

Large Language Models Are Competitive with Dedicated Machine Translation Systems for Regulated Communications: A Case Study in the Pension Domain

Nicholas Radunović¹[0009–0004–1441–5191], Alina
Karakanta²[0000–0002–1029–1337], Flor Miriam
Plaza-del-Arco¹[0000–0002–3020–5512], and Peter van der
Putten¹[0000–0002–6507–6896]

¹ LIACS, Leiden University, The Netherlands

`nick.radunovic@gmail.com,`

`{f.m.plaza.del.arco,p.w.h.van.der.putten}@liacs.leidenuniv.nl`

² Leiden Centre for Linguistics, Leiden University, The Netherlands

`a.karakanta@hum.leidenuniv.nl`

Abstract. Pension communication must clearly convey decisions and required actions, but language barriers can reduce comprehension and engagement. We evaluate off-the-shelf machine translation of Dutch pension client texts into seven languages in a two-track pilot study combining human and automated evaluation. Track A is reference-free: native speakers evaluate translations of six representative excerpts by GPT-4 and Azure Translator on meaning preservation, register, and readability, supplemented by an LLM-as-a-judge. Track B benchmarks nine systems (three commercial NMT engines and six LLMs) on a fixed five-paragraph document against professional references using overlap, edit, and semantic metrics, bootstrap uncertainty, and paragraph-level fidelity probes. Native speakers consistently prefer GPT-4 over Azure Translator across all languages, with the largest gains in register. Reference-based results show a tight top cluster, strong cross-language heterogeneity, and occasional disagreement between overlap and semantic metrics. Paragraph-level diagnostics highlight omission and formatting risks, especially for Arabic and Ukrainian. LLM-as-a-judge aligns with human preference direction but exhibits score compression. Overall, our findings support deployment workflows combining rubric-aligned acceptability checks, controlled audits with uncertainty, and explicit terminology and structure governance for regulated multilingual communication.

Keywords: Large Language Models · Neural Machine Translation · Translation Evaluation · LLM-as-a-Judge · Regulated Communications

1 Introduction

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 [10, 11, 21]. Research in Dutch pension communications also highlights that mandated disclosure and legal precision can conflict with comprehensibility, and that revisions can change information-finding and understanding outcomes [31]. To support informed decision-making, information and calls to action must be understandable and clear. For some populations, including migrant workers, language barriers may contribute to low engagement, underscoring the importance of clear, accessible multilingual communication [49].

Organizations increasingly adopt machine translation (MT) to deliver such communication at scale, including in financial services, the public sector and other regulated domains where quality requirements emphasize continuity of institutional discourse and terminology [30, 38, 43]. Over the past decade, neural MT (NMT) has become the dominant commercial approach, improving fluency and meaning preservation relative to earlier paradigms [2, 51, 53]. More recently, large language models (LLMs) have entered the MT landscape as general-purpose systems capable of translation via prompting and fine-tuning [35, 60, 61].

Regulated use imposes requirements beyond general-purpose translation. Translations must preserve meaning while maintaining institutional register, consistent domain terminology, and document structure [14, 20, 43]. LLMs are attractive because their behavior can be guided at inference time via prompts and structured constraints [3, 8, 60], yet LLMs may partly ignore instructions, and both LLM and NMT systems can omit or hallucinate content [9, 18, 28]. Performance also varies by language and domain, especially when in-domain parallel data and expert validation are limited [7, 28, 61].

A second challenge is evaluation. Professional reference translations are costly and may be unavailable for frequently changing content, which has prompted research into reference-free evaluation [50, 58]. When references are available, automatic metrics provide reproducible comparisons but may penalize valid paraphrases or miss problems with register and terminology [4].

We examine these issues in a two-track pilot study of Dutch pension communication translated into Turkish, Spanish, English, Polish, Ukrainian, Arabic, and German. Given budget constraints, we prioritized methodological breadth over the number of source documents, examining a modest set of texts through multiple complementary evaluation perspectives. This trade-off motivates the pilot design. Track A compares GPT-4 with Azure Translator on six pension-domain excerpts [36, 39]. Native speakers assess meaning, register, and readability without reference translations. Track A also evaluates GPT-4 as an LLM judge by comparing its scores on the anonymized translations with the native-speaker ratings. Track B compares nine LLM and NMT systems on one pension document using professional reference translations, several metric families, and uncertainty estimates. Both tracks include GPT-4 and Azure Translator, but they use different source texts. Their shared comparison therefore provides complementary evidence rather than a direct replication.

We address two research questions in this study:

1. How do native speakers rate GPT-4 and Azure Translator for meaning, register, and readability across seven target languages?
2. How do nine LLM and NMT systems rank across target languages and reference-based metric families?

To support reproducibility and reuse, the study resources, including the analysis code, prompts, scoring scripts, table and figure generation code, and publicly shareable text materials, are made available through an archived GitHub repository [45].

2 Background and Related Work

Pension communication is a regulated, action-oriented genre. Client letters, emails, and webpages combine specialized terminology, an institutional register, and instructions concerning benefits or required actions. Machine translation systems can omit or hallucinate content and mishandle terminology or protected document elements [9, 14, 18, 20]. Because recipients may rely on these texts when interpreting benefits and deciding what to do, translation quality in this setting includes semantic accuracy, readable wording, stable register, consistent terminology, and preservation of document structure [29, 43]. These requirements distinguish acceptability for client-facing use from general fluency.

Prior work has not established a consistent advantage for either NMT or LLMs across languages and deployment conditions. NMT performs well on many general-domain benchmarks, but domain shift can reduce quality when training data do not match the terminology and genre of deployment text [7, 28]. Adaptation can reduce this mismatch, but in-domain parallel or monolingual data may be scarce, costly, or sensitive in regulated organizations. Organizations therefore often begin with off-the-shelf commercial systems. Prompted LLMs offer an alternative. Studies report competitive translation quality from GPT-family models but also find prompt sensitivity, lexical inconsistency, omissions, and occasional meaning errors [1, 9, 18, 24, 56, 60]. Both approaches allow some control over output. NMT can use constrained decoding or control signals, while LLMs can receive terminology, register, and formatting instructions at inference time [22, 37, 42, 47, 57]. Control does not guarantee compliance, and LLMs may over-paraphrase or ignore instructions. Differences in training and evaluation resources across languages further complicate comparison because lower-resource language directions may behave differently under domain shift [19]. Their relative performance must therefore be evaluated in the target domain and languages.

Automatic scores alone cannot determine whether a translation is suitable for client-facing pension communication. When professional translations are available, reference-based metrics support reproducible system comparisons. Lexical metrics such as BLEU [40], chrF++ [41], and TER [48] are inexpensive but sensitive to reference wording [4, 6]. Learned metrics such as BERTScore [59] and COMET [46] correlate more closely with expert judgments but remain difficult to interpret [16]. When score differences are small, bootstrap estimates can show whether an apparent gap is stable within the test set [27]. However,

a high document-level score can still conceal a catastrophic error in one segment [23, 54]. Human evaluation assesses meaning, register, readability, and terminology more directly. Multidimensional Quality Metrics (MQM) offers a detailed error taxonomy, and expert evaluations show that protocol and evaluator expertise can change system rankings [15, 33]. Full MQM annotation is costly, so applied multilingual studies often rely on smaller native-speaker panels, rubrics, and qualitative comments [32]. When reference translations are unavailable, quality-estimation systems and prompted LLM evaluators can provide reference-free signals, but their reliability varies with the model, prompt, and calibration setting [25, 26, 44, 50]. Recent work in asylum reception centres combines automatic metrics with expert human evaluation to assess multilingual NMT for operational messages [34]. We therefore compare LLM and NMT systems using native-speaker judgments, reference-based metrics, and an exploratory LLM-as-a-judge assessment.

3 Methods

This section describes the two-track design and methods used to assess translation quality for Dutch pension-domain client communication. Track A (reference-free) compared GPT-4 with Azure Translator without professional reference translations. Native speakers assessed meaning, register, and readability. Track B (reference-based) compared nine LLM and NMT systems with professional translations of one controlled document. Reference-based metrics measured different aspects of similarity, and paragraph-level analyses identified localized failures. Because the tracks used different amounts and types of source material, we treat them as complementary rather than directly comparable experiments.

3.1 Materials, Languages, and Systems

Track A used six Dutch excerpts supplied by an external pension-sector stakeholder: four client letters and two informational webpages, selected mainly by length from pools of 20 and 10 respectively. The 200–500-word excerpts covered common informing and instructing functions, pension terminology, and institutional phrasing. The sample was not designed to represent the stakeholder’s full communication portfolio. Personal names were replaced with placeholders.

Track B used one of the six client letters. The 197-word document contains five paragraphs and explains a change in premium-free pension accrual following updated information about disability status and WAO/WIA benefits.³ The document was short enough to translate professionally into all seven languages while retaining institutional phrasing and pension terminology.

We commissioned one professional translation per target language from Translyte [52], preserving the five-paragraph structure used for alignment. Appendix 3 provides the Dutch source and the archived materials contain the references [45].

³ WAO and WIA are successive Dutch statutory disability-benefit schemes for employees.

Both tracks covered Turkish, Spanish, English, Polish, Ukrainian, Arabic, and German. The stakeholder selected these languages to reflect its multilingual communication needs in the Netherlands. The set combined widely used international languages with languages spoken by recipient groups that unpublished internal analyses had associated with lower response rates to Dutch pension messages containing calls to action.

Track A compared GPT-4 with Azure Translator. Track B included three commercial NMT baselines, Azure Translator, DeepL, and Google Translate, and six LLMs, Gemini 3 Pro, GPT-4, GPT-5.2, Llama 3.1 8B, Llama 3.1 405B, and Llama 3.3 70B. Public information places the three commercial services within the NMT tradition, although their production architectures are not fully disclosed [12, 13, 17, 36, 55]. The LLM set covers several providers, proprietary and open-weight systems, and different model sizes. Appendix Table 8 lists the model identifiers and available version information.

3.2 Translation Generation

GPT and Llama models were accessed through Microsoft Foundry, Azure Translator through its REST API, and Gemini 3 Pro, DeepL, and Google Translate through their web interfaces. Configurable decoding settings were aligned where possible; Table 9 in the Appendix reports the exposed settings.

We generated one output for every assigned source, language, and system combination. We did not estimate run-to-run variation, and systems without exposed decoding controls may retain residual stochasticity. No output received human post-editing. Normalization only removed Markdown markers and, when needed for Track B alignment, inserted empty lines between existing paragraph units. We did not change words, punctuation, or sentence content. GPT-4 translations in Track A used a deliberately simple, structured prompt targeting meaning, register, and readability (see repository for the prompt [45]).

3.3 Track A: Reference-Free Human Evaluation

Evaluators filled in a spreadsheet workbook remotely and independently. For each excerpt, the workbook displayed the Dutch source beside the two target-language outputs. The outputs were labeled “Translation 1” and “Translation 2”, and their system identities were hidden. This side-by-side layout allowed evaluators to compare all three texts without switching between workbook sheets. Candidate order was not counterbalanced, however, and one candidate may have helped an evaluator interpret the other.

Thirteen evaluators participated: one each for Spanish and English, three for Turkish, and two for each of the remaining languages. Each evaluator was a native speaker of the target language, proficient in Dutch, and employed by the stakeholder organization. They were familiar with pension communication and could assess each translation against the Dutch source. Evaluators received

the same written instructions and rubric but no formal calibration session. Participation was voluntary and based on informed consent, and the analysis used anonymized responses.

Evaluators scored meaning correctness and completeness, register and style, and readability and flow on a 1–10 anchored scale; Overall was their arithmetic mean (rubric and instructions in repository [45]).

Evaluator counts differed across languages, so we first averaged ratings within each language, system, and excerpt. Subsequent analyses used these excerpt-level means. We calculated paired differences between GPT-4 and Azure Translator for the six excerpts in each language. Within-language 95% confidence intervals came from bootstrap resampling of excerpts, and exact two-sided sign-flip tests used the six paired excerpt differences. Macro differences were the unweighted mean of the seven language-level differences. Macro confidence intervals resampled languages, and macro sign-flip tests used the seven language-level differences. This procedure gave each language equal weight.

As a supplementary analysis, a GPT-4 judge applied the same rubric to translations with system identity hidden. We compared its scores with the excerpt-level native-speaker means using Spearman rank correlation, mean absolute error (MAE), root mean squared error (RMSE), and a least-squares calibration fit of judge scores on native-speaker scores. We also calculated within-system correlations to distinguish item-level agreement from separation between the two systems. The judge results are provided in section 4.1 and Appendix 2.

3.4 Track B: Reference-Based Evaluation

The selected five-paragraph document was translated by each system into all seven languages, and aligned with the human reference at paragraph level. Because sentence boundaries differed across translations, we did not perform sentence-level alignment. Track B contains one document and reference per language, so it provides controlled comparative evidence rather than general system ratings.

We calculated BLEU, chrF++, and TER with `sacreBLEU` 2.6.0 [40, 41, 48]. We also calculated BERTScore F1 with `xlm-roberta-large` and COMET with `wmt22-comet-da` [46, 59]. BERTScore and COMET were multiplied by 100 for reporting. The main text focuses on chrF++ and COMET because they represent character-overlap and learned semantic evaluation. Appendix Table 14 reports the exact metric signatures, model identifiers, and settings for all metrics.

For BLEU, chrF++, and TER, document scores were calculated corpus-style over the five aligned paragraphs. BERTScore and COMET were calculated for each paragraph and averaged within the document. We then macro-averaged each system’s document scores across languages, weighting each language equally. Macro confidence intervals were obtained by resampling the seven languages. Appendix 3 gives the complete aggregation and bootstrap procedures, including the within-language paragraph resampling.

To further diagnose the outputs, we also calculated three paragraph-level probes. Length ratio identified unusually short or long output relative to the

Table 1. Track A native-speaker paired differences between GPT-4 and Azure Translator. Positive values favor GPT-4. The macro rows report confidence intervals obtained by resampling languages. The language-level rows report criterion differences by target language; brackets give the Overall confidence interval obtained by resampling the six excerpts.

<i>Macro differences across languages</i>				
Criterion	Mean difference		95% confidence interval	
Meaning correctness	+1.90		[1.37, 2.42]	
Register and style	+2.28		[1.12, 3.62]	
Readability	+1.77		[1.24, 2.40]	
Overall	+1.98		[1.31, 2.64]	
<i>Differences by target language</i>				
Language	Meaning	Register	Readability	Overall [95% CI]
Arabic	+2.08	+1.58	+1.83	+1.83 [1.14, 2.56]
English	+0.67	+0.00	+0.67	+0.44 [-0.11, 1.06]
German	+2.92	+2.25	+2.17	+2.44 [2.22, 2.72]
Polish	+1.33	+2.33	+1.33	+1.67 [0.94, 2.44]
Spanish	+2.00	+5.83	+1.50	+3.11 [2.44, 3.72]
Turkish	+2.72	+3.11	+3.39	+3.07 [2.61, 3.44]
Ukrainian	+1.58	+0.83	+1.50	+1.31 [0.75, 1.75]

reference. Acronym-preservation F1 checked exact retention of uppercase Latin-script acronyms. Punctuation preservation compared formatting-character counts. These probes are diagnostic signals rather than measures of translation quality. We note that the acronym probe may be unreliable for non-Latin or mixed-script realizations. See Appendix 3 for probe formulas and full paragraph-level results.

Finally, to link the two tracks, we compared the direction of the GPT-4 versus Azure differences and calculated exploratory Spearman correlations between Track A rubric means and Track B reference-based scores across the 14 shared system-language points. Because the tracks used different source texts, these comparisons provide directional triangulation rather than replication or metric validation.

4 Results

4.1 Track A: Reference-Free Evaluation

Track A comprises 84 language-excerpt-system combinations. Native-speaker ratings favor GPT-4 over Azure Translator on all three rubric dimensions. Across languages, the macro differences are +1.90 for meaning correctness, +2.28 for register and style, +1.77 for readability, and +1.98 Overall; all macro confidence intervals exclude zero, as shown in Table 1.

Overall point estimates favor GPT-4 in every language, with the largest differences for Spanish and Turkish and the smallest for English, whose confidence interval includes zero. The dominant criterion varies across languages, indicating a broad but non-uniform GPT-4 advantage.

Table 2. Illustrative issues reported in optional native-speaker comments. The examples explain possible sources of rubric penalties but do not estimate the prevalence of each issue.

Issue class	Representative observations	Potential consequence
Register and sensitive phrasing	Mixed formal and informal address in German, Spanish, and Polish. Overly direct wording about death in English, Ukrainian, and Arabic.	Could alter the institutional voice or make references to sensitive events too direct.
Pension terminology	Misleading or overly literal renderings of disability and partner-pension terms in Turkish and English.	Could change or obscure the intended pension concept.
Entities and template fields	Inconsistent treatment of UWV, translated URL components, and mishandling of PROEFDRUK, a Dutch “proof copy” watermark.	Could reduce consistency or corrupt protected content.
Structure and readability	Word-order, repetition, and right-to-left presentation problems in Arabic. Unnatural conditional constructions in Ukrainian.	Could make instructions harder to follow.

Table 3. Aggregate Track A rubric means by translation system and evaluator. Scores use a 1–10 scale. Overall is the mean of Correctness, Register and style, and Readability. Native-speaker ratings were first averaged within each language–system–excerpt item. Each row then averages the 42 language–excerpt items for its system and evaluator.

System	Evaluator	Correctness	Register and style	Readability	Overall
Azure Translator	Native speakers	6.0	5.7	6.1	5.9
GPT-4 (translator)	Native speakers	7.9	8.0	7.9	7.9
Azure Translator	GPT-4 judge	6.5	5.9	6.8	6.4
GPT-4 (translator)	GPT-4 judge	8.5	8.0	8.4	8.3

Optional comments identify issues involving formality, pension terminology, institutional entities, and structure or grammar (Table 2). Because comments were optional, these examples indicate failure types rather than prevalence; Appendix 2 provides language-specific examples.

The native-speaker results raise a practical question: can the auxiliary GPT-4 judge recover the same system-level contrast when human evaluation is unavailable? At the aggregate level, it does. Native speakers assign Overall means of 5.9 to Azure Translator and 7.9 to GPT-4, while the judge assigns means of 6.4 and 8.3, respectively. The aggregate results in Table 3 therefore show similar separation between the systems. Agreement on this broad contrast, however, does not establish that the judge can rank or calibrate individual outputs.

Across the 84 items, the pooled association between judge and native-speaker Overall scores is moderate (Spearman $\rho = 0.53$; MAE = 1.35). Within-system agreement is substantially weaker, with $\rho = 0.31$ for Azure outputs and $\rho = 0.09$ for GPT-4 outputs. The pooled correlation therefore partly reflects separation between the two translation systems rather than reliable ranking of outputs from the same system. Calibration slopes range from 0.52 to 0.71 across the rubric

Table 4. Agreement and calibration between GPT-4 judge scores and native-speaker excerpt-level means in Track A ($n = 84$). Spearman ρ measures rank agreement, while MAE and RMSE measure absolute error. The calibration fit $\hat{y} = a + bx$ predicts the judge score from the native-speaker score. Slopes below one indicate score compression. Nonzero intercepts indicate a score offset.

Rubric dimension	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

Table 5. Track B macro reference-based scores across seven target languages, reported as the mean and standard deviation across language-level document scores. BERTScore F1 and COMET are multiplied by 100. Higher values are better except for TER, where lower values are better. Systems are ordered by chrF++; bold marks the best mean.

System	Type	chrF++		BLEU		TER ↓		BERTScore F1		COMET	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Gemini 3 Pro	LLM	56.27	9.01	36.25	10.78	52.31	12.14	94.90	1.29	84.80	2.32
GPT-5.2	LLM	55.98	12.60	36.47	15.18	51.69	16.47	95.08	1.63	83.62	3.97
DeepL	NMT	54.67	11.19	34.79	14.29	53.17	14.94	94.66	1.62	83.33	3.35
GPT-4	LLM	51.18	11.84	29.39	13.83	57.83	15.78	94.31	1.71	82.93	3.27
Google Translate	NMT	49.65	8.77	27.75	9.09	60.43	11.92	93.72	1.15	83.18	3.20
Llama 3.1 405B	LLM	49.56	12.06	27.16	14.52	58.67	16.32	94.35	1.84	82.03	3.66
Azure Translator	NMT	48.83	10.06	26.86	10.99	60.31	12.98	94.00	1.69	80.12	3.24
Llama 3.3 70B	LLM	48.28	12.97	26.54	14.92	61.33	18.19	93.96	2.06	81.75	4.52
Llama 3.1 8B	LLM	42.60	12.99	20.48	14.04	69.31	20.93	93.15	2.21	77.60	5.75

dimensions, showing that the judge uses a narrower score range than the native speakers (Table 4).

Figure 1(a) shows the same pattern for Overall. Judge scores occupy a narrower range than the native-speaker means, and the separation between Azure and GPT-4 outputs contributes to the pooled association. The judge therefore preserves the broad system-level contrast but does not provide calibrated item-level acceptability scores. Appendix 2 reports the language-level judge scores and system differences.

4.2 Track B: Reference-Based Evaluation

Track B compares nine systems on the same five-paragraph document in seven languages. Gemini 3 Pro has the highest macro mean under both chrF++ and COMET. GPT-5.2 has the highest BLEU and BERTScore means and the lowest TER mean, while DeepL remains close to both systems across metrics. The identity of the leading system therefore depends on the metric family (Table 5).

The difference between Gemini 3 Pro and GPT-5.2 is 0.29 chrF++ points, while Gemini’s COMET advantage over GPT-5.2 is 1.18 points. These differences are small compared with variation across languages. Figure 1(b) shows the macro

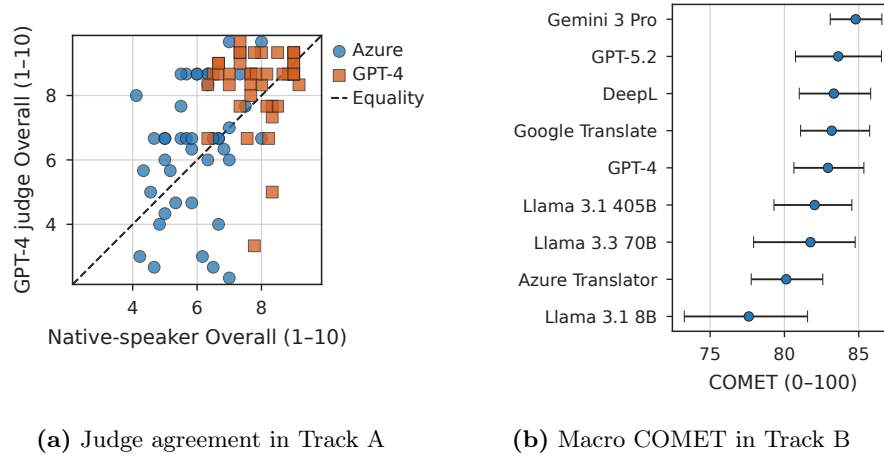


Fig. 1. Evaluation summaries for the two tracks. Panel (a) compares GPT-4 judge scores with native-speaker Overall means for 84 language–excerpt–system items. Marker shape identifies the translation system, and the dashed line marks equality. Panel (b) reports macro COMET by system across seven target languages. Error bars are 95% bootstrap confidence intervals obtained by resampling languages with replacement.

COMET estimates with confidence intervals obtained by resampling the seven target languages. The intervals overlap broadly among the leading systems, so the point estimates do not support a precise ordering of the top group.

The top group contains both LLMs and DeepL. Google Translate also scores close to the leading group under COMET, although its chrF++ score is lower. Given the small test set of one five-paragraph document per language, with COMET computed for each paragraph and then averaged, these scores do not establish a stable ordering of the leading systems. The observed order may reflect both sampling variation and genuine metric sensitivity. The benchmark therefore does not establish a general advantage for either NMT or LLMs.

Target language has a large effect. Averaged across systems, Arabic and Ukrainian receive the lowest reference-based scores, while German and Spanish receive the highest. The system-by-language chrF++ heatmap in Figure 2 also shows departures from the macro ordering, and no system leads every language.

The leading system under chrF++ and COMET is the same in four languages but differs in Turkish, Spanish, and Ukrainian. In Ukrainian, Llama 3.1 8B has the highest COMET score, but its 0.02-point lead over Gemini 3 Pro is not statistically significant. GPT-5.2 leads under chrF++, while Llama 3.1 8B ranks last in the macro evaluation. The Ukrainian result therefore illustrates metric-sensitive ranking on this document, not a general advantage for Llama 3.1 8B in Ukrainian.

Language variation is not the only source of ranking instability. Within the same target language, chrF++ and COMET sometimes assign different rela-

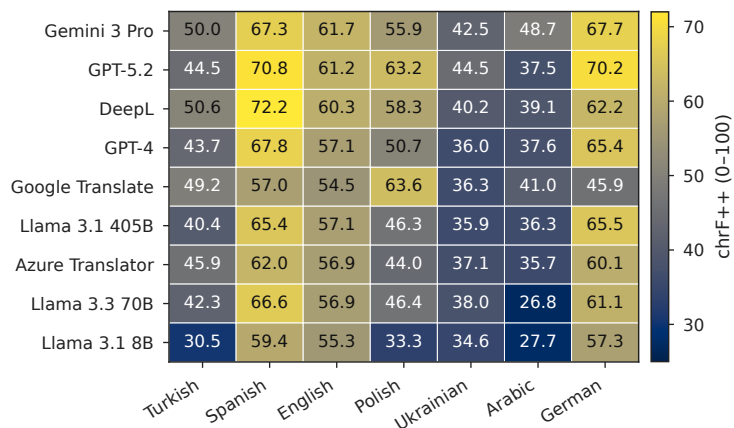


Fig. 2. Corpus-level chrF++ scores in Track B by translation system and target language. Scores range from 0 to 100, with higher scores indicating greater overlap with the professional reference. Rows are ordered by each system’s macro-average chrF++ across the seven languages, and cells show exact scores. Departures from this overall order within individual language columns show that the relative chrF++ ranking varies by target language.

tive scores to a system. Figure 3 plots the 63 language–system combinations and marks the ten largest differences between the within-language standardized metric scores. These cases identify strong metric disagreement, but they do not show which metric better represents translation quality. The figure therefore supports the narrower conclusion that metric choice can affect the apparent relative performance of systems.

Paragraph-level diagnostics identify possible localized risks. Ukrainian has the lowest mean length ratio at 0.732, followed by Arabic at 0.779. One Azure Ukrainian paragraph has a length ratio of 0.46, which is consistent with marked compression. Arabic has the lowest punctuation-preservation score at 0.396. These probes identify passages for review. They do not establish that every short or differently formatted translation contains a substantive error. Appendix 3 provides the full probe and paragraph-level results.

4.3 Cross-track Analysis

The shared GPT-4 versus Azure comparison examines whether the two tracks favor the same system within each language. Native-speaker Overall point estimates in Track A favor GPT-4 in all seven languages. Table 6 reports the language-level differences and the number of languages with matching directions. Because the columns use different score scales, the cross-track comparison concerns direction rather than magnitude.

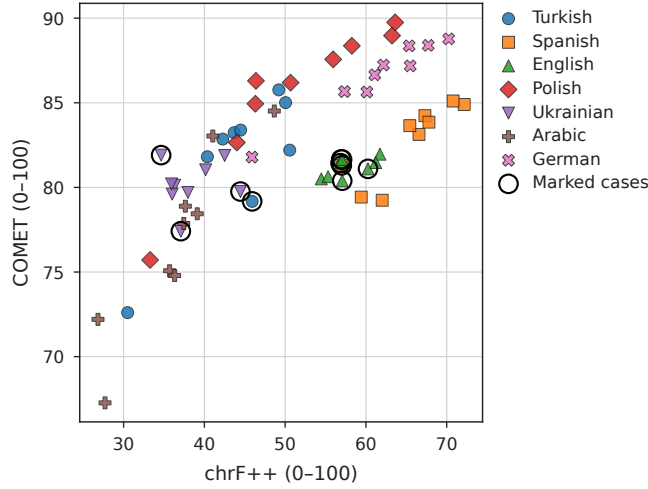


Fig. 3. Corpus-level chrF++ against COMET in Track B. Each point represents one of the 63 language–system combinations, and colors identify target languages. Within each language, the nine system scores were standardized separately for chrF++ and COMET. Black rings mark the five largest positive and five largest negative differences between these standardized scores. The marked points identify strong metric disagreement; they do not indicate which metric better represents translation quality.

COMET agrees with the Track A direction most often, differing only in English, which also has the smallest Track A Overall difference. chrF++ produces the opposite point-estimate direction in Turkish and Ukrainian, while TER does so in Polish and Ukrainian. These results provide directional triangulation rather than replication. Because the tracks used different source texts and evaluation methods, the analysis cannot determine whether the mismatches arise from the source texts, the evaluation methods, or both.

The directional comparison raises a more tentative question: across the 14 shared system–language points, do higher Track B metric scores tend to coincide with higher Track A native-speaker scores on the three rubric dimensions? COMET has the largest correlations, reaching $\rho = 0.50$ for Readability and $\rho = 0.36$ for Correctness and Register and style. The correlations for BLEU, chrF++, –TER, and BERTScore range from -0.05 to 0.23 (Table 7).

These associations combine differences between systems and languages and compare scores obtained from different source texts. They therefore do not validate any metric against human acceptability. Across this pilot, the tracks often agree on the direction of the GPT-4 versus Azure difference, but their disagreements show why reference-based metrics should not replace native-speaker assessment.

Table 6. Directional comparison of GPT-4 and Azure Translator across tracks. Track A reports native-speaker differences in Overall scores on the 10-point rubric, averaged over six excerpts per language. Track B reports corpus-level metric differences for a separate five-paragraph document. All differences are oriented so that positive values favor GPT-4. TER is reported as Azure minus GPT-4. The columns use different scales, so their signs, not their magnitudes, should be compared across tracks. Negative Track B point estimates are bold and indicate directions opposite to the Track A Overall point estimate.

Language	Track A: native speakers	Track B: reference metrics		
	Δ Overall	Δ chrF++	Δ COMET	$\Delta(-\text{TER})$
Arabic	+1.83	+1.95	+3.81	+2.48
English	+0.44	+0.13	− 1.26	+1.22
German	+2.44	+5.25	+2.73	+10.10
Polish	+1.67	+6.65	+3.54	− 0.53
Spanish	+3.11	+5.78	+4.61	+6.01
Turkish	+3.07	− 2.23	+4.06	+1.28
Ukrainian	+1.31	− 1.07	+2.19	− 3.14
Directional agreement with Track A		5/7	6/7	5/7

5 Discussion

5.1 Interpretation and Implications

The two evaluation tracks provide complementary evidence on the performance of off-the-shelf machine translation for regulated pension communication. Native speakers consistently prefer GPT-4 over Azure Translator across all seven languages, particularly for register and style. In the reference-based evaluation the strongest systems form a relatively tight cluster and rankings vary by language and metric. Established NMT systems remain competitive, while the strongest closed LLMs achieve best or near-best performance on several metrics. Thus,

Table 7. Exploratory Spearman correlations between Track A native-speaker rubric means and Track B corpus-level metrics across 14 shared system–language points, comprising seven languages and two systems. TER is multiplied by -1 so that higher scores indicate better performance. Because the tracks use different source texts, the correlations describe cross-track association and do not validate the metrics against human acceptability.

Track B metric	Track A: native-speaker rubric score		
	Correctness	Register and style	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

the relative advantage of LLM and NMT approaches appears to be system-, language-, and evaluation-dependent.

A central contribution of the study is the linkage between the two tracks through their shared GPT-4 versus Azure comparison. Because they use different source texts, this comparison constitutes triangulation rather than direct replication. Nevertheless, both point broadly in the same direction: human judgments consistently favor GPT-4, while semantic reference-based metrics often do so as well, even where overlap-based metrics diverge. This divergence is particularly relevant for regulated communication. Overlap metrics can penalize acceptable paraphrases, whereas native speakers also consider institutional register and pragmatic appropriateness. Learned semantic metrics, particularly COMET, align more closely with human preferences in our cross-track comparison, although this evidence remains exploratory. Paragraph-level analysis further shows that acceptable document averages can conceal localized omission or formatting risks, with Arabic and Ukrainian presenting particular challenges.

Taken together, these findings argue against using any single metric as an acceptance criterion. Controlled reference-based audits can support system comparison, quantify uncertainty, and identify localized failures, while human rubric-based assessment remains important for client-facing acceptability. Where references are unavailable, LLM-as-a-judge can provide a supplementary monitoring signal. The GPT-4 judge broadly reproduced the direction of human preferences, consistent with previous findings on LLM-based evaluation [26], but its scores were compressed and imperfectly calibrated. It is therefore more suitable for monitoring and triage than for final acceptance decisions.

Beyond evaluation and system selection, many observed failures are amenable to governance rather than improvements in model quality alone. Regulated MT pipelines should enforce language-specific register policies, compact termbases and do-not-translate lists, approved phrasing where required, protected template elements, and post-generation structure validation. Such controls directly target the terminology drift, register inconsistency, entity corruption, and formatting failures observed here [5, 42, 47]. LLM-based translation may offer an advantage in this respect because such constraints can easily be encoded in prompt and system instructions, or be specified interactively at inference time.

5.2 Limitations and Future Work

This study is an applied pilot in a focused domain rather than a definitive benchmark. Track A contains six excerpts with small and uneven native-speaker panels, without formal evaluator calibration; its fixed side-by-side presentation may also have influenced judgments. Track B contains only one five-paragraph document and one professional reference per language, making rankings sensitive to document and reference choice. Moreover, systems were accessed through different interfaces with only partially controllable generation settings, and each translation was generated once, so run-to-run variability was not measured. The evaluated systems are also a sample rather than an exhaustive comparison of current translation approaches.

Automatic evaluation introduces additional constraints. Single-reference metrics may penalize legitimate alternatives, while paragraph-level application of metrics commonly evaluated at sentence level may affect calibration. The operational probes identify potential omission, acronym, and formatting risks but cannot establish semantic correctness. The cross-track comparison is triangulation rather than replication because the tracks use different texts. Finally, the LLM judge is exploratory and may be affected by calibration or self-preference because GPT-4 was also evaluated as a translator.

Future work should expand the benchmark across more documents, templates, and professional references, and use larger calibrated native-speaker panels with risk-weighted error annotation. Repeated generation would enable robustness estimates. Evaluation should also test whether terminology constraints, protected-term lists, register instructions, and structure-preserving prompts reduce the observed failure modes. For production use, reference-free quality estimation, structural probes, and calibrated LLM judging could provide automated triage, with high-risk cases escalated to human review. Longitudinal monitoring could additionally detect regressions as models, prompts, and provider systems change.

6 Conclusion

This paper evaluated machine translation for regulated client communications with the pension domain as a case study, using two complementary tracks: native-speaker assessment of enterprise acceptability and a controlled reference-based benchmark. Native speakers consistently preferred GPT-4 over Azure Translator across all seven target languages, particularly for register and style. In the broader benchmark, rankings varied by language and metric, with the strongest systems forming a relatively tight top cluster. Paragraph-level diagnostics additionally revealed localized omission and formatting risks that document-level averages can obscure.

Overall, the study shows the value of combining human acceptability judgments, reference-based metrics, and localized risk diagnostics when evaluating machine translation for regulated communication. The results also suggest that deployment quality depends not only on selecting a strong translation system, but on controlling terminology, register, protected content, and document structure. LLMs may be particularly attractive in this respect because terminology, register, and structural constraints can be expressed directly through prompts and system instructions. Despite its pilot scale, the study provides a reproducible evaluation framework for assessing and governing multilingual pension communication.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Appendix

1 System and Generation Details

Table 8 lists the system types, model identifiers, and available version information for the systems evaluated in the two tracks. Table 9 reports the configurable generation settings for systems accessed through APIs. Parameters that providers did not expose could not be controlled. Table 10 summarizes the complete assignment of systems to evaluation tracks and target languages.

2 Track A Methods and Results

LLM-as-a-Judge Procedure and Supplementary Results The auxiliary judge analysis used GPT-4 to evaluate each translation with the system identity hidden. The judge received the Dutch source text, one target-language translation, and the scoring instructions (see rubric in repository). It returned separate scores for meaning correctness and completeness, register and style, and readability and flow. Overall was calculated as the arithmetic mean of the three scores.

Native-speaker ratings were first averaged within each language, system, and excerpt. Judge scores were then compared with these excerpt-level native-speaker means across all 84 language–excerpt–system items. We calculated Spearman rank correlation, mean absolute error, root mean squared error, and the least-squares calibration model shown in Eq. (1).

$$\hat{J}_i = a + bH_i. \quad (1)$$

In this model, H_i is the native-speaker mean and J_i is the judge score for item i . A slope below one indicates that the judge used a narrower score range than the native speakers.

Table 8. Systems evaluated in Tracks A and B, with access routes, model identifiers, and version information where available. Azure Translator was accessed through its REST API; GPT-4, GPT-5.2, and the Llama models through the Microsoft Foundry API; and Gemini 3 Pro, DeepL, and Google Translate through their web interfaces.

Type	System	Model identifier	Version information
NMT	Azure Translator	—	API v3.0
	DeepL Translator	—	Not exposed
	Google Translate	—	Not exposed
LLM	Gemini 3 Pro	—	Not exposed
	GPT-4	<code>gpt-4</code>	<code>turbo-2024-04-09</code>
	GPT-5.2	<code>gpt-5.2</code>	<code>2025-12-11</code>
	Llama 3.1 8B	<code>Meta-Llama-3.1-8B-Instruct</code>	6
	Llama 3.1 405B	<code>Meta-Llama-3.1-405B-Instruct</code>	1
	Llama 3.3 70B	<code>Llama-3.3-70B-Instruct</code>	9

Table 9. Generation settings for systems accessed through APIs. Sampling parameters do not apply to Azure Translator. Reasoning effort was available only for GPT-5.2. Azure Translator was accessed through Microsoft Translator Text API, others through Chat Completions API.

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

Table 10. System coverage in the two evaluation tracks.

System family	System	Track A	Track B
		Reference-free rubric-based	Reference-based metric-based
NMT	Azure Translator	✓	✓
	DeepL	—	✓
	Google Translate	—	✓
LLM	GPT-4	✓	✓
	Gemini 3 Pro	—	✓
	GPT-5.2	—	✓
	Llama 3.3 70B	—	✓
	Llama 3.1 405B	—	✓
	Llama 3.1 8B	—	✓

Note. Within each track, every included system was evaluated in all seven target languages: Turkish, Spanish, English, Polish, Ukrainian, Arabic, and German. Checkmarks indicate inclusion; a dash indicates exclusion. The tracks used different source materials and evaluation methods.

Figure 4 reports the native-speaker and judge scores by language, system, and rubric criterion. It supplements the aggregate comparison in the main text by showing how the judge’s scoring pattern varied across languages.

Table 11 reports the language-level GPT-4 minus Azure differences derived from judge scores. The corresponding native-speaker differences appear in the main Track A results.

Table 4 and Figure 1 in the main text report the item-level agreement and calibration results.

Detailed Qualitative Examples The optional evaluator comments identify problems involving institutional register, sensitive phrasing, pension terminology, protected entities, document structure, and readability. Tables 12 and 13 provide the detailed language-specific examples from which the compact failure classes in the main text were derived. Because comments were optional, the examples describe observed problems but do not estimate their prevalence.

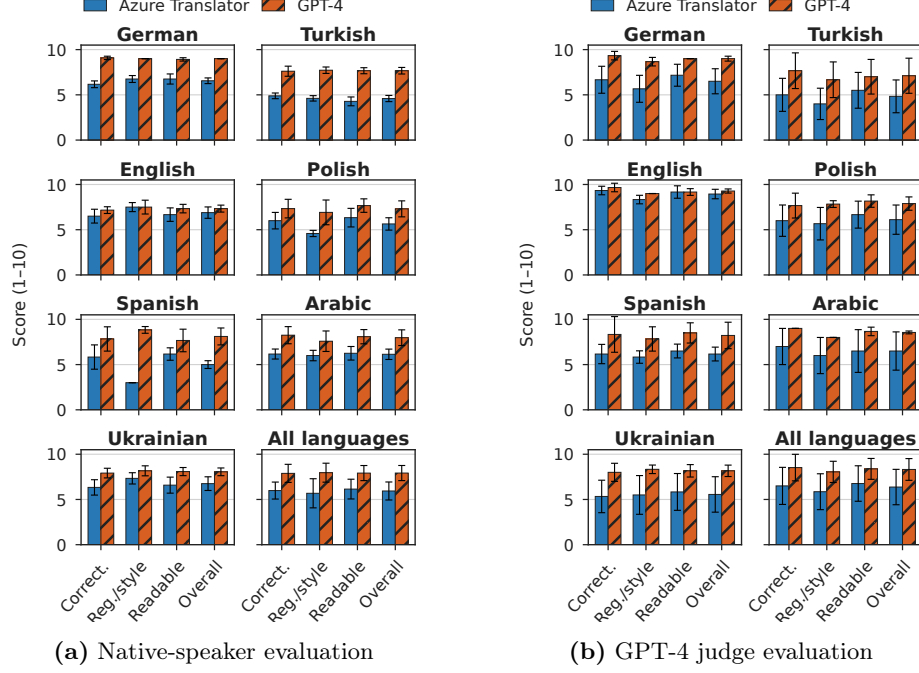


Fig. 4. Track A rubric results by language on a 1–10 scale. Panel (a) reports native-speaker evaluation, and panel (b) reports GPT-4 judge evaluation. Bars show mean scores for Azure Translator and GPT-4 on correctness, register and style, readability, and Overall. Error bars indicate ± 1 SD across the six excerpts within each language.

Table 11. LLM-as-a-judge paired deltas (GPT-4 minus Azure) by language for the reference-free track. Deltas are computed on excerpt-level means (paired by excerpt). 95% CIs are bootstrap over excerpts within each language. Significance markers: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$ (exact sign-flip test).

Language	Correctness Δ	Register and style Δ	Readability Δ	Overall Δ
German	2.67* [1.33, 3.83]	3.00** [2.00, 4.00]	1.83* [0.83, 2.83]	2.50** [1.44, 3.56]
Turkish	2.67** [1.50, 4.17]	2.67** [1.67, 4.17]	1.50 [0.50, 2.67]	2.28** [1.28, 3.61]
English	0.33 [0.00, 0.67]	0.67 [0.33, 1.00]	0.00 [-0.50, 0.50]	0.33 [0.00, 0.61]
Polish	1.67 [-0.33, 3.50]	2.17 [1.00, 3.33]	1.50 [0.17, 2.67]	1.78* [0.33, 3.00]
Spanish	2.17* [1.17, 2.83]	2.00* [1.00, 2.83]	2.00** [1.50, 2.50]	2.06** [1.28, 2.61]
Arabic	2.00 [0.67, 3.67]	2.00 [0.67, 3.67]	2.17 [0.67, 4.00]	2.06 [0.67, 3.78]
Ukrainian	2.67** [1.67, 3.83]	2.83** [1.33, 4.50]	2.33** [1.17, 3.67]	2.61** [1.39, 3.89]

Table 12. Track A native-speaker comments identified these issues for German, Turkish, Spanish, and English. The Dutch triggers quote Documents 1–6 verbatim and locate the source passage associated with each issue. **[quote]** marks evaluator-cited wording, and **[pattern]** marks paraphrased observations. Rubric tags identify correctness (C), register and style (T), and readability (R). Evaluators saw anonymized outputs with the system identity hidden.

Lang.	Dutch (Doc.)	trigger	System	Evaluator notes and examples
DE	“Beste meneer . . .” (Doc. 3); “Bij deze brief zit een folder.” (Doc. 1)		Azure	[quote] Mixed address forms (<i>Du/Sie</i>) 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 “ <i>Değerli Bay/Bayan</i> ”; suggested “ <i>Sayın</i> ”. (T) [quote] <i>arbeidsongeschiktheid</i> rendered literally; suggested “ <i>malulen emeklilik</i> ”. (C) [quote] <i>bijzonder partnerpensioen</i> rendered with “ <i>özel</i> ” 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 (<i>usted/tú</i>) 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] <i>arbeidsongeschikt</i> → “ <i>disabled</i> ” flagged; preferred “ <i>inability/incapacity for work</i> ”. (C) [quote] Framing shift: “ <i>survivor’s pension</i> ” vs. <i>partner pension</i> . (T) [quote] Blunt “ <i>when you die</i> ” penalized; suggested “ <i>when you pass away</i> ”. (T,R)

Table 13. Track A native-speaker comments identified these issues for Polish, Ukrainian, and Arabic. The Dutch triggers quote Documents 1–6 verbatim and locate the source passage associated with each issue. **[quote]** marks evaluator-cited wording, and **[pattern]** marks paraphrased observations. Rubric tags identify correctness (C), register and style (T), and readability (R). Evaluators saw anonymized outputs with the system identity hidden.

Lang.	Dutch (Doc.)	trigger	System	Evaluator notes and examples
PL	“Beste mevrouw . . .” (Doc. 5); “Dat betekent dat u geld mist.” (Doc. 1); PROEFDRUK (Doc. 3/5)	Azure		[quote] Switch into informal address (<i>ty</i>) in an official letter. (T) [pattern] Literal-but-unidiomatic constructions; layout-string handling reduces acceptability. (R,T)
		GPT		[quote] Mixed formal strategies (e.g., <i>Pan/Pani</i> vs. <i>Państwo</i>). (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 <i>Bu/Bau</i> 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/grammatical 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 14. 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

3 Track B Methods and Results

Dutch Source Document Track B used the following 197-word Dutch client letter. The document contains five paragraphs concerning a change in premium-free pension accrual following updated information about disability status and WAO/WIA benefits. The professional reference translations and system outputs retained these five paragraphs as the alignment units for evaluation.

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.

Metric Configurations Table 14 reports the implementations, signatures, and model identifiers used for BLEU, chrF++, TER, BERTScore, and COMET. BERTScore and COMET were multiplied by 100 for reporting.

For BLEU, chrF++, and TER, document scores were calculated corpus-style over the five aligned paragraphs. BERTScore and COMET were calculated

separately for each paragraph and then averaged within each document. System-level macro means assign equal weight to each of the seven target languages.

Macro confidence intervals were calculated by resampling the seven target languages. Within-language comparisons used 3,125 nonparametric bootstrap samples of the five aligned paragraphs with a fixed seed of 42. For BLEU, chrF++, and TER, the corpus score was recalculated for every bootstrap sample. For BERTScore and COMET, the mean of the resampled paragraph scores was recalculated.

Operational Probe Definitions Three paragraph-level probes were used to identify possible compression, acronym loss, and formatting drift. These probes identify passages for review but do not measure semantic correctness.

For a system paragraph s and its professional reference r , we defined the length ratio by Eq. (2).

$$\text{LenRatio}(s, r) = \frac{|s|_{\text{char}}}{|r|_{\text{char}}}. \quad (2)$$

Values below one indicate that the system output is shorter than the reference. Values above one indicate that it is longer. We defined length deviation by Eq. (3).

$$\text{LenDev}(s, r) = |1 - \text{LenRatio}(s, r)|. \quad (3)$$

Acronym preservation was calculated as exact-match set-based F_1 over uppercase Latin-script acronym tokens. Acronyms were extracted with the regular expression `\b[A-Z]{2,}(\?:/[A-Z]{2,})*\b`. This pattern captures forms such as UWV, WAO, and WAO/WIA.

Let A_s and A_r denote the acronym sets extracted from the system and reference paragraphs. We calculated precision and recall with Eq. (4).

$$P_{\text{acr}} = \frac{|A_s \cap A_r|}{|A_s|}, \quad R_{\text{acr}} = \frac{|A_s \cap A_r|}{|A_r|}. \quad (4)$$

We calculated acronym-preservation F_1 with Eq. (5).

$$F_{1,\text{acr}} = \frac{2P_{\text{acr}}R_{\text{acr}}}{P_{\text{acr}} + R_{\text{acr}}}. \quad (5)$$

The value of $F_{1,\text{acr}}$ was set to one when both sets were empty. Because extraction was limited to uppercase Latin-script forms, this probe may be unreliable for non-Latin or mixed-script realizations.

Punctuation preservation compared the counts of punctuation and formatting characters in the system and reference paragraphs. Let $c_k(x)$ denote the count of punctuation type k in text x . We calculated the preservation score with Eq. (6).

$$\text{PunctPres}(s, r) = \max\left(0, \min\left(1, 1 - \frac{\sum_k |c_k(s) - c_k(r)|}{\sum_k c_k(r) + 1}\right)\right). \quad (6)$$

Table 15. Reference-based track: language-level difficulty snapshot for the evaluated Dutch source document. We report the mean *corpus-level* metric value across all nine systems for each target language (higher is better except TER). These averages are descriptive (they compare languages on this document and reference set) and should not be interpreted as general-language benchmarks.

Language	BLEU	chrF++	TER	BERTScore-F1	COMET
Spanish	48.23	65.39	35.48	96.60	82.76
German	36.82	61.71	46.24	95.81	86.63
English	40.10	57.88	47.20	95.18	81.20
Polish	28.58	51.30	60.90	93.50	85.61
Turkish	23.71	44.12	71.58	94.01	81.78
Ukrainian	14.61	38.35	71.76	92.66	80.18
Arabic	14.59	36.72	75.21	91.91	76.90

Table 16. Bootstrap comparison of the top two systems per language. Differences in mean paragraph-level chrF++ (A–B) under paragraph-level bootstrap resampling ($n = 3125$). Positive values favor Model A. Checkmarks (✓) indicate 95% CIs excluding 0. Since corpus chrF++ is not equal to the arithmetic mean of paragraph-level chrF++, the corpus-best system (Model A) can occasionally have a slightly negative mean paragraph-level difference.

Language	Model A (corpus chrF++ best)	Model B	Mean diff.	95% CI	p	Sig.
Turkish	deepl	gemini-3-pro	0.23	[-6.77, 4.66]	0.804	–
Spanish	deepl	gpt-5-2	2.06	[-1.38, 5.51]	0.224	–
English	gemini-3-pro	gpt-5-2	0.29	[-2.52, 3.36]	0.846	–
Polish	google-translate	gpt-5-2	-0.80	[-7.66, 6.05]	0.855	–
Ukrainian	gpt-5-2	gemini-3-pro	2.55	[0.85, 4.25]	<0.001	✓
Arabic	gemini-3-pro	google-translate	6.23	[2.57, 10.31]	<0.001	✓
German	gpt-5-2	gemini-3-pro	2.09	[-1.53, 6.04]	0.309	–

A score of one indicates identical punctuation counts. Lower scores indicate larger count differences relative to the reference.

Language Effects and Ranking Uncertainty Table 15 reports the mean score for each target language across the nine systems. These values describe the controlled document and its reference translations; they are not general estimates of translation difficulty for the languages.

Table 16 reports paragraph-level bootstrap comparisons between the two leading chrF++ systems in each language.

Table 17 compares the leading system under chrF++ with the leading system under COMET for each language.

Metric Disagreement Table 18 reports paragraph-level disagreement between lexical and semantic metric composites. Each metric was standardized across the complete set of paragraph-level observations. The lexical and semantic composites were calculated as

Table 17. 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

Table 18. 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 $\times$ 7 languages). Right: disagreement rate aggregated per language over 45 paragraph pairs per language (9 systems $\times$ 5 paragraphs).

System	Disagree. rate	Language	Disagree. rate
gpt-4	11.4%	Spanish	15.6%
gpt-5-2	11.4%	Arabic	8.9%
meta-llama-3-1-8b-instruct	8.6%	Turkish	8.9%
deepl	5.7%	English	4.4%
llama-3-3-70b-instruct	5.7%	German	2.2%
meta-llama-3-1-405b-instruct	5.7%	Polish	2.2%
azure-translator	2.9%	Ukrainian	2.2%
gemini-3-pro	2.9%		
google-translate	2.9%		

$$\begin{aligned}
 z_{\text{lex}} &= \frac{z_{\text{BLEU}} + z_{\text{chrF++}} + z_{\text{-TER}}}{3}, \\
 z_{\text{sem}} &= \frac{z_{\text{BERTScore}} + z_{\text{COMET}}}{2}.
 \end{aligned}
 \tag{7}$$

TER was inverted before standardization so that higher values were better for every component. We classified an observation as a large disagreement when

$$|z_{\text{lex}} - z_{\text{sem}}| \geq 1. \tag{8}$$

The reported rates give the proportion of paragraph-level observations meeting this threshold by system and by language. This calculation differs from the main-text scatter plot: the table combines five paragraph-level metrics standardized over all observations, whereas the scatter compares corpus-level chrF++ and COMET standardized within each language.

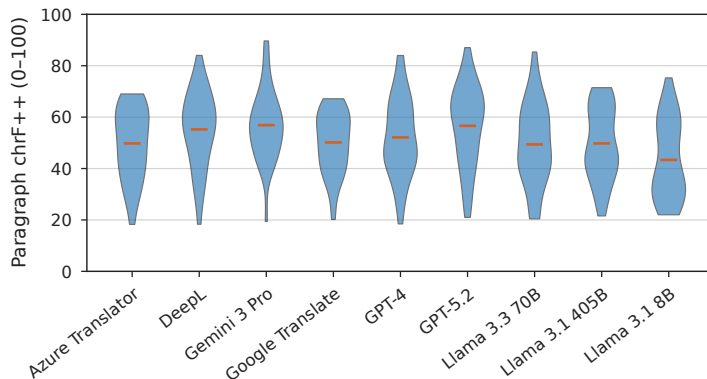


Fig. 5. Track B (reference-based): distribution of paragraph-level chrF++ (0–100) per system, pooled across all target languages (35 paragraph scores per system = 7 languages $\times$ 5 paragraphs). Violin width reflects score density; the mean is shown for each system. This view highlights within-document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

Paragraph-Level Diagnostics Figure 5 shows the distribution of paragraph-level chrF++ scores for each system. The distributions indicate whether low scores are broadly distributed or concentrated in a small number of paragraphs.

Table 19 reports the language-level probe averages. Figure 6 presents the corresponding system-level and language-level patterns.

Figure 7 shows how often each paragraph produced the lowest within-document chrF++ score. Table 20 reports mean chrF++ by language and paragraph.

Table 21 reports the lowest-scoring paragraph observed in each language together with its TER, COMET, length-ratio, and punctuation-preservation values.

Table 19. 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

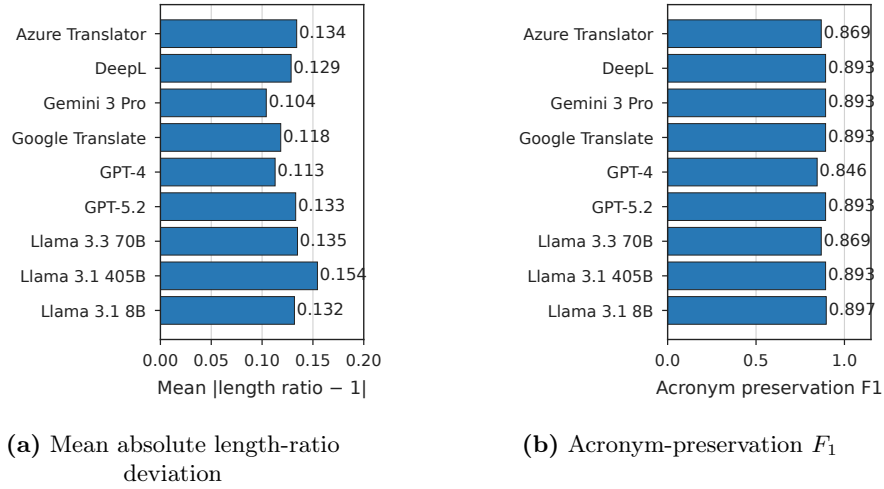


Fig. 6. System-level operational probes for Track B, the reference-based evaluation, macro-averaged over seven target languages. Panel (a) shows the mean absolute deviation of the normalized character-length ratio between the system output and the reference from 1, averaged across paragraphs. Lower values indicate a closer length match. Panel (b) shows exact-match set-based F_1 for preservation of uppercase Latin-script acronyms. Higher values indicate better preservation.

Table 20. 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

Table 21. 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. pres.
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

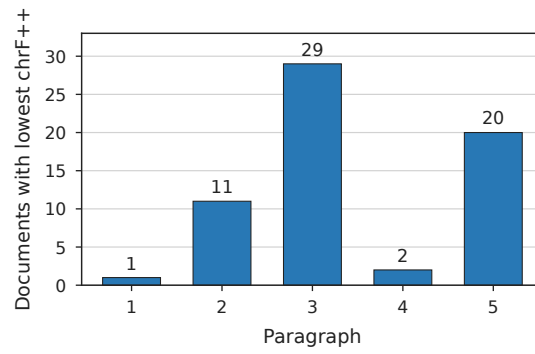


Fig. 7. Track B (reference-based): frequency with which each paragraph is the lowest-scoring segment (by paragraph-level chrF++) within a system–language document (63 documents = 9 systems $\times$ 7 languages; 5 paragraphs each). Concentration on specific paragraphs indicates localized difficulty and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

References

1. Ataman, D., Birch, A., Habash, N., Federico, M., Koehn, P., Cho, K.: Machine translation in the era of large language models: A survey of historical and emerging problems. *Information* **16**, 723 (08 2025). <https://doi.org/10.3390/info16090723>
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: *International Conference on Learning Representations (ICLR)* (2015), <https://arxiv.org/abs/1409.0473>
3. Bogoychev, N., Chen, P.: Terminology-aware translation with constrained decoding and large language model prompting. In: *Proceedings of the Eighth Conference on Machine Translation*. pp. 890–896 (2023). <https://doi.org/10.18653/v1/2023.wmt-1.80>, <https://aclanthology.org/2023.wmt-1.80/>
4. Callison-Burch, C., Osborne, M., 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*. pp. 249–256. Association for Computational Linguistics, Trento, Italy (Apr 2006), <https://aclanthology.org/E06-1032/>
5. Castilho, S., Knowles, R.: A survey of context in neural machine translation and its evaluation. *Natural Language Processing* (2024). <https://doi.org/10.1017/nlp.2024.7>, <https://www.cambridge.org/core/journals/natural-language-processing/article/survey-of-context-in-neural-machine-translation-and-its-evaluation/875C3E2DEAAB8845A0F05988D5A18B55>, published online 17 May 2024
6. Chatzikoumi, E.: How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering* **26**(2), 137–161 (2020). <https://doi.org/10.1017/S1351324919000469>
7. Chu, C., Wang, R.: A survey of domain adaptation for neural machine translation. In: *Proceedings of the 27th International Conference on Computational Linguistics*. pp. 1304–1319. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), <https://aclanthology.org/C18-1111/>
8. Dabre, R., Song, H., Exel, M., Buschbeck, B., Eschbach-Dymanus, J., 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)*. pp. 73–87 (2024), <https://aclanthology.org/2024.amta-research.8/>
9. Dale, D., Voita, E., Lam, J., Hansanti, P., Ropers, C., Kalbassi, E., 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), <https://aclanthology.org/2023.emnlp-main.42/>
10. Debets, S., Prast, H., Rossi, M., van Soest, A.: Pension communication in The Netherlands and other countries. *SSRN Electronic Journal* (01 2018). <https://doi.org/10.2139/ssrn.3284409>
11. Debets, S., Prast, H., Rossi, M., van Soest, A.: Pension communication, knowledge, and behaviour. *Journal of Pension Economics and Finance* **21**(1), 99–118 (2022). <https://doi.org/10.1017/S1474747220000232>, published online 8 September 2020
12. DeepL: About DeepL language models. <https://support.deepl.com/hc/en-us/articles/14241705319580-About-DeepL-language-models> (2026), DeepL Help Center. Accessed: 2026-04-12
13. DeepL Team: How does DeepL work? <https://www.deepl.com/en/blog/how-does-deepl-work> (2021), DeepL blog post. Last updated: 2021-10-31. Accessed: 2026-04-12

14. Dinu, G., Mathur, P., Federico, M., Al-Onaizan, Y.: Training neural machine translation to apply terminology constraints. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3063–3068 (2019). <https://doi.org/10.18653/v1/P19-1294>, <https://aclanthology.org/P19-1294/>
15. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., Macherey, W.: Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics* **9**, 1460–1474 (2021). https://doi.org/10.1162/tacl_a_00437, <https://aclanthology.org/2021.tacl-1.87/>
16. Freitag, M., Rei, R., Mathur, N., Lo, C.k., Stewart, C., Avramidis, E., 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). pp. 46–68. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022). <https://doi.org/10.18653/v1/2022.wmt-1.2>, <https://aclanthology.org/2022.wmt-1.2/>
17. Google Cloud: Neural Machine Translation (NMT). <https://docs.cloud.google.com/translate/docs/advanced/nmt-model> (2026), Google Cloud documentation. Accessed: 2026-04-12
18. Guerreiro, N.M., Alves, D.M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., et al.: Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics* **11**, 1500–1517 (2023). https://doi.org/10.1162/tacl_a_00615, <https://aclanthology.org/2023.tacl-1.85/>
19. Haddow, B., Bawden, R., Miceli Barone, A.V., Helcl, J., Birch, A.: Survey of low-resource machine translation. *Computational Linguistics* **48**(3), 673–732 (2022). https://doi.org/10.1162/coli_a_00446, <https://aclanthology.org/2022.cl-3.6/>
20. Hanneman, G., Dinu, G.: How should markup tags be translated? In: Proceedings of the Fifth Conference on Machine Translation (2020), <https://aclanthology.org/2020.wmt-1.138.pdf>
21. 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* **17**(3), 475–518 (Sep 2024). <https://doi.org/10.1007/s12062-023-09420-x>, published online 2 June 2023
22. Hokamp, C., Liu, Q.: Lexically constrained decoding for sequence generation using grid beam search. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1535–1546. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). <https://doi.org/10.18653/v1/p17-1141>, <https://aclanthology.org/P17-1141/>
23. Isabelle, P., Cherry, C., Foster, G.: A challenge set approach to evaluating machine translation. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2486–2496. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). <https://doi.org/10.18653/v1/d17-1263>, <https://aclanthology.org/D17-1263/>
24. Jiao, W., Wang, W., Huang, J.t., Wang, X., Shi, S., Tu, Z.: Is ChatGPT a good translator? yes with GPT-4 as the engine (2023). <https://doi.org/10.48550/arxiv.2301.08745>, <https://doi.org/10.48550/arXiv.2301.08745>
25. 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). pp. 147–165. Association for Computational Linguistics, Vienna, Austria (Jul 2025). <https://doi.org/10.18653/v1/2025.acl-industry.12>, <https://aclanthology.org/2025.acl-industry.12/>

26. Kocmi, T., Federmann, C.: Large language models are state-of-the-art evaluators of translation quality. In: Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S.A., Aranberri, N., Nunziatini, M., Escartín, C.P., Forcada, M., Popovic, M., Scarton, C., Moniz, H. (eds.) *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. pp. 193–203. European Association for Machine Translation, Tampere, Finland (Jun 2023), <https://aclanthology.org/2023.eamt-1.19/>
27. Koehn, P.: Statistical significance tests for machine translation evaluation. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. pp. 388–395. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), <https://aclanthology.org/W04-3250/>
28. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Luong, T., Birch, A., Neubig, G., Finch, A. (eds.) *Proceedings of the First Workshop on Neural Machine Translation*. pp. 28–39. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). <https://doi.org/10.18653/v1/w17-3204>, <https://aclanthology.org/W17-3204/>
29. Koponen, M., Nurminen, M.: Risk management for content delivery via raw machine translation. In: Winters, M., Deane-Cox, S., Böser, U. (eds.) *Translation, Interpreting and Technological Change: Innovations in Research, Practice and Training*, pp. 111–135. Bloomsbury Academic, London (2024). <https://doi.org/10.5040/9781350212978.0013>, <https://trepo.tuni.fi/handle/10024/209574>
30. Läubli, S., Amrhein, C., Düggelein, P., Gonzalez, B., Zwahlen, A., 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*. pp. 267–272. European Association for Machine Translation, Dublin, Ireland (Aug 2019), <https://aclanthology.org/W19-6626/>
31. Lentz, L., Nell, L., Pander Maat, H.: Begrijpelijkheid van pensioencommunicatie: effecten van wetgeving, geletterdheid en revisies. *Tijdschrift voor Taalbeheersing* **39**(2), 191–208 (2017). <https://doi.org/10.5117/TVT2017.2.LENT>, <https://dspace.library.uu.nl/bitstream/handle/1874/358493/s5.pdf?sequence=1>
32. Li, A., Zhou, W., Hoda, R., Bain, C., Poon, P.: Comparing large language models and traditional machine translation tools for translating medical consultation summaries: A pilot study (2025). <https://doi.org/10.48550/arXiv.2504.16601>, <https://arxiv.org/abs/2504.16601>
33. Lommel, A.R., Burchardt, A., Uszkoreit, H.: Multidimensional quality metrics: a flexible system for assessing translation quality. In: *Proceedings of Translating and the Computer 35*. Aslib, London, UK (Nov 2013), <https://aclanthology.org/2013.tc-1.6/>, november 28–29
34. Macken, L., Fonteyne, M., Tezcan, A., van Hest, E., Maryns, K., De Wilde, J.: Machine translation in asylum reception centres: System selection and multilingual quality evaluation. *Tradumàtica: Tecnologies de la Traducció* **23**, 326–349 (2025). <https://doi.org/10.5565/rev/tradumatica.522>, <https://doi.org/10.5565/rev/tradumatica.522>
35. Manakhimova, S., Avramidis, E., Macketanz, V., Lapshinova-Koltunski, E., Bagdasarov, S., 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*. pp. 224–245 (2023), <https://aclanthology.org/2023.wmt-1.23/>

36. Microsoft: Azure Translator documentation. <https://learn.microsoft.com/en-us/azure/ai-services/translator/> (2026), microsoft Learn documentation. Accessed: 2026-04-12
37. Moslem, Y., Romani, G., Molaei, M., Kelleher, J.D., Haque, R., Way, A.: Domain terminology integration into machine translation: Leveraging large language models. In: Koehn, P., Haddow, B., Kocmi, T., Monz, C. (eds.) Proceedings of the Eighth Conference on Machine Translation. pp. 902–911. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.wmt-1.82>, <https://aclanthology.org/2023.wmt-1.82/>
38. Nunziatini, M.: Machine translation in the financial services industry: A case study. In: Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks. pp. 57–63. European Association for Machine Translation, Dublin, Ireland (Aug 2019), <https://aclanthology.org/W19-6709/>
39. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
40. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). <https://doi.org/10.3115/1073083.1073135>, <https://aclanthology.org/P02-1040/>
41. Popović, M.: chrF++: words helping character n-grams. In: Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Kreutzer, J. (eds.) Proceedings of the Second Conference on Machine Translation. pp. 612–618. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). <https://doi.org/10.18653/v1/W17-4770>, <https://aclanthology.org/W17-4770/>
42. Post, M., Vilar, D.: Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1314–1324. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). <https://doi.org/10.18653/v1/n18-1119>, <https://aclanthology.org/N18-1119/>
43. 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* **14**, 1–29 (2025). <https://doi.org/10.14762/jll.2025.001>, <https://www.languageandlaw.eu/jll/article/download/201/97/748>
44. Qian, S., Sindhuja, A., Kabra, M., Kanojia, D., Orasan, C., Ranasinghe, T., 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. pp. 3660–3674. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). <https://doi.org/10.18653/v1/2024.emnlp-main.214>, <https://aclanthology.org/2024.emnlp-main.214/>
45. Radunovic, N.: Reproducibility materials for: A two-track evaluation of llm and nmt translation for regulated pension communication. Zenodo (Apr 2026). <https://doi.org/10.5281/zenodo.19821399>, version 1.0.0, software and data archive, MIT License
46. Rei, R., Stewart, C., Farinha, A.C., Lavie, A., Martins, A.F.T.: COMET: A neural framework for MT evaluation. In: Webber, B., Cohn, T., He, Y., Liu, Y.

- (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2685–2702. Association for Computational Linguistics, Online (Nov 2020). <https://doi.org/10.18653/v1/2020.emnlp-main.213>, <https://aclanthology.org/2020.emnlp-main.213/>
47. Sennrich, R., Haddow, B., Birch, A.: Controlling politeness in neural machine translation via side constraints. In: Knight, K., Nenkova, A., Rambow, O. (eds.) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 35–40. Association for Computational Linguistics, San Diego, California (Jun 2016). <https://doi.org/10.18653/v1/n16-1005>, <https://aclanthology.org/N16-1005/>
 48. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., 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. pp. 223–231. Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA (Aug 2006), <https://aclanthology.org/2006.amta-papers.25/>
 49. Sobczyk-Turek, A.: De Nederlandse taalvaardigheid onder arbeidsmigranten: De resultaten van het 6e arbeidsmigrantenpanel. Report, Het Kenniscentrum Arbeidsmigranten (Oct 2023), <https://hetkenniscentrumarbeidsmigranten.nl/wp-content/uploads/2023/10/Eindrapport-De-Nederlandse-taalvaardigheid-onder-arbeidsmigranten.pdf>, edited by Jan Cremers, based on research by I&O Research
 50. Specia, L., Scarton, C., Paetzold, G.H.: Quality Estimation for Machine Translation. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2018). <https://doi.org/10.2200/s00854ed1v01y201805hlt039>
 51. Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., et al.: Neural machine translation: A review of methods, resources, and tools. AI Open (2020), <https://arxiv.org/abs/2012.15515>
 52. Translayte: Professional translation services. <https://translayte.com/professional-translations> (2026), translayte website. Accessed: 2026-08-20
 53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017), <https://arxiv.org/abs/1706.03762>
 54. Vilar, D., Xu, J., D’Haro, L.F., Ney, H.: Error analysis of statistical machine translation output. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06). European Language Resources Association (ELRA), Genoa, Italy (May 2006), <https://aclanthology.org/L06-1141/>
 55. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., et al.: Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016), <https://arxiv.org/abs/1609.08144>
 56. Yan, J., Yan, P., Chen, Y., Li, J., Zhu, X., Zhang, Y.: Benchmarking gpt-4 against human translators: A comprehensive evaluation across languages, domains, and expertise levels (2024). <https://doi.org/10.48550/arxiv.2411.13775>, <https://arxiv.org/abs/2411.13775>
 57. Yang, M., Li, F.: Improving machine translation formality with large language models. Computers, Materials & Continua **82**(2), 2061–2075 (2025). <https://doi.org/10.32604/cmc.2024.058248>, <https://www.techscience.com/cmc/v82n2/58355>
 58. Zerva, C., Blain, F., C. De Souza, J.G., Kanojia, D., Deoghare, S., Guerreiro, N.M., 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. pp. 82–109. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). <https://doi.org/10.18653/v1/2024.wmt-1.3>, <https://aclanthology.org/2024.wmt-1.3/>
59. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (2020), <https://openreview.net/forum?id=SkeHuCVFDr>
60. Zhang, X., Rajabi, N., Duh, K., 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). pp. 468–481. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.wmt-1.43>, <https://aclanthology.org/2023.wmt-1.43/>
61. Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., et al.: Multilingual machine translation with large language models: Empirical results and analysis. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 2765–2781 (2024). <https://doi.org/10.18653/v1/2024.findings-naacl.176>, <https://aclanthology.org/2024.findings-naacl.176/>