ISSN 2669-2023 (print) / ISSN 2538-8959 (on-line) **/ https://doi.org/10.47459/cndcgs.2026.110**

Evaluating AI Translation Capabilities for Defence and Security Use

Eva Štěpánková[1], Karel Pešek[2], Denisa Sarı[2]*

[1]Department of Resources Management, Faculty of Military Leadership, University of Defence, Address: Kounicova 65, 662 10 Brno, Czech Republic
[2]Institute of Intelligence Studies, University of Defence, Address: Kounicova 65, 662 10 Brno, Czech Republic

*Correspondence:*eva.stepankova@unob.cz*

Abstract

This study evaluates the translation capabilities of selected AI language models in a defence and security context through a multistage multilingual workflow combining translation, language identification, and back-translation. Using 32 English inputs across eight target languages, the study compares offline and online models in terms of translation quality, processing stability, and task performance across varying levels of sentence complexity. The findings indicate that all models perform strongly on basic translation tasks, but notable differences emerge in language identification, back-translation quality, and output stability. The results help clarify the conditions under which AI translation can support security-relevant multilingual workflows.

KEY WORDS: *AI translation; defence and security; multilingual processing; back-translation; large language models.*

Citation: Štěpánková, E.; Pešek, K.; Sari, D. Evaluating AI Translation Capabilities for Defence and Security Use: A Comparative Study of Multistage Multilingual Processing. In Proceedings of the Challenges to National Defence in Contemporary Geopolitical Situation, Brno, Czech Republic, 7-10 September 2026. ISSN 2538-8959. https://doi.org/10.47459/cndcgs.2026.110

1. Introduction

The application of artificial intelligence in the field of Natural Language Processing (NLP) has significantly expanded the possibilities of automated translation and multilingual text processing in environments where rapid and reliable language mediation is operationally important. In defence and security contexts, these capabilities are particularly relevant, as military organisations, intelligence structures, and security institutions routinely operate across multilingual information environments in which the timely interpretation of foreign-language content may affect situational awareness, analytical accuracy, and decision-making. Although contemporary deep learning-based language models are capable of producing fluent and semantically rich translations across a broad range of languages, their practical utility in defence-related settings depends not only on output fluency, but also on consistency, robustness, and the ability to preserve meaning under multi-step processing conditions.

This study therefore examines the translation capabilities of selected artificial intelligence models in a defence and security context, with particular attention to their suitability for multilingual workflows relevant to intelligence support, operational communication, and the processing of foreign-language materials. The analysis includes three offline models - Mistral-7B, Gemma-3-12B, and OpenAI-20B - as well as the online model ChatGPT-5.2. Their performance is assessed through a three-stage procedure consisting of translation from English into a target language, language identification, and back-translation into English. This design makes it possible to evaluate not only translation accuracy and linguistic fluency, but also the extent to which individual models preserve the meaning of the original message throughout a chained processing sequence, which more closely reflects real-world security and defence use cases than isolated one-step translation alone.

The aim of the study is to compare the behaviour of the selected AI models in terms of output quality, processing demands, and performance stability across different languages and varying levels of sentence complexity. The evaluation combines quantitative indicators, such as processing time and back-translation correspondence scores, with qualitative analysis focused on semantic shifts, stylistic deviations, and the preservation of communicative intent. In a defence and security setting, such distinctions are not merely linguistic in nature - even minor distortions in meaning may affect the interpretation of actors, actions, intent, or context. The study therefore seeks to provide a transparent and empirically grounded comparison of selected AI models as translation-support tools for security-relevant multilingual environments,

Peer-review under responsibility of General Jonas Žemaitis Miltitary Academy of Lithuania and University of Defence, Czech Republic

rather than as general-purpose consumer applications. Its findings may serve as a basis for more informed decisions regarding the responsible deployment of AI translation systems in contexts where both linguistic quality and operational reliability are of critical importance.

2. Theoretical Background

Recent research suggests that AI translation in defence and security contexts should be understood not as a stand-alone convenience, but as an enabling capability for multilingual intelligence exploitation, coalition interoperability, and rapid orientation in foreign-language information environments. This logic is already visible in the DARPA LORELEI programme, which prioritised timely and operationally useful linguistic support for low-resource languages rather than stylistically polished translation outputs [1]. Defence-oriented NLP research reaches a similar conclusion by identifying multilingual operations, peacekeeping, and intelligence sharing as environments in which low-resource and domain-specific languages remain a persistent operational bottleneck [2].

At the capability level, large language models have substantially improved machine translation. Recent studies show that contemporary LLMs can achieve strong multilingual translation performance across many directions, although results remain sensitive to model scale, language pair, and prompting strategy [3]. Comparative research further indicates that GPT-based systems are often competitive for high-resource languages but remain weaker than specialised MT systems in many low-resource scenarios [4]. This distinction is particularly important for defence and intelligence practice, where operationally relevant languages are often underrepresented and domain conditions differ markedly from general-purpose benchmarks.

The literature also shows that translation quality depends not only on the model itself, but on the inference protocol through which the model is used. Prompt design, task framing, and decoding settings can significantly affect output quality, robustness, and the incidence of hallucinations [5]. In security contexts, this means that evaluation should not focus solely on model identity, but also on the wider procedural environment in which the model is deployed.

This issue becomes even more significant in military, doctrinal, and intelligence-related texts. Research on military translation highlights that such materials require the precise rendering of technical vocabulary, relational structure, and domain-specific meaning [6]. More broadly, work on low-resource MT identifies domain similarity as a major predictor of translation performance [7]. Strong benchmark performance on generic corpora therefore cannot automatically be treated as evidence of reliable behaviour on doctrinal, operational, or intelligence materials.

Another relevant strand of scholarship concerns discourse and context. Whereas traditional machine translation (MT) evaluation often focuses on sentence-level quality, security-relevant materials frequently require document-level coherence, reference consistency, and preservation of temporal and actor relations across longer passages. Recent findings suggest that LLM-based document-level MT can outperform conventional systems in some discourse-sensitive settings, particularly when long-context processing is effectively used [8]. For defence practice, this is important because operational misunderstanding can arise not only from lexical mistranslation, but also from the loss of coherence across a document.

The risk dimension is equally important. Hallucinations remain a substantial problem in multilingual translation, especially across diverse translation directions and low-resource conditions [9]. In defence and intelligence settings, even a small semantic shift may distort judgments regarding actor identity, intent, capability, time, or location. Translation error in such contexts is therefore not merely a linguistic defect, but a potential analytical and operational liability.

This makes evaluation a central issue. Traditional automatic metrics such as BLEU remain influential in MT research [10], while newer neural approaches such as COMET have improved correlation with human judgment [11]. At the same time, large-scale studies of human evaluation show that automatic metrics alone are insufficient and may produce misleading conclusions if they are not complemented by expert-based assessment [12,13]. The same caution applies to the use of LLMs as evaluators: recent evidence suggests that LLM-based MT evaluation benefits from reference translations but does not always provide stable or fully reliable judgments [14]. For high-stakes security use, credible evaluation therefore requires a combined framework incorporating automatic metrics, expert review, terminology checks, and workflow-based validation.

The multilingual benchmark literature reinforces this conclusion. FLORES-101 was developed precisely to address weaknesses in multilingual and low-resource evaluation [15], while the NLLB programme demonstrated both the scale of recent progress and the continuing challenge of achieving robust quality across very large language sets [16]. These developments confirm that multilingual evaluation must be treated as a structured empirical task rather than an impressionistic comparison of outputs.

Within the broader defence and intelligence literature, AI language tools are increasingly understood as instruments of decision support rather than decision substitution. Experimental evidence suggests that AI assistance can improve aspects of military intelligence analysis under time pressure, while also revealing limitations under ambiguous or contradictory informational conditions [17]. At the policy level, NATO's revised AI strategy formalises reliability, explainability and traceability, governability, and bias mitigation as core conditions for operational AI adoption [18]. In the case of AI translation, this implies that the most suitable model is not necessarily the one that produces the most fluent output, but the one whose behaviour can be assessed, constrained, and integrated into secure workflows.

Overall, the literature indicates that AI translation is already useful for analytical support, multilingual screening, and selected defence-related tasks, but remains uneven in low-resource, terminology-dense, and security-sensitive scenarios. The core research problem is therefore no longer whether AI can translate, but under what conditions translation quality,

stability, and evaluation are sufficient for responsible operational use. In this respect, studies focusing on chained tasks such as translation, language identification, and back-translation are especially valuable, because they reflect more closely the realities of defence and intelligence workflows [19,20].

3. Methodology

The study provides a systematic assessment of the translation quality generated by selected AI language models, comparing their ability to process texts of varying linguistic and stylistic complexity. The analysis focuses on the behaviour of models with different parameter scales when performing three sequential language tasks — translation, language detection, and back-translation — applied to a fixed dataset of 32 English sentences or short texts across eight target languages, serving as a standardised input for all four evaluated models. Particular attention is devoted to the time requirements of individual tasks, as well as to the qualitative evaluation of translation accuracy, meaning preservation, and the stylistic naturalness of the resulting text across different languages and levels of sentence complexity. The individual stages of the analysis are as follows:

1. Translation from English into the Target Language - 32 English texts were translated by each selected AI model into one of eight predefined target languages. In total, each model therefore translated 256 sentences. The target languages were Czech, French, German, Spanish, Russian, Ukrainian, Japanese, and Chinese. For each sentence, the time required to complete the translation was recorded as an indicator of processing speed.
2. Language Detection - each AI model was prompted to identify the language in which the submitted sentences were written. Two aspects were evaluated: (1) the accuracy of language identification and (2) time required to perform the detection task.
3. Back-translation into English - the third phase of the experiment consisted of translating sentences written in a non-English language back into English. For each sentence, both the time required to perform the translation, and the translation quality were recorded. Back-translation represents a more complex task, as it combines comprehension of a foreign-language input with the generation of an appropriate English sentence. It therefore allows a more detailed assessment of the ability of AI models to preserve meaning across linguistic transformations.
4. Evaluation of Translation Quality - finally, a manual assessment was conducted comparing the original English sentence with its back-translated version derived from the non-English text. The evaluation was carried out manually by one of the article's authors on the basis of predefined criteria: (1) degree of literalness; (2) semantic equivalence; (3) stylistic and linguistic naturalness, and (4) overall similarity (i.e., a summary judgement of correspondence between the original and resulting sentence).

ChatGPT was employed in this phase as a supportive tool for assessing linguistic naturalness and stylistic appropriateness. Based on the criteria outlined above, each sentence was assigned a numerical score on a scale from 0 to 1, where 1.0 represented complete correspondence between the original and back-translated sentence, while 0.0 indicated a complete absence of semantic or formal correspondence (the back-translated sentence either did not reflect the original sentence at all or conveyed an entirely different meaning). This scale enabled a quantitative comparison of the performance of individual AI models across all languages and categories of textual difficulty.

The evaluation of translation performance also incorporated the time required to complete individual tasks, with results further aggregated according to sentence complexity. For each category (simple, moderately complex, and complex sentences), average, minimum, and maximum values for processing time and translation quality were calculated. This made it possible to capture not only overall model performance, but also its variability.

In summary, the following parameters were examined in the assessment of the translation performance of the individual AI models: (1) The average time required to translate from English into the specified target language; (2) Language detection performance, including detection quality and average processing time; (3) Translation from a non-English language back into English, including translation quality and the average time required for back-translation.

The proposed methodological framework enables a comprehensive evaluation of language models by combining time-based performance analysis with a qualitative assessment of back-translation outputs. It therefore provides an integrated view of AI model behaviour when processing different languages and varying levels of input complexity (sentences or short texts).

Dataset – Typology and Linguistic Characteristics of the Analysed Texts

For the purposes of this case study, a dataset of 32 original English sentences and short texts was compiled. These items differ in their degree of linguistic and terminological complexity, making it possible to observe how this variable influences the performance of AI models across the evaluated tasks. The dataset consists of three categories of texts and sentences:

- Simple Sentences (8 items) - short, structurally simple sentences and phrases typical of everyday informal communication, such as greetings and routine conversational expressions. They are characterised by straightforward syntactic structure, common lexical items, the absence of metaphorical or specialised

terminology, and minimal dependence on broader context. These sentences serve as a control baseline, enabling verification of whether artificial intelligence models can accurately translate elementary communicative content without error.

- Moderately Complex Sentences / Texts (16 items) - A linguistically and stylistically diverse set of sentences representing standard English communication commonly found in discussions, descriptions, instructions, popular science texts, or mildly specialised contexts. These items are longer, structurally more varied, and thematically diverse, including explanations, narrative passages, descriptive content, as well as wordplay or humour. This category was designed to test the ability of models to process natural language in realistic communicative situations, where interpretation of meaning, style, and context is required.
- Specialised and Academic Sentences (8 items) - The most linguistically and cognitively demanding category, consisting of sentences adapted from academic articles, professional studies, technical publications, or specialist descriptions. These sentences are characterised by high informational density, precise terminology, and complex syntactic structure. They were included to assess the extent to which AI models are capable of preserving specialist meaning, terminological accuracy, and stylistic adequacy, particularly during multi-stage translation and back-translation into English.

The division of sentences into these categories makes it possible to examine AI model performance in relation to linguistic complexity, degree of contextual dependence, and demands for semantic precision and stylistic naturalness. This typology creates a methodologically balanced corpus that reflects both everyday language use and demanding professional contexts in which translation quality has a direct impact on intelligibility and communicative accuracy.

AI Models Used for Translation

The translations of the input materials (sentences and short texts) were carried out using the following AI models:

- *Offline Language Models* - Recent offline generative artificial intelligence language models were included in the experiment. These models differ primarily in the number of parameters, which determines their structural complexity and capacity to capture linguistic patterns. Parameters represent internal weights learned from data during training; the higher their number, the greater the volume of information and relational patterns that the model is capable of representing. The number of parameters is commonly indicated in the model name using the format nb, where n denotes the number of billions of parameters. One of the principal variables examined in this study is precisely the effect of differing parameter scales. The objective was therefore to compare models of different parametric sizes, which also informed their selection. It should, however, be emphasised that larger variants of the same model family (for example, AI Mistral) would, with a high degree of probability, achieve different, typically stronger, results than their smaller counterparts. The specific offline models tested were:

    1. Mistral AI Mistral-7B-Instruct-v0.3
    2. Google Gemma-3-12B
    3. OpenAI gpt-oss-20B

- *GPT-5.2 Model* - in addition, the online model OpenAI GPT-5.2 was included in the comparison.

All models were tested under identical conditions, using the same input data and equivalent task instructions. They differ, however, in architecture, training data, and model scale (number of parameters). The observed differences in results may therefore be interpreted, among other factors, as a consequence of these underlying distinctions.

The findings of the experiment should be understood strictly within the context of the specific experimental design and do not constitute a general evaluation of the overall translation capabilities of the individual models. Rather, the purpose of the study is to provide a transparent and reproducible comparison, not a normative ranking.

4. Findings

The aim of the analysis was to evaluate the quality of multi-step translation performed by the individual AI models. This process was divided into three phases: (1) translation from English into selected target non-English languages (specified to the model); (2) identification of the non-English language; (3) back-translation into English. In each phase, certain parameters are monitored (see above), on the basis of which the overall translation "competence" of the individual AI models is evaluated. The following subsections present the results of the individual phases of the translation process.

Time Required for Translation Tasks

First, the time demands of the individual language tasks were compared across all the evaluated AI models, specifically the time required for translation from English into the target language and for back-translation into English. The aim is to identify differences in processing speed between the individual models and to assess how the time demands vary

depending on the type of task (translation and back-translation) and the complexity of the input (the individual languages and sentences). The results of the comparison of AI models in the phase of translation from English into the specified language are presented in Table 1, which shows the average translation times for a set of 32 sentences in 8 languages for the individual AI models.

Table 1.

Time required for translation from English into the specified language according to the individual AI models and sentence complexity

	Simple sentences			Moderately complex sentences			Complex sentences		
	Mistral	Gemma	Open AI	Mistral	Gemma	Open AI	Mistral	Gemma	Open AI
Czech	0.119	0.171	1.098	0.600	0.761	1.470	1.835	1.681	2.464
German	0.109	0.162	0.453	0.495	0.702	1.052	1.295	1.482	2.099
Spanish	0.102	0.152	0.497	0.475	0.620	1.020	1.063	1.390	1.906
French	0.117	0.201	0.609	0.467	0.671	1.125	1.101	1.540	2.103
Chinese	0.111	0.133	0.683	0.478	0.500	0.977	1.318	1.101	1.794
Japanese	0.234	0.135	0.853	0.797	0.611	1.279	1.895	1.270	2.549
Ukrainian	0.179	0.181	0.911	0.590	0.770	1.347	1.914	1.772	2.706
Russian	0.127	0.150	0.585	0.547	0.650	1.112	1.410	1.441	2.075
Average	0.137	0.161	0.711	0.556	0.661	1.173	1.479	1.459	2.212

Across all the evaluated models, the expected trend was confirmed: translation time increases with the complexity of the input sentences. There are, however, differences between individual models. The offline Mistral 7B model achieves the shortest average translation times for simple sentences and remains relatively fast even for more complex constructions. At the same time, the differences between languages are relatively stable and do not exhibit marked variability beyond the increase caused by sentence complexity. The Gemma 3-12B model is characterised by balanced behaviour across all languages and levels of complexity, with the increase in translation times being gradual and the differences between languages being minimal. The OpenAI 20B model is the slowest in all categories and exhibits relatively greater variability between languages, particularly for complex sentences.

Overall, it may be concluded that the dominant factor affecting translation time is sentence complexity, whereas the influence of the specific language is secondary, though not negligible.

The time required for back-translation (from the non-English language back into English) was subsequently analysed for the individual AI models. The results are presented in Table 2.

Table 2.

Time required for back-translation according to the individual AI models and sentence complexity

	Simple sentences			Moderately complex sentences			Complex sentences		
	Mistral	Gemma	Open AI	Mistral	Gemma	Open AI	Mistral	Gemma	Open AI
Czech	0.080	0.142	0.779	0.320	0.489	1.074	0.734	1.058	1.604
German	0.083	0.147	0.601	0.293	0.505	1.017	0.663	1.046	1.591
Spanish	0.082	0.140	0.604	0.287	0.493	0.925	0.666	1.057	1.593
French	0.078	0.159	0.731	0.289	0.509	0.958	0.714	1.106	1.625
Chinese	0.081	0.147	0.614	0.328	0.509	0.937	1.425	1.089	1.545
Japanese	0.112	0.129	0.849	0.282	0.489	1.144	0.935	1.038	1.594
Ukrainian	0.078	0.141	0.780	0.293	0.497	1.038	0.737	1.077	1.594
Russian	0.083	0.134	0.778	0.301	0.487	0.970	0.664	1.073	2.127
Average	0.084	0.142	0.717	0.299	0.497	1.008	0.817	1.068	1.659

Across all the offline models, the average translation times again increase as sentence complexity rises. At the same time, more complex inputs lead to a slight increase in differences between languages, though not to marked variability, and extreme values are rare. A notable exception is the back-translation of complex Chinese sentences by Mistral 7B (1.425), which represents the only value in this task exceeding the processing times of the OpenAI-20B model and may reflect the typological distance of Chinese combined with the limited capacity of the low-parameter model. The higher-parameter model, Gemma-3-12B, exhibits stable and balanced behaviour in this task across languages and levels of complexity. Differences between languages remain relatively small for simple and moderately complex sentences and become more apparent only in

the case of complex sentences. The occurrence of extreme maxima is not pronounced in the data. The OpenAI-20B model is also the most time-demanding in back-translation and at the same time exhibits greater variability between languages, especially for more complex sentences."

From the perspective of the aim of the article, the time costs of the translation tasks are of primary importance. Nevertheless, for the sake of completeness, it should be noted that the third language task, namely language identification, is the least time-demanding for all models and is also the least sensitive to sentence complexity. The offline Mistral and Gemma models perform language identification within hundredths of a second. Differences between languages and between sentence categories are minimal, and the time demands increase only very slightly (although the 7B model makes many errors in this task - see below). The model with the highest number of parameters exhibits higher absolute language-identification times than the other two offline models.

The temporal analysis shows that the influence of language is markedly weaker than that of sentence complexity. However, in back-translation, linguistic differences nevertheless become apparent. For simple sentences, the times for the individual languages are very similar, whereas for complex sentences systematic differences begin to emerge. The highest average and maximum back-translation times occur in Russian and Japanese, which may be related to their rich morphology, more flexible word order, and higher degree of syntactic variability. Another interesting finding is that some languages (e.g. Chinese) exhibit a relatively balanced temporal profile even for complex sentences, despite being typologically very distant from English. This suggests that time demands are determined not only by linguistic distance, but also by how consistently a language is represented in the models and by how "predictable" its structures are from the perspective of generation.

For the ChatGPT model, time demands were not measured empirically. Its response as an online service also includes external factors and is therefore not directly comparable with offline models. It is therefore not possible to draw conclusions regarding its relative speed, but only regarding the stability and consistency of its behaviour. In view of the nature of an online service, it may only be assumed that the overall response time, particularly for simple tasks, will not be lower than that of locally operated offline models. However, this assumption cannot be quantitatively verified on the basis of the available data. In this context, translation speed is not the main optimisation priority.

Language Identification

The results in the area of non-English language identification and the comparison of model performance in this area are summarised below. The ChatGPT model and the offline OpenAI 20B model achieved completely error-free performance in language identification. No incorrect identifications were recorded across any of the languages or levels of sentence complexity. Among the higher-capacity offline models, the Google Gemma-3-12B model likewise demonstrated very high accuracy, with only a single error, which was not systematic in nature. These results confirm that language identification is a stable task for more powerful models and is not markedly affected either by linguistic variability or by increasing complexity of the input text.

A different picture emerges in the case of the model with a low number of parameters (Mistral 7B). This model exhibits systematic error rates in language identification. The errors are likely related to the model's capacity and lower number of parameters, which is manifested particularly in languages that are closely related to one another. Table 3 presents the numbers of incorrect identifications for the individual languages and indicates the languages with which they were confused.

Table 3.

Error rates of the low-parameter model (7B) in language identification

Language	Total number of errors	Simple sentences (8)	Moderately complex sentences (16)	Complex sentences (8)	Incorrectly identified languages
Czech	11	0	6	5	9x English, 1x Ukrainian, 1x Swedish
French	2	0	1	1	2 x English
German	8	0	3	5	7 x English, 1x Dutch
Spanish	6	0	3	3	5 x English, 1x Swedish
Russian	6	0	1	5	5 x English, 1x only a transcription of the original sentence (i.e. an error)
Ukrainian	20	3	10	7	12 x Russian, 7 x English, 1x Swedish
Chinese	13	1	5	7	12 x English, 1 x Swedish
Japanese	6	0	4	2	5 x English, 1 x Swedish

The most frequent confusions occur between closely related Slavic languages, especially between Russian and Ukrainian. At the same time, however, it is apparent that the most frequent incorrect identification across languages is the assignment of English, which also occurs in languages with different scripts, such as Chinese or Japanese. With regard to sentence complexity, a consistent trend is confirmed: for simple sentences, the error rate in language identification is very

low, whereas a marked increase occurs in moderately complex sentences, and in complex sentences the error rate is either maintained or increases slightly further. This increase is evident in most of the languages under consideration and suggests that, for this model, language identification is sensitive to the increasing complexity of the input text.

Overall, it may be concluded that language identification represents a task in which the differences between the models are manifested more markedly than in translation tasks. Larger and more powerful models achieve almost perfect accuracy, whereas the 7B model encounters its capacity limits in this area. These limits are manifested not in random errors, but in systematic confusions between languages that are typologically and lexically similar, particularly when processing more complex sentences. The number of parameters and the depth of linguistic representation play a key role, especially in tasks requiring fine-grained discrimination of linguistic signals.

Quality of Back-Translation

The main aim of the analysis is to provide a comprehensive view of the differences between individual AI models in terms of output quality and performance stability in the task of back-translation. The process comprises three steps: (1) translation of sentences from English into a non-English language, (2) detection of the target language, and (3) translation back into English. The key evaluation criterion is the quality of the back-translation, which serves as an indicator of the model's ability to preserve the meaning, linguistic structure, and stylistic adequacy of the text during multi-step processing. It is precisely this phase that best reveals the cumulative errors, shifts in meaning, and stylistic deviations arising in the individual steps of the process. The comparison reflects the differing character of the tested models, in particular the distinction between offline models with local deployment and the online ChatGPT model. It does not seek to establish a ranking of the models, but rather to interpret their strengths and limitations depending on the number of parameters, the specific task, and the context of use.

Table 4 presents the quality of back-translation for the individual models. For each language, the minimum value achieved (i.e. the lowest-rated translation) and the average quality value for all sentences in that language are reported. The combination of these two indicators makes it possible to assess not only the overall level of quality, but also the stability of the model's performance across different inputs.

Table 4.

Quality of back-translations produced by the individual models

	Mistral 7B		Gemma 12B		OpenAI 20B		Chat GPT 5.2	
	Min	Avg	Min	Avg	Min	Avg	Min	Avg
Czech	0.60	0.85	0.60	0.92	0.70	0.94	0.80	0.95
French	0	0.81	0.90	0.97	0.80	0.98	0.80	0.96
German	0.70	0.88	0.70	0.95	0.70	0.94	0.80	0.95
Spanish	0.60	0.91	0.70	0.95	0.80	0.97	0.80	0.95
Russian	0.50	0.91	0.70	0.93	0.60	0.92	0.80	0.95
Ukrainian	0	0.83	0.80	0.96	0.70	0.93	0.80	0.97
Chinese	0.60	0.85	0.80	0.93	0.70	0.93	0.90	0.97
Japanese	0.50	0.86	0.50	0.90	0.60	0.87	0.80	0.95
Average	0.44	0.86	0.71	0.94	0.70	0.94	0.80	0.96

It may be stated that all evaluated models achieved a high level of correspondence between the original English sentences and their back-translated versions. The differences between models appeared primarily in average and minimum values, which indicate the stability and reliability of performance. ChatGPT achieved the highest average translation quality and the highest minimum values, suggesting stable and robust behaviour across languages. The OpenAI 20B model produced results close to ChatGPT, although with slightly greater variability and lower minima. Gemma-3-12B also demonstrated high and relatively balanced back-translation quality, comparable to OpenAI 20B particularly in average values. The 7B model, which had the lowest number of parameters, achieved the lowest quality values and showed the greatest gap between average and minimum results, indicating lower performance stability.

The differences between models therefore do not lie primarily in maximum attainable performance, but in consistency. More powerful models maintained high quality across inputs, while smaller models showed greater variability and more frequent semantic or stylistic deviations. The quantitative results were therefore complemented by qualitative evaluation, which showed that translation errors were not primarily associated with text length or formal complexity, but with narrowly specialised technical terminology. The models often handled long and syntactically complex passages successfully when the vocabulary was familiar or well established. The most problematic texts were those containing highly specific technical terms, such as specialised botanical names, which were sometimes replaced by more general or semantically related but factually incorrect expressions.

These errors generally did not appear as obvious failures or unintelligible outputs, but as systematic and semantically plausible substitutions. The models tended to replace infrequent or highly specialised terms with more general or contextually related expressions, preserving fluency while reducing technical precision. This type of error is especially difficult to detect without expert domain knowledge. Even smaller offline models could produce high-quality outputs, but with greater variability and a higher risk of individual failures. Higher-capacity models were characterised mainly by fewer extreme deviations rather than by a fundamentally different translation logic.

The analysis also shows that languages differ in their level of difficulty for AI models. The most stable and highest-quality results were achieved in French, Spanish, and German, which showed high average correspondence values and relatively high minima. This suggests both strong back-translation quality and low performance variability, possibly due to their proximity to English, similar syntactic structures, and broad representation in model training data. By contrast, typologically more distant languages, especially Japanese, Chinese, and to some extent Russian, produced lower average correspondence values and lower minima. Paraphrasing, sentence restructuring, and occasional shifts in meaning occurred more frequently, although the core semantic content was usually preserved. This indicates not a general failure, but a higher risk of cumulative minor deviations, particularly in complex and technical sentences.

Linguistic difficulty, however, cannot be explained solely by typological distance from English. Some structurally distant languages displayed relatively stable behaviour, while others, especially languages with rich morphology and flexible word order, produced greater variability in quality and processing time. This suggests that the predictability of linguistic structures in training data may be as important as formal linguistic difference.

Overall, the results show that back-translation represents the most sensitive test of the models' ability to preserve meaning. It is at this stage that differences between model approaches become most visible. While offline models may provide good and fast performance together with the advantage of independent deployment, GPT-based models clearly dominate in the quality, stability, and linguistic balance of outputs.

5. Conclusions

The article assessed the translation capabilities of selected artificial intelligence language models through a chained multilingual procedure combining translation from English into a target language, language identification, and back-translation into English. This design enabled the evaluation not only of translation quality, but also of model stability and meaning preservation across successive language-processing tasks. Such a perspective is particularly relevant for defence and security contexts, where multilingual processing often forms part of interconnected analytical workflows rather than isolated translation acts.

The findings broadly correspond with existing research indicating that contemporary large language models can achieve strong translation performance, especially in high-resource languages, while their reliability remains shaped by model scale, language pair, and task design [3, 4, 19]. All evaluated models performed well in basic translation and generally produced semantically accurate and comprehensible outputs. The main differences appeared in the consistency with which they preserved meaning, style, terminology, and linguistic naturalness across more complex inputs and chained operations. From a defence and intelligence perspective, this is significant because even minor semantic shifts may affect the interpretation of actors, actions, intent, context, location, or technical content. The most visible differences emerged in language identification and back-translation: higher-capacity models achieved almost error-free language identification, whereas the lower-parameter model showed more errors, particularly with related languages and complex sentences. Back-translation proved the most sensitive stage, revealing cumulative semantic and structural deviations that were less visible in the initial translation, consistent with research warning that fluent multilingual output may conceal subtle semantic deviations or hallucinations in high-stakes contexts [9].

The results also indicate variation in language-specific difficulty. French, Spanish, and German appeared relatively less demanding; Czech and Ukrainian occupied an intermediate position, performing comparably to the Romance languages in higher-capacity models but showing greater variability in the low-parameter model. Japanese, Russian, and Chinese produced greater variability, especially in complex constructions. These differences, however, cannot be explained solely by typological distance from English, as model performance also appears to be influenced by training data, structural predictability, morphology, and sentence complexity.

Overall, the study suggests that differences between models are limited in baseline translation performance but become operationally significant when stability, input complexity, and chained-task reliability are considered. Model selection should therefore be understood as a trade-off between speed, stability, translation quality, deployment autonomy, data protection, and technical constraints. Offline models may be useful where independent deployment and data control are priorities, whereas more advanced GPT-based systems appear better suited to linguistically complex or security-sensitive tasks requiring stronger output stability.

The study has several limitations. The dataset comprised only 32 English inputs across eight target languages and therefore provides an indicative comparison rather than a comprehensive benchmark. The analysis focused primarily on sentence- and short-text-level processing, while longer documents, noisy social media content, speech-derived transcripts, and specialised intelligence reporting may produce different results. The evaluation also relied on manual assessment, which

enabled qualitative sensitivity but introduced a degree of evaluator dependence. One limitation concerns the use of an evaluated model as a partial evaluator of its own outputs, which may introduce bias into the qualitative assessment. Finally, the online GPT-based model was not directly comparable with offline models in terms of response time.

Despite these limitations, the study suggests that AI translation can already support multilingual screening, analytical preparation, and selected defence and security tasks. In high-stakes settings, however, its use should remain subject to validation, human oversight, terminology control, and context-sensitive evaluation. AI translation should therefore be understood not as a replacement for expert judgement, but as a decision-support capability whose value depends on responsible integration into secure and professionally supervised workflows, consistent with broader defence-oriented discussions on responsible AI adoption and human oversight [17, 18].

Acknowledgements. The authors acknowledge support from the Ministry of Defence of the Czech Republic and the University of Defence, Faculty of Military Leadership, under the project 'LANDOPS'.

References

1. **Christianson C., Duncan J., Onyshkevych B.** Overview of the DARPA LORELEI Program. Machine Translation, 2018, 32 (1-2), p. 3-9.
2. **Teze V., Nazaruka E.** Future Directions in Defence NLP: Investigating Research Gaps for Low-Resource Languages. In Digital Business and Intelligent Systems; Lupeikienė A., Ralyté J., Dzemyda G., Eds.; Springer: Cham, 2024; Communications in Computer and Information Science, Vol. 2157, pp. 93-105.
3. **Zhu W., Liu H., Dong Q., Xu J., Huang S., Kong L., Chen J., Li L.** Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In Findings of the Association for Computational Linguistics: NAACL 2024; Association for Computational Linguistics: Mexico City, Mexico, 2024; pp. 2765-2781.
4. **Robinson N., Ogayo P., Mortensen D.R., Neubig G.** ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In Proceedings of the Eighth Conference on Machine Translation; Association for Computational Linguistics: Singapore, 2023; pp. 392-418.
5. **Peng K., Ding L., Zhong Q., Shen L., Liu X., Zhang M., Ouyang Y., Tao D.** Towards Making the Most of ChatGPT for Machine Translation. In Findings of the Association for Computational Linguistics: EMNLP 2023; Association for Computational Linguistics: Singapore, 2023; pp. 5622-5633.
6. **Liu X., Tang L., Ma X., Hu J.** Research on Military Text Machine Translation Based on Deep Neural Network. In Advances in Intelligent Automation and Soft Computing; Li X., Ed.; Springer: Cham, 2021; Lecture Notes on Data Engineering and Communications Technologies, Vol. 80, pp. 361-368.
7. K**hiu E., Toossi H., Anugraha D., Liu J., Li J., Flores J., Roman L., Doğruöz A.S., Lee E.-S.** Predicting Machine Translation Performance on Low-Resource Languages: The Role of Domain Similarity. In Findings of the Association for Computational Linguistics: EACL 2024; Association for Computational Linguistics, 2024; pp. 1474-1486.
8. **Wang L., Lyu C., Ji T., Zhang Z., Yu D., Shi S., Tu Z.** Document-Level Machine Translation with Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Singapore, 2023; pp. 16646-16661.
9. **Guerreiro N.M., Alves D.M., Waldendorf J., Haddow B., Birch A., Colombo P., Martins A.F.T.** Hallucinations in Large Multilingual Translation Models. Transactions of the Association for Computational Linguistics, 2023, 11, p. 1500-1517.
10. **Papineni K., Roukos S., Ward T., Zhu W.-J.** Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Philadelphia, Pennsylvania, USA, 2002; pp. 311-318.
11. **Rei R., Stewart C., Farinha A.C., Lavie A.** COMET: A Neural Framework for MT Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Online, 2020; pp. 2685-2702.
12. **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, 2021, 9, p. 1460-1474.
13. **Kocmi T., Federmann C., Grundkiewicz R., Junczys-Dowmunt M., Matsushita H., 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; Association for Computational Linguistics: Online, 2021; pp. 478-494.

14. **Qian S., Sindhujan A., Kabra M., Kanojia D., Orasan C., Ranasinghe T., Blain F.** What do Large Language Models Need for Machine Translation Evaluation? In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Miami, Florida, USA, 2024; pp. 3660-3674.
15. **Goyal N., Gao C., Chaudhary V., Chen P.-J., Wenzek G., Ju D., Krishnan S., Ranzato M.-A., Guzmán F., Fan A.** The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. Transactions of the Association for Computational Linguistics, 2022, 10, p. 522-538.
16. **NLLB Team.** Scaling neural machine translation to 200 languages. Nature, 2024, 630, p. 841-846.
17. **Nitzl C., Cyran A., Krstanovic S., Borghoff U.M.** The use of artificial intelligence in military intelligence: an experimental investigation of added value in the analysis process. Frontiers in Human Dynamics, 2025, 7, 1540450.
18. **North Atlantic Treaty Organization.** Summary of NATO's Revised Artificial Intelligence (AI) Strategy. Available online: https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/summary-of-natos-revised-artificial-intelligence-ai-strategy (accessed on 23 April 2026).
19. **Hendy A., Abdelrehim M., Sharaf A., Raunak V., Gabr M., Matsushita H., Kim Y.J., Afify M., Awadalla H.H.** How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv, 2023, arXiv:2302.09210.
20. **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. arXiv, 2023, arXiv:2301.08745.