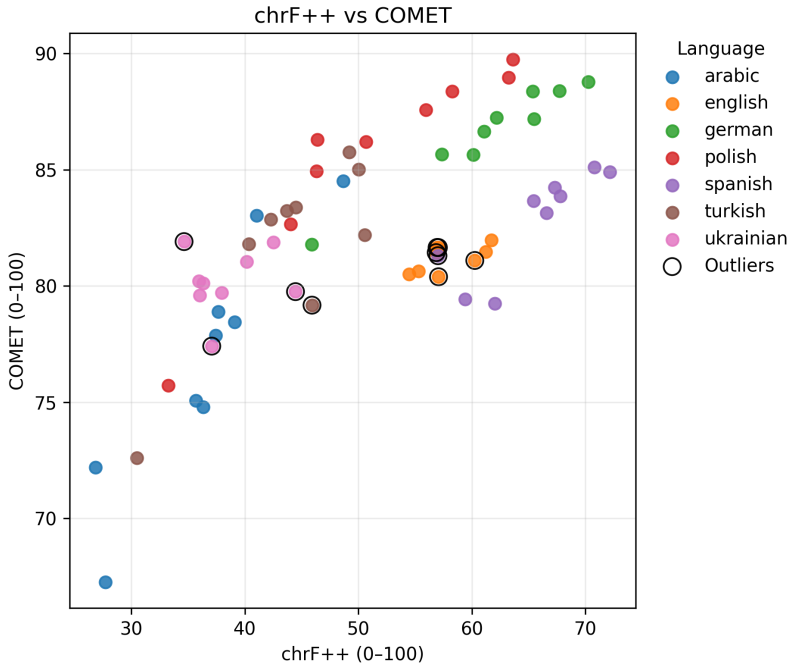


Figure 5: Track B (reference-based): corpus-level chrF++ vs. COMET (both 0–100). Each point is a language×system pair ($7 \times 9 = 63$ points); colors indicate target language. Black rings highlight the largest within-language chrF++–COMET disagreements, which can reflect paraphrase tolerance (high COMET, lower overlap) or overlap without semantic adequacy (high chrF++, lower COMET).

Worst-segment localization further shows that low overlap is concentrated in specific document parts. Paragraphs 3 and 5 most often drive the within-document minimum chrF++ (Figure 9), with paragraph 3 accounting for the largest share of minima. Table 17 supports this pattern: paragraph 3 has the lowest mean chrF++ in English, Ukrainian, Polish, and German, and is close to the lowest value in Arabic. This suggests reference-sensitive handling of institutional acronyms, legal descriptions, and compressed administrative phrasing. Manual inspection of the paragraph-level outputs indicates that this pattern is especially visible in English and Ukrainian, where expanded reference formulations for `WAO/WIA`-related content contrast with shorter acronym-preserving system outputs. Because chrF++ is based on character n-gram overlap, such differences can produce low scores even when the meaning is broadly similar.

Paragraph 5 reflects a different difficulty pattern. It contains an obligation under a condition and links that condition to a consequence for the contribution amount. This makes the paragraph sensitive to paraphrase and register. The data support this interpretation: paragraph 5 has the lowest mean chrF++ for Turkish in Table 17, and it is the single worst localized paragraph for German and Polish in Table 18. Paragraph 2 is also diagnostically important because it is terminology-dense and shows low mean chrF++ in several languages, especially Spanish, Arabic, Ukrainian, and Polish. Overall, the worst-paragraph frequency plot in Figure 9 identifies localized review points and metric-sensitive segments; it does not show that every low-scoring paragraph is a substantive translation failure.

Table 18 helps distinguish possible failure types. Some worst cases are consistent with omission-like behavior, as indicated by very low length ratios, such as Ukrainian paragraph 3 for `azure-translator` with length ratio 0.46. Others point to formatting drift, as indicated by punctuation preservation of 0.00 in Arabic paragraph 3 and German paragraph 5.

4.3 Cross-track triangulation (GPT-4 vs. Azure)

Because the tracks evaluate different source texts (six excerpts vs. one controlled five-paragraph document), cross-track comparison is triangulation rather than replication. Still, the shared contrast GPT-4 vs. Azure

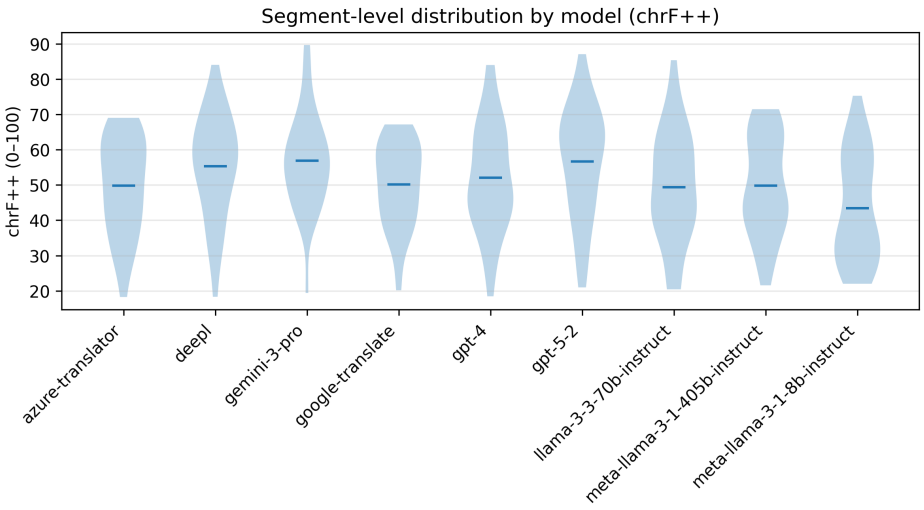


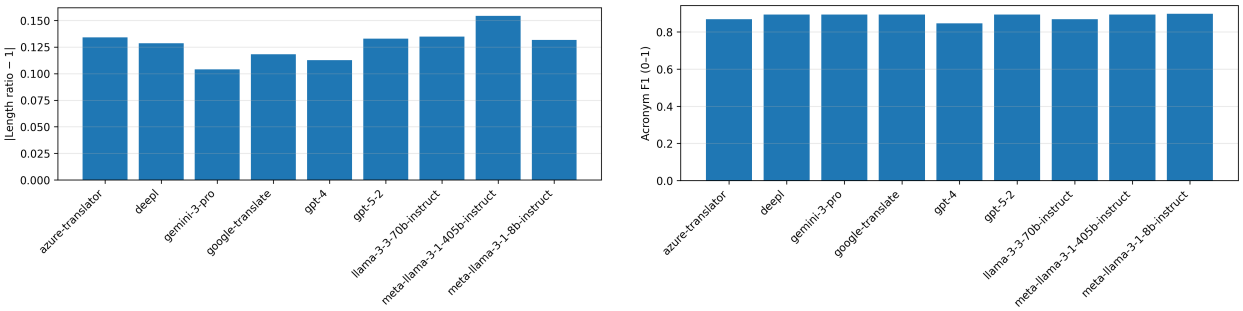
Figure 6: Track B (reference-based): distribution of paragraph-level chrF++ (0–100) per system, pooled across all target languages (35 paragraph scores per system = 7 languages × 5 paragraphs). Violin width reflects score density; the mean is shown for each system. This view highlights within-document tail risk that corpus-level averages can obscure.

Table 16: Track B (reference-based): operational probe averages by language (macro-averaged across the nine systems; 5 paragraphs per language). Len. ratio is candidate/reference character length; Len. abs. dev. is $|1 - \text{Len. ratio}|$ (lower is better). Acr. F1 measures acronym preservation; Punct. pres. measures punctuation preservation. Numeral preservation F1 was 1.00 for all languages/systems on this dataset and is omitted.

Language	Len. ratio	Len. abs. dev. ↓	Acr. F1	Punct. pres.
Turkish	0.899	0.101	1.000	0.586
Spanish	0.960	0.040	1.000	0.579
English	0.867	0.133	0.920	0.728
Polish	0.893	0.107	1.000	0.525
Ukrainian	0.732	0.268	0.407	0.641
Arabic	0.779	0.221	0.957	0.396
German	0.975	0.025	0.896	0.602

Table 17: Track B (reference-based): mean paragraph-level chrF++ (0–100) by target language and paragraph, averaged across the nine systems. Lower values indicate paragraphs that are systematically harder under overlap-based scoring, helping localize where failures concentrate within the controlled document.

Language	P1	P2	P3	P4	P5
Turkish	50.23	47.06	51.34	42.50	32.23
Spanish	69.24	60.85	74.96	63.94	63.45
English	66.44	55.54	41.58	65.32	62.85
Polish	64.28	48.47	45.64	53.66	47.34
Ukrainian	41.83	34.17	21.45	45.72	52.44
Arabic	43.85	31.31	32.03	37.48	42.83
German	71.75	58.83	58.81	60.53	61.67



(a) Length preservation: mean |len ratio − 1|

(b) Acronym preservation: F1

Figure 7: Track B (reference-based): system-level operational probes, macro-averaged over target languages. (a) Mean |len ratio − 1| across paragraphs (lower indicates closer length match to the reference; potential omission/verbosity deviations). (b) Acronym preservation (F1; higher is better).

Figure 8: Track B (reference-based): operational probe averages by language. Probes were computed over the five aligned paragraphs for each language–system pair and then macro-averaged across the nine systems. Len. ratio is the system/reference ratio of normalized character length; Len. abs. dev. is |1−Len. ratio| (lower is better). Acr. F1 is exact-match set-based F1 over extracted uppercase Latin acronyms. Punct. pres. is a bounded punctuation-count preservation score. Numeral preservation F1 was 1.00 for all languages/systems on this dataset and is omitted.

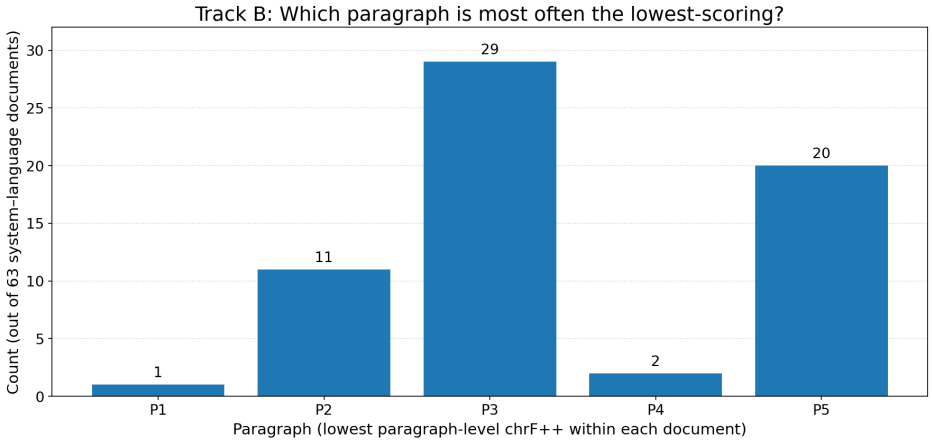


Figure 9: Track B (reference-based): frequency with which each paragraph is the lowest-scoring segment (by paragraph-level chrF++) within a system–language document (63 documents = 9 systems × 7 languages; 5 paragraphs each). Concentration on specific paragraphs indicates localized difficulty and tail risk.

Table 18: Track B (reference-based): worst localized failure per language, defined as the minimum paragraph-level chrF++ observed across all systems and paragraphs. We report the paragraph index, responsible system, and complementary diagnostics (TER, COMET, length ratio, punctuation preservation) for the same paragraph to distinguish overlap-driven divergence from omission- or formatting-like issues.

Language	P#	System	chrF++↓	TER↑	COMET↑	Len. ratio	Punct. preservation
Arabic	3	`llama-3-3-70b-instruct`	20.43	91.89	64.19	0.68	0.00
English	3	`meta-llama-3-1-8b-instruct`	33.03	62.00	69.54	0.57	0.83
German	5	`google-translate`	41.19	85.29	85.12	1.15	0.00
Polish	5	`meta-llama-3-1-8b-instruct`	28.31	83.87	64.82	0.69	0.36
Spanish	2	`google-translate`	50.01	56.67	78.69	0.87	0.59
Turkish	1	`meta-llama-3-1-8b-instruct`	23.81	91.67	76.48	0.88	0.44
Ukrainian	3	`azure-translator`	18.26	83.67	60.60	0.46	0.25

Table 19: Cross-track triangulation for GPT-4 vs. Azure Translator. Track A deltas are native-speaker rubric differences (GPT-4 minus Azure), aggregated over the six excerpts per language. Track B deltas are corpus-level metric point estimates on the controlled five-paragraph document: ΔchrF++ and ΔCOMET are (GPT-4 minus Azure), and ΔTER is (Azure minus GPT-4) so that positive values consistently favor GPT-4. "Notes" flags languages where a reference-based metric opposes the Track A direction.

Language	ΔCorrectness	ΔRegister	ΔchrF++	ΔCOMET	ΔTER	Notes
Turkish	2.7	3.1	-2.23	4.06	1.28	opposes: chrF++
Spanish	2.0	5.8	5.78	4.61	6.01	–
English	0.7	0.0	0.13	-1.26	1.22	opposes: COMET
Polish	1.3	2.3	6.65	3.54	-0.53	opposes: TER
Ukrainian	1.6	0.8	-1.07	2.19	-3.14	opposes: chrF++, TER
Arabic	2.1	1.6	1.95	3.81	2.48	–
German	2.9	2.2	5.25	2.73	10.10	–

Table 20: Indicative association between Track A rubric scores and Track B reference-based metrics across shared system–language points ($N = 14$: 7 languages × 2 systems). We report Spearman ρ between native-speaker rubric means and corpus-level metrics; for TER we use −TER so that higher is better. Because the tracks evaluate different source texts, these correlations should be interpreted as exploratory rather than as metric validation.

Reference-based metric	ρ(Correctness)	ρ(Register/style)	ρ(Readability)
BLEU	-0.05	0.08	-0.04
chrF++	0.05	0.13	0.10
−TER	0.05	0.22	0.07
BERTScore (F1)	0.07	0.23	0.12
COMET	0.36	0.36	0.50

tests whether reference-based metrics tend to agree with human rubric preferences on the *direction* of the difference.

4.3.1 Direction agreement and informative mismatches

Table 19 summarizes this bridge. In Track A, GPT-4 receives higher native-speaker scores than Azure in every language. The largest register/style gains appear in Spanish and Turkish, while English shows the smallest rubric differences.

The reference-based metrics mostly point in the same direction, but not always. GPT-4 scores higher than Azure on chrF++ in five of seven languages and on COMET in six of seven languages. The mismatches are useful for interpretation: in Turkish and Ukrainian, chrF++ favours Azure even though native-speaker correctness scores favour GPT-4, which is consistent with overlap metrics being sensitive to wording differences or alternative terminology. In English, COMET favours Azure while the human rubric slightly favours GPT-4. TER also favours Azure in Polish and Ukrainian. These cases reinforce that the reference-based track is best read as supporting evidence, not as a replacement for the native-speaker judgments.

4.3.2 Which reference-based metrics best track rubric preferences

Table 20 reports exploratory Spearman correlations between native-speaker rubric means and corpus-level reference-based metrics across shared system–language points ($N = 14$). COMET shows the strongest monotonic association among the metrics considered, especially for Readability ($\rho = 0.50$), but its associations with Correctness and Register/style are more modest (both $\rho = 0.36$). Overlap- and edit-based metrics correlate weakly with the rubric dimensions in this small setting. These results are indicative (tracks use different texts), but they support the broader observation that learned semantic metrics can align better than overlap metrics when paraphrase and stylistic variation are common.

4.4 Summary answers to the research questions

The results provide the following answers to the four research questions introduced in Section 1.

RQ1: Native-speaker comparison of GPT-4 and Azure Translator. Under the pension-domain rubric, native speakers rated GPT-4 higher than Azure Translator across all seven target languages and across all three rubric dimensions: meaning correctness, register/style, and readability. The largest and most consistent gains were observed for register and style, suggesting that GPT-4 was more successful at producing translations that sounded appropriate for client-facing pension communication. The effect was smallest for English, where both systems performed relatively strongly.

RQ2: Reference-based rankings across languages, metrics, and paragraphs. In the controlled reference-based setting, system rankings depended on both target language and metric family. The strongest systems formed a close top cluster, with Gemini 3 Pro, GPT-5.2, and DeepL often ranking highly, while Arabic and Ukrainian appeared more difficult on average than German and Spanish. Paragraph-level diagnostics showed that failures were not evenly distributed across documents: localized under-translation, formatting drift, and weak preservation of structured elements could affect specific paragraphs even when document-level scores looked acceptable.

RQ3: Alignment between reference-based metrics and human acceptability judgments. The shared GPT-4 versus Azure comparison showed partial alignment between the two tracks. Native speakers preferred GPT-4 in all languages, and reference-based semantic metrics often pointed in the same direction, especially COMET. However, overlap-based metrics such as chrF++ and BLEU sometimes diverged from human preferences, especially where acceptable wording differed from the reference.

RQ4: LLM-as-a-judge as a proxy for human evaluation. The GPT-4 judge provided a useful coarse proxy signal, but not a replacement for native-speaker evaluation. Its scores showed moderate agreement with human ratings and generally preserved the direction of the GPT-4 versus Azure contrast, but the judge also compressed score ranges and tended to shift scores upward. This makes LLM-as-a-judge more suitable for broad monitoring or regression detection than for final acceptability decisions in regulated pension communication.

5 Discussion

We structure the discussion around the main practical questions raised by the results, focusing on what the two-track evidence implies for deployment decisions in regulated pension communication. We discuss (i) how human rubric preferences translate into client-facing acceptability (meaning preservation, stable institutional voice, and readable calls to action), (ii) which recurring error patterns drive operational risk (e.g., condition omissions, terminology drift, entity corruption, and pragmatic misalignment in sensitive phrasing), and (iii) when reference-based metrics provide supportive evidence versus when they diverge from human judgments due to paraphrase and institutional phrasing. We close by summarizing language-specific challenges, the implications of model availability (open-weight vs. closed systems), the role of LLM-as-a-judge as a monitoring signal, and the governance measures that most directly reduce risk in regulated multilingual communication.

5.1 Client-facing acceptability: what human preferences mean in practice

In our pension-domain client-communication setting, GPT-4 received higher human rubric scores than Azure Translator across all languages and rubric categories. This does not support a general claim that LLMs outperform NMT systems across domains or language pairs. Rather, within the conditions of this pilot study, it suggests that a strong closed LLM can produce translations that are perceived as more acceptable for client-facing pension communication than a widely used commercial NMT baseline.

Across languages, evaluators most often described the practical advantage of LLM-based translations (relative to NMT baselines) as improvements in tone of voice and overall "client-readiness". This is consistent with observations that LLM translations can exhibit more human-like stylistic characteristics than conventional NMT outputs [70]. The weight of tone versus strict literal adequacy is context-dependent, but for client-facing pension communication it can be as consequential as meaning preservation.

5.2 Operational error profiles: register, terminology, and pragmatic risk

Across languages, human evaluators most frequently reduced scores due to three recurring classes of issues, consistent with fine-grained translation-quality taxonomies (e.g., MQM dimensions such as Style/Register, Terminology, and Locale convention) [56]: (1) register and address instability (intra-document shifts in formality conventions, e.g., DE *Du/Sie*, ES *usted/tú*, PL informal *ty*, and analogous politeness conventions in Ukrainian) [41, 71, 72], (2) terminology and entity governance failures (drift in domain terms and inconsistent handling of protected entities such as `UWV`) [40, 72, 73], and (3) pragmatic misalignment in sensitive contexts (mortality phrasing that is too direct where euphemistic register is expected) [71, 72].

Crucially, the systems differ not only in aggregate quality but also in *error profile*. Azure (NMT) is penalized predominantly for consistent failures, including literal or choppy realizations, register switching, and (for Arabic) structural/formatting problems, affecting register and readability across large portions of a document. GPT-4 (LLM) is more often described as letter-like and usable, but is penalized for more localized, undesirable choices: non-standard or misleading term selections (e.g., *arbeidsongeschikt* rendered as "disabled"), framing shifts (e.g., "survivor's pension" vs. partner pension), culturally infelicitous genre conventions (e.g., salutations in Turkish), occasional over-literal handling of structured strings (e.g., URL fragments), and insufficient pragmatic softening around death-related statements.

5.3 Metric–human divergence under paraphrase and institutional phrasing

Prior comparative studies that rely primarily on reference-based metrics (e.g., BLEU, chrF++, TER) frequently find that strong NMT systems remain superior for some language pairs and settings [44–46, 74, 75]. This contrasts with our human-evaluation results and suggests a domain-dependent mismatch between what overlap-style metrics reward and what matters for client-facing communication (tone, register stability, institutional phrasing, and pragmatic acceptability).

In our experiments, NMT systems such as DeepL and Google Translate remained highly competitive on reference-based metrics and outperformed several LLM systems on COMET [51, 76]. This is consistent with metric-based comparisons reporting strong performance of established NMT services relative to GPT-4 in some translation settings [74]. At the same time, the strongest closed LLMs in our setup achieved the best or near-best scores on most metrics, including COMET, suggesting that the gap between NMT and LLM-based translation is increasingly system- and setting-dependent rather than paradigm-fixed. This pattern is in line with shared-task evidence showing that top-tier proprietary LLM systems can lead MT-related rankings under controlled protocols [77, 78], while also reinforcing concerns that LLM-era MT may change how existing metrics behave [79, 80]. Overall, our results emphasize that reference-based metric superiority and human preference can diverge, especially when governance-critical constraints such as register stability, terminology consistency, and institutional phrasing dominate perceived quality.

5.4 Language-specific challenges

Arabic and Ukrainian were the hardest languages in our set. Both also use non-Latin scripts (Arabic script and Cyrillic), which may interact with formatting, punctuation conventions, and mixed-script template elements common in client communication (e.g., IDs, URLs, and numeric spans), amplifying structural and consistency errors.

5.5 Model availability: open-weight performance and scaling trends

Open-weight LLMs performed worse overall than the strongest closed systems in our set. Within an open-weight family, we observe improvements with increased capacity, while some newer releases match older

larger models, suggesting non-size-related gains (training data, optimization, and architectural refinements). Similar gaps between closed and open systems and strong size effects among open baselines are also reported in recent multilingual shared-task evaluations [77].

5.6 LLM-as-a-judge: useful proxy, limited calibration

We find a positive association between LLM-as-a-judge scores and human judgments, suggesting this approach can be a practical proxy for large-scale assessment. Prior work similarly finds that strong LLMs can be competitive evaluators at the system level [61]. At the same time, LLM judges often exhibit calibration artifacts (overscoring and compressed score ranges) and systematic biases [62, 77]. These effects matter most in high-quality regimes, where differences between systems are subtle and decision thresholds are tight.

5.7 Deployment governance: controlling decisive failure modes

A key implication is that many decisive failures in this setting are better characterized as *governance gaps* than as irreducible model limitations. The highest-leverage mitigations are therefore procedural and constraint-based: enforce a single register strategy per language (via templates and automatic checks, and where applicable explicit style/politeness control signals) [41, 71]; maintain a compact termbase and do-not-translate list for protected entities (with within-document consistency constraints and, where feasible, lexically constrained decoding) [40, 73]; codify approved phrasing for sensitive-event content; and treat structured/template fields (URLs, IDs, `PROEFDRUK`, and RTL number spans) as protected with post-generation validation. More broadly, these observations emphasize that progress in enterprise MT for regulated domains is not only about improving model capability, but also about integrating models into controlled translation pipelines that explicitly encode institutional policy, domain semantics, context, and template constraints [72].

6 Limitations

This study should be interpreted as an applied pilot rather than as a definitive benchmark of machine translation systems or paradigms. Its main limitations concern the size and composition of the evaluation materials, the human evaluation setup, the controllability of generation settings, the use of single-reference and paragraph-level automatic evaluation, and the exploratory nature of the LLM-as-a-judge analysis.

First, the empirical scope is limited. Track A uses six Dutch pension-domain excerpts, and Track B uses one short five-paragraph Dutch client-facing pension letter, with one professional reference translation for each target language. Although the materials were selected to reflect realistic pension communication and recurring communicative functions, they do not cover the full diversity of pension-domain genres, templates, legal formulations, or user journeys. As a result, the findings should not be generalized to all pension communication, to all regulated domains, or to all Dutch-to-target-language translation settings. The controlled reference-based results in Track B are especially sensitive to the selected document and should be read as complementary diagnostic evidence rather than as a standalone ranking of systems.

Second, the target-language evaluation is based on small and uneven native-speaker panels. Track A prioritized obtaining at least one qualified evaluator per language over enforcing a balanced evaluator count across languages. This design was appropriate for a pilot study with domain-familiar evaluators, but it limits the reliability of language-level estimates and prevents strong claims about inter-annotator agreement. In addition, no formal calibration session was conducted. Although all evaluators received the same workbook, rubric, and written instructions, differences in individual strictness, interpretation of score anchors, and sensitivity to register or terminology may have influenced the ratings.

Third, the Track A presentation format may have affected human judgments. Evaluators saw the Dutch source text together with two anonymized candidate translations in the same worksheet view. This side-by-side setup made the task manageable and supported direct comparison with the source, but it may also have allowed evaluators to use one translation to interpret or recover from problems in the other. Moreover, the presentation order was fixed rather than counterbalanced. Therefore, the resulting GPT-4 versus Azure differences should be interpreted within this specific evaluation setup, rather than as fully presentation-independent preference estimates.

Fourth, translation generation was not repeated. For each source text, target language, and system, we generated one output. This means that run-to-run variance was not estimated directly. The issue is particularly relevant for LLM-based systems, where residual stochasticity may remain even under low-temperature or deterministic-looking settings, and for systems accessed through web interfaces where decoding parameters were not exposed. Consequently, some observed differences may reflect a single sampled realization rather than stable system behavior across repeated generations.

Fifth, generation settings and access routes were not fully uniform across systems. API-accessed LLMs were configured as consistently as possible, but commercial web interfaces and NMT services did not expose the same controls. Gemini 3 Pro, DeepL, and Google Translate were accessed through public web interfaces, while Azure Translator was accessed through an API and most LLMs through Microsoft Foundry. These different access routes may introduce differences in hidden preprocessing, model versioning, decoding behavior, or interface-level updates. The reported system comparisons therefore reflect the systems as accessed in this study, not necessarily all possible configurations of those systems.

Sixth, the study does not include all relevant model classes or configurations. The system set was designed to provide a representative pilot sample of current NMT and LLM approaches, but it is not exhaustive. In particular, the experiments did not include reasoning-model configurations with non-zero reasoning effort, nor did they evaluate dedicated translation-oriented LLM variants beyond the systems listed in the study. The results therefore should not be interpreted as a complete comparison of all available LLM-based translation approaches.

Seventh, Track B relies on a single professional reference translation per language. Single-reference evaluation can penalize valid paraphrases, alternative terminology choices, or stylistic variants, especially when comparing LLM outputs with more literal NMT outputs. Although the study uses several metric families, including learned semantic metrics, all reference-based scores remain partly dependent on the wording and style of the single professional reference. This is particularly important in regulated client communication, where more than one translation may be acceptable if it preserves meaning, institutional voice, and readability.

Eighth, automatic metrics and operational probes provide only partial evidence of translation acceptability. BLEU, chrF++, TER, BERTScore, and COMET are useful for reproducible comparison, but they do not fully capture whether a translation is suitable for client-facing regulated communication. In Track B, these metrics were computed over paragraph-level units, not sentence-aligned units. Although the five paragraph boundaries were preserved across the Dutch source, professional references, and system outputs, the evaluation units were paragraph-level rather than sentence-level; in the Dutch source, these units ranged from 28 to 52 words and included multi-sentence administrative content. This matters because COMET is commonly used and interpreted at sentence level, and longer multi-sentence inputs may be less well calibrated for this use. We did not use a document-aware variant such as docCOMET in this pilot. Similarly, string-based metrics such as BLEU, chrF++, and TER were computed over paragraph units, so their overlap estimates may partly reflect sentence-boundary effects within paragraphs. The Track B scores should therefore be interpreted as paragraph-level diagnostic evidence rather than as sentence-aligned or fully document-aware estimates of translation quality. Similarly, the lightweight probes for length, acronyms, and punctuation are useful indicators of possible risk, but they are not substitutes for semantic error analysis or professional review. Some probe results may also be affected by language-specific writing systems and conventions, especially for non-Latin scripts.

Ninth, the cross-track comparison is triangulation rather than direct replication. Track A and Track B share the GPT-4 versus Azure contrast, but they use different source materials and different evaluation methods. Therefore, agreement between the two tracks strengthens the overall interpretation, but disagreement cannot be attributed cleanly to either the systems, the metrics, or the texts without further controlled experiments on shared materials.

Finally, the LLM-as-a-judge analysis is exploratory. The GPT-4 judge broadly tracked the direction of human preferences, but its scores were compressed and imperfectly calibrated relative to native-speaker ratings. In addition, because GPT-4 was also one of the translation systems under evaluation, the setup creates a potential self-evaluation conflict: even with system identity withheld, the judge may exhibit stylistic self-preference toward outputs resembling its own generation style. The judge should therefore be interpreted as a possible coarse monitoring signal, not as a replacement for human evaluation. This limitation is especially important in regulated pension communication, where small but meaning-critical errors can have

consequences that are not reliably captured by aggregate scores.

Taken together, these limitations mean that the study provides pilot evidence and a reproducible evaluation framework rather than final deployment recommendations. The findings are most useful for identifying evaluation design requirements, recurring risk patterns, and promising directions for controlled multilingual translation assessment in pension-domain communication.

7 Future research

This work was intentionally scoped as a two-track pilot. The current results already surface three themes that any follow-up should address: (i) strong cross-language heterogeneity, (ii) paragraph-level tail risk concentrated in a subset of document parts, and (iii) metric sensitivity and uncertainty in single-reference, small-sample evaluation. The next phase should therefore evolve the pilot into a deployment-grade evaluation and monitoring program. We outline concrete extensions below, grouped by evaluation expansion, controllability interventions, and operationalization.

7.1 Scale the controlled reference-based evaluation to stabilize conclusions

The controlled reference-based track provides an auditable pipeline with per-language rankings, uncertainty estimates, and paragraph-level diagnostics, but it currently rests on a single short document per language. Future research should expand scope along three axes:

Increase document coverage and template diversity. Evaluate multiple Dutch source documents per language that reflect common pension communication templates (e.g., informational webpages, action-required emails, benefit explanations, legal notices, authentication and account instructions). This enables (i) more stable per-language rankings, (ii) analysis by template type and terminology density, and (iii) stronger estimates of paragraph-level tail risk beyond one document.

Reduce single-reference sensitivity. Single-reference evaluation can over-penalize acceptable paraphrase and style variation, which is especially relevant when comparing LLMs to more literal NMT outputs. Two feasible extensions are: (i) multi-reference evaluation with two or more professional references per language for a subset of documents, and/or (ii) post-edit oriented evaluation, where professional translators measure acceptance with minimal edits (e.g., edit rate or MQM-style severity). These designs improve alignment between reference-based scores and enterprise acceptability.

Pre-registered comparisons and broader uncertainty analysis. With more data, bootstrap testing can be extended beyond top-versus-runner-up to a pre-registered set of comparisons (e.g., Azure vs. the strongest LLM per language; DeepL vs. top LLMs; prompt variants within an LLM), including correction for multiple comparisons. This reduces the risk of over-interpreting small, unstable gaps and better matches decision-making needs in procurement and governance.

7.2 Strengthen the reference-free rubric track for reliability and auditability

The rubric track directly measures stakeholder-relevant criteria (meaning, register/style, readability, terminology consistency), but reliability can be improved without abandoning feasibility:

Larger, better-calibrated native-speaker panels. Increase the number of evaluators per language and introduce lightweight calibration (shared anchors, short training, periodic consistency checks). A practical hybrid approach is to use a small expert panel to define anchors and then scale through a broader native-speaker pool.

Risk-weighted error annotation. Augment the rubric with a compact, MQM-inspired error taxonomy tailored to pension-domain risk (e.g., incorrect conditions, dates, eligibility criteria, identity/benefit entities, required actions). For raw or minimally reviewed MT outputs, these categories should be weighted not only by linguistic severity, but also by their likely consequences for specific recipients and use contexts [15]. In pension communication, this means distinguishing errors that mainly affect fluency or style from errors that may change a recipient's understanding of rights, obligations, eligibility, deadlines, or required actions. Such annotation would make the rubric more auditable and would support risk treatment decisions, such as which templates require human review, terminology controls, or post-generation validation. Recording severity categories alongside scores would preserve interpretability while enabling more systematic aggregation than free-text comments alone.

Template-specific acceptance checklists. Some message types have non-negotiable constraints (e.g., required legal phrasing, authentication instructions, privacy disclaimers). Adding template-specific checklists alongside the rubric would better reflect production acceptance criteria and allow clearer go/no-go decisions for high-impact templates.

7.3 Evaluate controllability interventions that target observed failure modes

Results across both tracks emphasize governance-related failures (terminology drift, register inconsistency, formatting deviations) and paragraph-level tail risk. Future research should therefore test interventions that improve *controllability* rather than only switching models:

Protected terms, glossaries, and terminology governance. Introduce (i) a protected-term list for non-translatable entities (fund names, service names, acronyms) and (ii) preferred translations for core pension concepts. Implement and compare glossary integration for NMT systems (where supported) and structured glossary conditioning for LLMs (prompt or constrained decoding). Evaluate gains in both tracks to quantify how much improvement is attributable to governance controls versus model choice.

Formatting and structure constraints. Many high-impact failures are structural (missing paragraphs, altered list formatting, punctuation drift). Evaluate simple structure constraints (preserve paragraph count, preserve bullets and numbering, preserve punctuation categories) using existing probes and targeted human checks. This directly targets the paragraph-level tail risk highlighted in the controlled track.

Prompt and configuration ablations for LLM translation. LLM translation behavior can change substantially with prompt design. A systematic ablation study (baseline prompt vs. register-controlled vs. glossary-controlled vs. structure-controlled) would clarify which advantages are due to model capability versus prompting, and would yield a reproducible enterprise translation recipe.

7.4 Operationalize monitoring for reference-free production streams

Most enterprise translation occurs without trusted references, so monitoring must rely on reference-free signals and escalation policies:

Probe-based triage without references. Redefine the probes used here to operate without references (e.g., length ratio relative to source, paragraph-count preservation, number consistency between source and target, punctuation structure preservation, protected-entity retention). Combine these with reference-free quality estimation (QE) models and/or rubric-aligned LLM judging with guardrails to route high-risk segments to human review.

Drift detection and regression monitoring. Providers, models, and prompts change over time. Future research should implement ongoing monitoring that detects drift in output characteristics (register, terminology usage, structural fidelity) and flags regressions after updates. Evaluation should become a recurring snapshot with versioned system configurations and reproducible artifacts.

Explainable error localization for auditability. Learned metrics provide useful semantic signals but remain hard to audit. Investigate explainable evaluators or span-level error localization methods that can highlight likely error regions and categories. Where such methods are too expensive at scale, run them selectively on outliers, high-impact templates, or disputed cases where overlap and semantic metrics disagree.

7.5 Summary

The next phase should move from a benchmarking pilot to a deployment-grade evaluation and monitoring program. Concretely, this means expanding controlled evaluation across documents and templates, strengthening rubric reliability, testing terminology and structure controls that target observed failure modes, and operationalizing reference-free monitoring with drift detection and auditability. These steps would enable system selection and maintenance based not only on average quality, but also on controllability, robustness, and risk reduction for pension-domain client communication.

8 Conclusion

This paper examined the quality of machine translation for Dutch pension-domain client communication under enterprise constraints using a two-track evaluation design. Track A assessed *enterprise acceptability* directly through native-speaker ratings of meaning correctness and completeness, register and style, and readability and flow. Track B complemented this with a controlled, reference-based benchmark on one fixed five-paragraph document translated into each target language, using overlap- and edit-based metrics (BLEU, chrF++, TER), learned semantic metrics (BERTScore, COMET), bootstrap uncertainty, and paragraph-level probes aimed at omission- and formatting-related risk.

Four main conclusions follow from this pilot. First, in the reference-free human evaluation, native speakers consistently preferred GPT-4 over Azure Translator across all seven target languages, with the largest gains in register and style. The qualitative feedback indicates that register stability and consistent handling of protected entities and terminology are decisive for perceived quality in pension communication. Second, the controlled reference-based benchmark showed that system rankings depend on both language and metric family. The strongest systems formed a relatively tight top cluster, and several small score gaps were not robustly separable given the limited size of the controlled test set. Third, the paragraph-level diagnostics showed that risk was often concentrated in specific parts of a document rather than distributed evenly across it. In particular, Arabic and Ukrainian exhibited more signs of under-translation and formatting drift, illustrating why document-level averages alone are insufficient for regulated communication. Fourth, GPT-4 used as an *LLM-as-a-judge* broadly tracked the direction of human preferences, but its scores were compressed and imperfectly calibrated. This supports its use as a coarse monitoring signal, not as a substitute for human evaluation.

A central contribution of this study is the linkage between the two tracks through the shared GPT-4 versus Azure comparison. Because the tracks used different source texts, this comparison does not provide a direct like-for-like replication across methods. Instead, it shows whether two different evaluation setups point to the same overall conclusion. In that respect, the pattern was informative: human judgments consistently favored GPT-4, while semantic reference-based metrics often pointed in the same direction, even when overlap-based metrics diverged. This suggests that learned semantic metrics may better reflect human acceptability in settings where paraphrase, institutional phrasing, and stylistic variation matter.

Taken together, the findings support a practical evaluation workflow for regulated multilingual communication. Controlled reference-based audits are useful for shortlisting systems, quantifying uncertainty, and surfacing localized failures. Human rubric-based assessment remains necessary for deciding whether translations are acceptable for client-facing use, especially where institutional voice, readability, and terminology governance are critical. For production settings in which references are unavailable, carefully constrained LLM-based judging may provide useful coarse monitoring signals, provided it is used with explicit guardrails and escalation to human review.

Beyond system selection, the strongest practical implication is that many decisive failures observed in this study are better treated as governance gaps than as model limitations alone. Register shifts, terminology drift, entity corruption, sensitive-event phrasing, and structured-field errors are not problems that can be solved reliably by average quality improvements alone. For practitioners, the actionable takeaway is that

enterprise MT in regulated domains should be deployed as a controlled translation pipeline: language-specific register policies, compact termbases and do-not-translate lists, approved phrasing for sensitive content, protected handling of template fields, and post-generation validation should be treated as core quality controls rather than optional refinements.

This study is necessarily limited in scope. The controlled benchmark uses one short document per language, Track A relies on small evaluator panels, and the cross-track analysis compares different source texts. The results should therefore be read as applied pilot evidence rather than as definitive rankings of MT paradigms. Even so, the study provides a reproducible framework for evaluating translation quality in pension-domain communication and shows how human acceptability, controlled automatic evaluation, and localized risk diagnostics can be combined to support safer multilingual deployment.

Data and code availability

The code and publicly shareable materials used in this study are available in the archived repository release corresponding to the version used in this paper [29]. This release includes the analysis code, prompts, and other reproducibility materials associated with the paper. The live development repository is hosted on GitHub at `https://github.com/nickradunovic/machine-translation-project`.

References

1. Debets S, Prast H, Rossi M, and soest A van. Pension Communication in the Netherlands and Other Countries. SSRN Electronic Journal 2018.
2. Haupt M. The Effects of Pension Communication on Knowledge, Attitudes, and Behaviour: An Integrative Review of Evidence and Directions for Future Research. Journal of Population Ageing 2023;17:1–44.
3. Debets S, Prast H, Rossi M, and Soest A van. Pension communication, knowledge, and behaviour. Journal of Pension Economics and Finance 2022;21. Published online 8 September 2020:99–118.
4. Lentz L, Nell L, and Pander Maat H. Begrijpelijkheid van pensioencommunicatie: effecten van wetgeving, geletterdheid en revisies. Tijdschrift voor Taalbeheersing 2017;39:191–208.
5. Sobczyk A. De Nederlandse taalvaardigheid onder arbeidsmigranten: De resultaten van het 6e arbeidsmigrantenpanel. Ed. by Cremers J. Vol. 6. Het Kenniscentrum Arbeidsmigranten, 2023.
6. Nunziatini M. Machine Translation in the Financial Services Industry: A Case Study. In: *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*. Dublin, Ireland: European Association for Machine Translation, 2019:57–63. URL: https://aclanthology.org/W19-6709/.
7. Läubli S, Amrhein C, Düggelin P, Gonzalez B, Zwahlen A, and Volk M. Post-editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain. In: *Proceedings of Machine Translation Summit XVII: Research Track*. Dublin, Ireland: European Association for Machine Translation, 2019:267–72. URL: https://aclanthology.org/W19-6626/.
8. Prieto Ramos F. Embracing New Translation Technologies at International Organizations: Tools, Risks and Competences in Institutional Translation with Neural Machine Translation. Journal of Language and Law 2025;14:1–29.
9. Bahdanau D, Cho K, and Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. In: *International Conference on Learning Representations (ICLR)*. 2015. URL: https://arxiv.org/abs/1409.0473.
10. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017. URL: https://arxiv.org/abs/1706.03762.
11. Tan Z, Wang S, Yang Z, et al. Neural Machine Translation: A Review of Methods, Resources, and Tools. AI Open 2020.
12. Zhang X, Rajabi N, Duh K, and Koehn P. Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. In: *Proceedings of the Eighth Conference on Machine Translation (WMT)*. Singapore: Association for Computational Linguistics, 2023:468–81. DOI: 10.18653/v1/2023.wmt-1.43. URL: https://aclanthology.org/2023.wmt-1.43/.
13. Zhu W, Liu H, Dong Q, et al. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. 2024:2765–81. DOI: 10.18653/v1/2024.findings-naacl.176. URL: https://aclanthology.org/2024.findings-naacl.176/.
14. Manakhimova S, Avramidis E, Macketanz V, Lapshinova-Koltunski E, Bagdasarov S, and Möller S. Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can GPT-4 Outperform NMT? In: *Proceedings of the Eighth Conference on Machine Translation*. 2023:224–45. URL: https://aclanthology.org/2023.wmt-1.23/.
15. Koponen M and Nurminen M. Risk Management for Content Delivery via Raw Machine Translation. In: *Translation, Interpreting and Technological Change: Innovations in Research, Practice and Training*. Ed. by Winters M, Deane-Cox S, and Böser U. London: Bloomsbury Academic, 2024:111–35. DOI: 10.5040/9781350212978.0013. URL: https://trepo.tuni.fi/handle/10024/209574.
16. Dinu G, Mathur P, Federico M, and Al-Onaizan Y. Training Neural Machine Translation to Apply Terminology Constraints. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019:3063–8. DOI: 10.18653/v1/P19-1294. URL: https://aclanthology.org/P19-1294/.

17. Hanneman G and Dinu G. How Should Markup Tags Be Translated? In: *Proceedings of the Fifth Conference on Machine Translation*. 2020. URL: https://aclanthology.org/2020.wmt-1.138.pdf.

18. Bogoychev N and Chen P. Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting. In: *Proceedings of the Eighth Conference on Machine Translation*. 2023:890–6. DOI: 10.18653/v1/2023.wmt-1.80. URL: https://aclanthology.org/2023.wmt-1.80/.

19. Dabre R, Song H, Exel M, Buschbeck B, Eschbach-Dymanus J, and Tanaka H. How Effective is Synthetic Data and Instruction Fine-tuning for Translation with Markup using LLMs? In: *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. 2024:73–87. URL: https://aclanthology.org/2024.amta-research.8/.

20. Dale D, Voita E, Lam J, et al. HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023. URL: https://aclanthology.org/2023.emnlp-main.42/.

21. Guerreiro NM, Alves DM, Waldendorf J, et al. Hallucinations in Large Multilingual Translation Models. Transactions of the Association for Computational Linguistics 2023;11:1500–17.

22. Koehn P and Knowles R. Six Challenges for Neural Machine Translation. In: *Proceedings of the First Workshop on Neural Machine Translation*. Ed. by Luong T, Birch A, Neubig G, and Finch A. Vancouver, Canada: Association for Computational Linguistics, 2017:28–39. DOI: 10.18653/v1/w17-3204. URL: https://aclanthology.org/W17-3204/.

23. Chu C and Wang R. A Survey of Domain Adaptation for Neural Machine Translation. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018:1304–19. URL: https://aclanthology.org/C18-1111/.

24. Specia L, Scarton C, and Paetzold GH. Quality Estimation for Machine Translation. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2018. DOI: 10.2200/s00854ed1v01y201805hlt039.

25. Zerva C, Blain F, C. De Souza JG, et al. 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*. Miami, Florida, USA: Association for Computational Linguistics, 2024:82–109. DOI: 10.18653/v1/2024.wmt-1.3. URL: https://aclanthology.org/2024.wmt-1.3/.

26. Freitag M, Foster G, Grangier D, Ratnakar V, Tan Q, and Macherey W. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. Transactions of the Association for Computational Linguistics 2021;9:1460–74.

27. OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 2023.

28. Microsoft. Azure Translator documentation. https://learn.microsoft.com/en-us/azure/ai-services/translator/. Microsoft Learn documentation. Accessed: 2026-04-12. 2026.

29. Radunovic N. Reproducibility materials for: A Two-Track Evaluation of LLM and NMT Translation for Regulated Pension Communication. Version 1.0.0. 2026. DOI: 10.5281/zenodo.19821399. URL: https://github.com/nickradunovic/machine-translation-project.

30. Ekaterina K, Klara H, and Agneta R. How can pension communication be easier for individuals? A summary of the literature. International Journal of Pensions Management 2024;22:107–25.

31. Hoekstra M, Daven E, and Maarten A. De Nederlandse pensioensector onder de loep. TPEdigitaal 2024;10:15–30.

32. Li A, Zhou W, Hoda R, Bain C, and Poon P. Comparing Large Language Models and Traditional Machine Translation Tools for Translating Medical Consultation Summaries: A Pilot Study. 2025. DOI: 10.48550/arXiv.2504.16601. arXiv: 2504.16601 [cs.CL]. URL: https://arxiv.org/abs/2504.16601.

33. Haddow B, Bawden R, Miceli Barone AV, Helcl J, and Birch A. Survey of Low-Resource Machine Translation. Computational Linguistics 2022;48:673–732.

34. Ataman D, Birch A, Habash N, Federico M, Koehn P, and Cho K. Machine Translation in the Era of Large Language Models:A Survey of Historical and Emerging Problems. Information 2025;16:723.

35. Liao B, Herold C, Khadivi S, and Monz C. IKUN for WMT24 General MT Task: LLMs Are here for Multilingual Machine Translation. 2024. DOI: 10.48550/arxiv.2408.11512. arXiv: 2408.11512. URL: https://arxiv.org/abs/2408.11512.

36. Jiao W, Wang W, Huang Jt, Wang X, Shi S, and Tu Z. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. 2023. DOI: 10.48550/arxiv.2301.08745. arXiv: 2301.08745 [cs.CL]. URL: https://doi.org/10.48550/arXiv.2301.08745.

37. Yan J, Yan P, Chen Y, Li J, Zhu X, and Zhang Y. Benchmarking GPT-4 against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels. 2024. DOI: 10.48550/arxiv.2411.13775. arXiv: 2411.13775 [cs.CL]. URL: https://arxiv.org/abs/2411.13775.

38. Finkelstein M, Caswell I, Domhan T, et al. TranslateGemma Technical Report. 2026. arXiv: 2601.09012 [cs.CL]. URL: https://arxiv.org/abs/2601.09012.

39. Hokamp C and Liu Q. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Barzilay R and Kan MY. Vancouver, Canada: Association for Computational Linguistics, 2017:1535–46. DOI: 10.18653/v1/p17-1141. URL: https://aclanthology.org/P17-1141/.

40. Post M and Vilar D. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Walker M, Ji H, and Stent A. New Orleans, Louisiana: Association for Computational Linguistics, 2018:1314–24. DOI: 10.18653/v1/n18-1119. URL: https://aclanthology.org/N18-1119/.

41. Sennrich R, Haddow B, and Birch A. Controlling Politeness in Neural Machine Translation via Side Constraints. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Knight K, Nenkova A, and Rambow O. San Diego, California: Association for Computational Linguistics, 2016:35–40. DOI: 10.18653/v1/n16-1005. URL: https://aclanthology.org/N16-1005/.

42. Moslem Y, Romani G, Molaei M, Kelleher JD, Haque R, and Way A. Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Koehn P, Haddow B, Kocmi T, and Monz C. Singapore: Association for Computational Linguistics, 2023:902–11. DOI: 10.18653/v1/2023.wmt-1.82. URL: https://aclanthology.org/2023.wmt-1.82/.

43. Yang M and Li F. Improving Machine Translation Formality with Large Language Models. Computers, Materials & Continua 2025;82:2061–75.

44. Papineni K, Roukos S, Ward T, and Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Isabelle P, Charniak E, and Lin D. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002:311–8. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040/.

45. Popović M. chrF: character n-gram F-score for automatic MT evaluation. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Bojar O, Chatterjee R, Federmann C, et al. Lisbon, Portugal: Association for Computational Linguistics, 2015:392–5. DOI: 10.18653/v1/w15-3049. URL: https://aclanthology.org/W15-3049/.

46. Snover M, Dorr B, Schwartz R, Micciulla L, and Makhoul J. A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, 2006:223–31. URL: https://aclanthology.org/2006.amta-papers.25/.

47. Callison-Burch C, Osborne M, and Koehn P. Re-evaluating the Role of BLEU in Machine Translation Research. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics, 2006:249–56. URL: https://aclanthology.org/E06-1032/.

48. Chatzikoumi E. How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering 2020;26:137–61.

49. Zhang T, Kishore V, Wu F, Weinberger KQ, and Artzi Y. BERTScore: Evaluating Text Generation with BERT. 2020. arXiv: 1904.09675 [cs.CL]. URL: https://openreview.net/forum?id=SkeHuCVFDr.

50. Sellam T, Das D, and Parikh AP. BLEURT: Learning Robust Metrics for Text Generation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020:7881–92. DOI: 10.18653/v1/2020.acl-main.704. URL: https://aclanthology.org/2020.acl-main.704/.

51. Rei R, Stewart C, Farinha AC, Lavie A, and Martins AFT. COMET: A Neural Framework for MT Evaluation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Webber B, Cohn T, He Y, and Liu Y. Online: Association for Computational Linguistics, 2020:2685–702. DOI: 10.18653/v1/2020.emnlp-main.213. URL: https://aclanthology.org/2020.emnlp-main.213/.

52. Freitag M, Rei R, Mathur N, et al. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, 2022:46–68. DOI: 10.18653/v1/2022.wmt-1.2. URL: https://aclanthology.org/2022.wmt-1.2/.

53. Vilar D, Xu J, D'Haro LF, and Ney H. Error Analysis of Statistical Machine Translation Output. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA), 2006. URL: https://aclanthology.org/L06-1141/.

54. Koehn P. Statistical Significance Tests for Machine Translation Evaluation. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, 2004:388–95. URL: https://aclanthology.org/W04-3250/.

55. Graham Y, Baldwin T, and Mathur N. Accurate Evaluation of Segment-level Machine Translation Metrics. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, 2015:1183–91. DOI: 10.3115/v1/n15-1124. URL: https://aclanthology.org/N15-1124/.

56. Lommel AR, Burchardt A, and Uszkoreit H. Multidimensional quality metrics: a flexible system for assessing translation quality. In: *Proceedings of Translating and the Computer 35*. November 28–29. London, UK: Aslib, 2013. URL: https://aclanthology.org/2013.tc-1.6/.

57. Lommel A, Burchardt A, and Uszkoreit H. Multidimensional quality metrics (MQM): a framework for declaring and describing translation quality metrics. Tradumàtica: Tecnologies de la Traducció 2014:455–63.

58. Fomicheva M, Lertvittayakumjorn P, Zhao W, Eger S, and Gao Y. The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results. In: *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Ed. by Gao Y, Eger S, Zhao W, Lertvittayakumjorn P, and Fomicheva M. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021:165–78. DOI: 10.18653/v1/2021.eval4nlp-1.17. URL: https://aclanthology.org/2021.eval4nlp-1.17/.

59. Guerreiro NM, Rei R, Stigt Dv, Coheur L, Colombo P, and Martins AFT. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. Transactions of the Association for Computational Linguistics 2024;12:979–95.

60. Treviso MV, Guerreiro NM, Agrawal S, et al. xTower: A Multilingual LLM for Explaining and Correcting Translation Errors. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Al-Onaizan Y, Bansal M, and Chen YN. Miami, Florida, USA: Association for Computational Linguistics, 2024:15222–39. DOI: 10.18653/v1/2024.findings-emnlp.892. URL: https://aclanthology.org/2024.findings-emnlp.892/.

61. Kocmi T and Federmann C. 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*. Ed. by Nurminen M, Brenner J, Koponen M, et al. Tampere, Finland: European Association for Machine Translation, 2023:193–203. URL: https://aclanthology.org/2023.eamt-1.19/.

62. Liu Y, Iter D, Xu Y, Wang S, Xu R, and Zhu C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Bouamor H, Pino J, and Bali K. Singapore: Association for Computational Linguistics, 2023:2511–22. DOI: 10.18653/v1/2023.emnlp-main.153. URL: https://aclanthology.org/2023.emnlp-main.153/.

63. Qian S, Sindhujan A, Kabra M, et al. What do Large Language Models Need for Machine Translation Evaluation? In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, 2024:3660–74. DOI: 10.18653/v1/2024.emnlp-main.214. URL: https://aclanthology.org/2024.emnlp-main.214/.

64. Kim A. RUBRIC-MQM: Span-Level LLM-as-judge in Machine Translation For High-End Models. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*. Vienna, Austria: Association for Computational Linguistics, 2025:147–65. DOI: 10.18653/v1/2025.acl-industry.12. URL: https://aclanthology.org/2025.acl-industry.12/.

65. Isabelle P, Cherry C, and Foster G. A Challenge Set Approach to Evaluating Machine Translation. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Palmer M, Hwa R, and Riedel S. Copenhagen, Denmark: Association for Computational Linguistics, 2017:2486–96. DOI: 10.18653/v1/d17-1263. URL: https://aclanthology.org/D17-1263/.

66. Google Cloud. Neural Machine Translation (NMT). https://docs.cloud.google.com/translate/docs/advanced/nmt-model. Google Cloud documentation. Accessed: 2026-04-12. 2026.

67. Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144 2016.

68. DeepL Team. How does DeepL work? https://www.deepl.com/en/blog/how-does-deepl-work. DeepL blog post. Last updated: 2021-10-31. Accessed: 2026-04-12. 2021.

69. DeepL. About DeepL language models. https://support.deepl.com/hc/en-us/articles/14241705319580-About-DeepL-language-models. DeepL Help Center. Accessed: 2026-04-12. 2026.

70. Sizov F, España-Bonet C, Van Genabith J, Xie R, and Dutta Chowdhury K. Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations. In: *Proceedings of the Ninth Conference on Machine Translation*. Ed. by Haddow B, Kocmi T, Koehn P, and Monz C. Miami, Florida, USA: Association for Computational Linguistics, 2024:1183–99. DOI: 10.18653/v1/2024.wmt-1.116. URL: https://aclanthology.org/2024.wmt-1.116/.

71. Niu X, Rao S, and Carpuat M. Multi-Task Neural Models for Translating Between Styles Within and Across Languages. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018:1008–21. URL: https://aclanthology.org/C18-1086/.

72. Castilho S and Knowles R. A survey of context in neural machine translation and its evaluation. Natural Language Processing 2024. Published online 17 May 2024.

73. Wiśniewski D, Pokrywka M, and Rostek Z. Do Not Change Me: On Transferring Entities Without Modification in Neural Machine Translation - a Multilingual Perspective. In: *Proceedings of Machine Translation Summit XX: Volume 1*. Geneva, Switzerland: European Association for Machine Translation, 2025:248–64. URL: https://aclanthology.org/2025.mtsummit-1.19/.

74. Son J and Kim B. Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. Information 2023;14:574.

75. Chow RC, Angeline V, Jayata G, Mujhid A, and Hidayaturrahman. Comparing LLMs and NMTs Performances in Translating English Indonesian Texts. Procedia Computer Science 2025;269:1455–65.

76. Kocmi T, Federmann C, Grundkiewicz R, Junczys-Dowmunt M, Matsushita H, and Menezes A. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In: *Proceedings of the Sixth Conference on Machine Translation*. Ed. by Barrault L, Bojar O, Bougares F, et al. Online: Association for Computational Linguistics, 2021:478–94. URL: https://aclanthology.org/2021.wmt-1.57/.

77. Kocmi T, Agrawal S, Artemova E, et al. Findings of the WMT25 Multilingual Instruction Shared Task: Persistent Hurdles in Reasoning, Generation, and Evaluation. In: *Proceedings of the Tenth Conference on Machine Translation*. Ed. by Haddow B, Kocmi T, Koehn P, and Monz C. Suzhou, China: Association for Computational Linguistics, 2025:414–35. DOI: 10.18653/v1/2025.wmt-1.23. URL: https://aclanthology.org/2025.wmt-1.23/.

78. Kocmi T, Avramidis E, Bawden R, et al. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In: *Proceedings of the Ninth Conference on Machine Translation*. Ed. by Haddow B, Kocmi T, Koehn P, and Monz C. Miami, Florida, USA: Association for Computational Linguistics, 2024:1–46. DOI: 10.18653/v1/2024.wmt-1.1. URL: https://aclanthology.org/2024.wmt-1.1/.

79. Freitag M, Mathur N, Deutsch D, et al. Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In: *Proceedings of the Ninth Conference on Machine Translation*. Ed. by Haddow B, Kocmi T, Koehn P, and Monz C. Miami, Florida, USA: Association for Computational Linguistics, 2024:47–81. DOI: 10.18653/v1/2024.wmt-1.2. URL: https://aclanthology.org/2024.wmt-1.2/.

80. Mathur N, Baldwin T, and Cohn T. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Jurafsky D, Chai J, Schluter N, and Tetreault J. Online: Association for Computational Linguistics, 2020:4984–97. DOI: 10.18653/v1/2020.acl-main.448. URL: https://aclanthology.org/2020.acl-main.448/.

Appendices

A Dutch source document for the reference-based track

Dutch source text (5 paragraphs)

Wij hebben informatie gekregen van het Uitvoeringsinstituut Werknemersverzekeringen (UWV). De informatie van het UWV gaat over uw arbeidsongeschiktheid. In deze brief leest u wat dit voor u betekent.

U bouwt 'premievrije pensioen' bij ons op
Dat komt omdat u arbeidsongeschikt bent. En omdat u een WAO/WIA-uitkering van het UWV krijgt. Daarom betalen wij (een deel van) uw pensioenpremie. Dit noemen we 'premievrije deelname (PVD)'. De hoogte van de premie die wij betalen hangt af van uw 'uitkeringspercentage'.

Uw PVD is veranderd
U heeft een WAO/WIA-uitkering. Het UWV heeft ons nieuwe informatie gegeven over uw WAO/WIA-uitkering. In het onderstaande overzicht leest u wat dat voor u betekent.

Wat betekent dit voor uw PVD?
In de kolom 'Premievrije pensioenopbouw' ziet u een percentage (%) staan. Dit is het deel waarover u PVD krijgt. Dit deel van de pensioenpremie betalen wij. Verandert uw 'uitkeringspercentage'? Dan verandert ook het percentage 'premievrije pensioenopbouw'. En dus ook het deel van de pensioenpremie dat wij betalen.

U moet veranderingen van uw arbeidsongeschiktheid zelf doorgeven
Krijgt u een nieuwe brief van het UWV? Dan moet u ons een kopie van die brief sturen. Het deel van de premie dat wij betalen verandert dan.

B LLM prompts

Translation prompt

```
Act as a diligent professional translator that translates text into {language}.
Preserve meaning completely, keep the original tone/register and style, and ensure fluent readability.

Translate the following text:
{doc}
```

Figure B.1: Translation prompt used for the LLM-based translations.

Grading prompt

```
Please act as an impartial judge.
Compare ORIGINAL with TRANSLATION, and evaluate the quality of the translation from source language (
ORIGINAL) to target language (TRANSLATION). This is the TRANSLATE_SCORE.

TRANSLATE_SCORE is divided into the following categories: ‘correctness and
completeness of the content’, ‘tone and style’, and ‘flow and readability’.

For ‘correctness and completeness of the content’, you grade from 1 to 10, where a higher grade
corresponds to the TRANSLATION having preserved the content of ORIGINAL more such that it includes the
same information without missing any parts. For example, if certain topics are not covered by the
translation, or if the translation talks about unrelated topics, then points are subtracted from ‘
correctness and completeness’.

For ‘tone and style’, you grade from 1 to 10, where a higher grade corresponds to the
TRANSLATION being more similar in tone and voice and style. For example, if the translation differs from
the original in terms of the sentiment, and tone of voice, then points are subtracted from ‘tone and
style’.

For ‘flow and readability’, you grade from 1 to 10, where a higher grade corresponds to the
TRANSLATION being more similar in the readability level of the text. For example, if the translation
differs from the original in terms of the extent of grammatical errors, then points are subtracted from ‘
flow and readability’.

All the <<...>> instances that appear in the text are placeholders; ignore them.

Start your evaluation by assigning TRANSLATE_SCORE on the previously disclosed scales for the categories
‘correctness and completeness of the content’, ‘tone and style’, and ‘flow
and readability’.

Follow the format of the following example:

-----------
‘correctness and completeness of the content’: 8
‘tone and style’: 5
‘flow and readability’: 7

Explanation: <give explanation>
-----------

Note, that each noticeable deviation in the translation can lead to subtraction of points in only one of
the categories at a time. Make sure that the grading of each of the categories is non-overlapping. So,
there should be a clear distinct grading for each of the categories based on deviations with the
translation.

Then, provide a brief explanation to support your scores for each of the areas.

ORIGINAL:
{original}

TRANSLATION:
{translation}
```

Figure B.2: Translation prompt used for the LLM-based translations.