Fuzzy Matching and Sentence Embeddings for Few-shot Machine Translation with Large Language Models

Miguel Rios, Claudia Plieseis, Dragoș Ciobanu, Alina Secară
Centre for Translation Studies, University of Vienna, Austria
{miguel.angel.rios.gaona, claudia.karin.plieseis
dragos.ioan.ciobanu, alina.secara}@univie.ac.at

Abstract

In-context learning is a method for improving machine translation in Large Language Models, but its performance is sensitive to the quality of the few-shot example selection. Current retrieval strategies use semantic similarity by computing sentence embeddings, and these methods often require significant computational overhead and specialised expertise. We evaluate the impact of retrieval strategies on translation performance in a specialised domain, comparing traditional, token-based fuzzy matching against semantic sentence embeddings. We use a medical corpus from the European Medicines Agency (EMEA) for the English-Romanian and English-German language pairs, and we evaluate translation quality with automatic metrics and manual evaluation. Our results show that 1-shot and 5-shot prompting significantly outperforms the 0-shot baselines for quality in automatic evaluations for both language pairs, and in manual evaluation for English-German. For the English-Romanian pair, the average scores of the manual evaluation for both quality and ranking follow the same trend, but statistical significance is not consistently reached for all few-shot prompting configurations. In general, token-based fuzzy matching overwhelmingly has higher automatic quality scores than embedding-based retrieval.

1 Introduction

In-context learning (ICL) serves as a robust method for enhancing the performance of Large Language Models (LLMs) for machine translation (MT), particularly in few-shot scenarios (Moslem et al., 2023; Bawden and Yvon, 2023). However, the performance of few-shot prompting depends on the selection of high-quality examples. Current retrieval strategies rely on computing the semantic similarity between sentence representations (Zebaze et al., 2025). Furthermore, these embedding-based approaches introduce challenges for translation practitioners, such as the required computational resources, sufficiently large local translation memories, and specialised expertise.

This study compares the integration of token-based fuzzy matching against selecting the closest semantic sentence embedding(s) for 1-shot and 5-shot examples selection. Both types of matching - token-based fuzzy matching (FM) and embedding-based matching (EM) - are done by comparing the string of tokens (or the embedding, respectively) of an English segment (English is the source language in our study) to all other English strings (or their corresponding embeddings, which we derive) which have a Romanian or a German translation in the European Medicines Agency (EMEA) translation memory (ELG, 2020). Thus we address the following research question: How do different retrieval strategies for few-shot examples impact the quality of LLM MT output in a specialised domain? Our domain is the medical domain, our source language is English, and our two target languages are Romanian and German. We evaluate the translation quality of both retrieval strategies using a combination of automatic metrics and human preference rankings and quality assessment of the resulting

MT outputs.

The 1-shot and 5-shot prompting significantly outperforms the 0-shot baseline LLMs according to BLEU and COMET scores in both language pairs. This also applies to English-German human evaluation, where 5-shot EM prompting generates the highest-scoring output. For English-Romanian, only the 5-shot FM prompting performs significantly better than 0-shot.

2 Background and Related Work

To produce MT outputs, an LLM is provided with a prompt specifying the task to be completed (i.e., translate), the source and target language(s) requested, followed by the source segment to be translated. ICL enables LLMs to perform specialised translation tasks without the need for fine-tuning (Brown et al., 2020). A prompt with a series of source-target examples retrieved from a translation memory referred to as few-shot demonstrations/examples influences the model generation for patterns and stylistic nuances relevant for MT (Bawden and Yvon, 2023; Moslem et al., 2023). Recent studies have shown that retrieval-based selection for translation examples based on semantic or syntactic similarity leads to better translation performance (Vilar et al., 2023; Zebaze et al., 2025). In contrast, 0-shot prompting involves translating without any translation examples in the prompt.

Dense Retrieval leverages sentence embeddings from models such as LASER (Artetxe and Schwenk, 2019) to map source segments and a selection pool (e.g., a translation memory) into a shared vector space. This allows for the retrieval of examples that are semantically similar to the source input (Vilar et al., 2023; Zebaze et al., 2025). Another model, SONAR (Duquenne et al., 2023), introduces a fixed-size sentence embedding representation designed for both multilingual and multimodal applications (e.g., MT and speech recognition). With a single text encoder, SONAR supports 200 languages, and outperforms LASER (Artetxe and Schwenk, 2019) and LabSE (Feng et al., 2022) in multilingual similarity search benchmarks.

Fuzzy-match Retrieval uses information retrieval metrics (e.g., BM25, fuzzy-match) to find lexical overlaps from examples from the selection pool with the source segment (Xu et al., 2020; Zebaze et al., 2025). For example, the SYSTRAN fuzzy-match library filters candidates via suffix arrays and ranks them using edit distance. This tokenization approach introduces a bias toward languages with low morphological complexity. On the other hand, languages with extensive morphology retrieve fewer matches because their word forms are more diverse (Nieminen et al., 2025).

Xu et al. (2020) present data augmentation techniques for Neural Machine Translation (NMT) that leverage similar translation pairs, which resembles how human translators use fuzzy matches. Moreover, they extend this framework to include semantically related translations retrieved based on sentence embeddings. Fuzzy matches provide the NMT model with explicit lexical information, whereas embedding-based similarities extend the translation context.

Bawden and Yvon (2023) show that LLMs frequently underperform compared to NMT in low-resource and highly specialised contexts, such as the medical domain. Moreover, few-shot prompting addresses over-generation in LLMs compared to 0-shot prompting. Moslem et al. (2023) use ICL to refine the generated translation based on provided examples. In particular, ICL improves MT performance in the medical domain (i.e., COVID). Vilar et al. (2023) show that the quality and selection of examples (e.g., using k-Nearest Neighbours to find similar sentences) are more important than the quantity of examples. Nieminen et al. (2025) introduce an embedding-based variant of Retrieval-Augmented Translation (RAT) for NMT. Neural Fuzzy Repair (NFR) enhances the input by appending the target side of similar source examples retrieved from a translation memory. Zebaze et al. (2025) perform a systematic study evaluating various LLMs and retrieval strategies for ICL. In contrast to previous work, their results show that sentence embedding similarity significantly improves MT performance, particularly for low-resource languages. Furthermore, they analyse the trade-off between the diversity of the selection pool and the quality of the retrieved examples.

We follow an experimental design similar to (Xu et al., 2020; Zebaze et al., 2025). However, our approach focuses on a specialised domain, fuzzy-match retrieval instead of BM25, and SONAR for dense retrieval. Moreover, we also perform a manual evaluation based on quality (accuracy and fluency) and perceived editing effort scoring, as well as ranking of the generated MT outputs.

3 Experimental Setup

3.1 Data

The EMEA corpus consists of automatically-aligned PDF documents from the European Medicines Agency (ELG, 2020). We use the English-Romanian (783,742 segments) and English-German (760,574 segments) sections from the EMEA parallel corpus. We use the complete EMEA corpus as a candidate pool for extracting few-shot examples, thus creating a test set with the first 2,000 segments that have at least 5 fuzzy matches. We use the corresponding Romanian and German target segments of these 2,000 segments as the reference translations in the subsequent evaluation with both automatic metrics and manual analysis.

We perform heuristic filtering[1] for each language pair to eliminate noise. Specifically, the tool discarded duplicates, source copied segments, sentences exceeding a 200-word limit, pairs with a length ratio greater than 2:1, and HTML tags.

3.2 Models

We use the HuggingFace transformers framework for the LLMs (Wolf et al., 2020). We selected open-source LLMs with fewer than 20B parameters to allow future accessibility for translators with limited computational resources. Our LLMs are as follows:

Tower+ closes the gap between translation performance and general multilingual reasoning by introducing enhanced training methods (Rei et al., 2025). The method integrates supervised fine-tuning (SFT), and preference optimisation for translation tasks. Baseline: Tower-Plus-9B[2].

EuroLLM introduces open-source LLMs for comprehensive text generation and multilingual coverage (Martins et al., 2025). The collection covers all 24 official European Union languages, alongside 11 additional languages (e.g., Chinese, Hindi, and Korean). Baseline: EuroLLM-9B-Instruct[3].

Qwen3 is a series of LLMs designed for model scale and inference efficiency (Yang et al., 2025). Qwen3 improves over previous Qwen models on both multilingual task performance and task-specific performance (e.g., translation, and question answering). Qwen uses a thinking mode to produce detailed explanations for reasoning tasks (e.g., math, code). We use the non-thinking mode of Qwen3 to avoid overgeneration. Baseline: Qwen3-8B[4].

Mistral is a series of parameter-efficient and dense LLMs for deployment in compute-constrained environments. Beyond base LLMs, the series includes instruction-tuned variants and specialised reasoning models for complex tasks (Liu et al., 2026). Baseline: Ministral-8B-Instruct-2410[5].

We use the default generation configuration for the LLMs: temperature 1.0, top-p 0.9, top-k 0, and maximum number of generated tokens after the prompt 2048.

3.3 Few-shot Prompting

We use the complete EMEA corpus as a selection pool for both retrieval strategies.

Fuzzy-match We use the fuzzy-match[6] library to index and query the candidate pool. For each source segment in the test set, we retrieve the highest-scoring fuzzy matches from the candidate pool. Our configuration uses a threshold of 0.7, and retrieves a maximum of five examples per segment.

Embedding Match We use SONAR sentence embeddings to generate our text representations. For each source segment in the test set, we compute the cosine similarity against the candidate pool and retrieve the five highest-ranking examples.

The prompt template for the few-shot configuration is defined as follows:

```
Translate the {src} source text to {tgt}.
Rules:
Output strictly the {tgt} translation.
Do NOT repeat the {src} text.
Do NOT include labels like "{src}:" or "{tgt}:".
Do NOT add quotation marks, explanations, or any extra text.
Examples:
{src}: {example-segment}
{tgt}: {example-segment}
...
{src}: {example-segment}
{tgt}: {example-segment}
Now translate the following:
{src}: {src-segment}
{tgt}:
```

[1] https://github.com/ymoslem/MT-Preparation
[2] Unbabel/Tower-Plus-9B
[3] utter-project/EuroLLM-9B-Instruct
[4] Qwen/Qwen3-8B
[5] mistralai/Ministral-8B-Instruct-2410
[6] https://github.com/SYSTRAN/fuzzy-match

	0-shot	1-shot FM	1-shot EM	5-shot FM	5-shot EM
BLEU↑					
Tower+	45.39	69.29	68.94	71.22	71.82
EuroLLM	40.50	65.66*	64.55	66.70*	59.82
Qwen3	29.27	63.22*	61.75	67.57	67.16
Mistral	24.19	66.33*	64.51	69.33	69.12
COMET↑					
Tower+	0.906	0.927	0.929	0.932	0.933
EuroLLM	0.901	0.923	0.923	0.928	0.922
Qwen3	0.863	0.922	0.922	0.931*	0.930
Mistral	0.804	0.916*	0.910	0.927	0.927

Table 1: Automated evaluation scores for English-Romanian with Fuzzy Match (FM) and Embedding Match (EM). Asterisks (*) denote statistical significance ($p < 0.05$) for comparisons between the FM and EM prompts.

	0-shot	1-shot FM	1-shot EM	5-shot FM	5-shot EM
BLEU↑					
Tower+	42.32	72.70	72.11	78.32*	77.04
EuroLLM	44.93	74.64*	73.41	75.66*	74.21
Qwen3	35.63	73.12*	71.15	78.92*	78.01
Mistral	34.52	72.67*	70.77	77.40	76.79
COMET↑					
Tower+	0.867	0.903	0.904	0.914	0.912
EuroLLM	0.866	0.902	0.901	0.906*	0.897
Qwen3	0.840	0.901	0.901	0.912	0.911
Mistral	0.830	0.895	0.894	0.908	0.907

Table 2: Automated evaluation scores for English-German with Fuzzy Match (FM) and Embedding Match (EM). Asterisks (*) denote statistical significance ($p < 0.05$) for comparisons between the FM and EM prompts.

In this the few-shot template, the *src* is the source language (*English*), *tgt* is the target language (*Romanian* or *German*), *example-segment* is the source or target retrieved examples, and *src-segment* is the source test segment. The 0-shot template is similar to the few-shot, except that there is no examples section. We use 1-shot and 5-shot prompting configurations for both retrieval methods.

To replicate our experiments, we release our scripts for few-shot LLM prompting, and a Docker container with the fuzzy-match library installed on GitHub[7].

4 Automatic Metrics Evaluation

We use BLEU (Papineni et al., 2002; Post, 2018), and COMET (Rei et al., 2020) for automatic evaluation. For evaluation, we establish a 0-shot baseline and compare it against 1-shot and 5-shot prompting configurations. We define fuzzy match retrieval as (**FM**), and embedding match retrieval as (**EM**) for each prompting configuration.

Table 1 shows the results for English-Romanian. The asterisks denote statistical significance ($p < 0.05$) between the FM and EM 1-shot and 5-shot prompting for both automatic metrics based on bootstrap resampling (Post, 2018). Tower+ achieves the highest performance in the 0-shot setting according to BLEU scores. For the 1-shot setting, EuroLLM, Qwen3, and Mistral demonstrate statistically significant improvements in BLEU with the FM setting, while Tower+ still obtains the highest BLEU score. Interestingly, BLEU scores for 5-shot are better than for 1-shot, which in turn are better than for 0-shot for Tower+, Qwen3, and Mistral, but EuroLLM shows a significant decrease in BLEU with the 5-shot EM setting. Regarding the COMET metric, Tower+ continues to obtain the

[7] https://github.com/HAITrans-lab/fuzzy-match_MT_LLM

highest scores, but only Qwen3 and Mistral show significant improvements with the FM prompting.

Table 2 shows the results for English-German. EuroLLM achieves the highest performance in the 0-shot and 1-shot settings measured by BLEU, while Qwen3 outperforms with the 5-shot setting. Consistent with the English-Romanian results, the 1-shot FM produces significant BLEU score improvements for EuroLLM, Qwen3, and Mistral. For the 5-shot FM, Tower+, EuroLLM, and Qwen3 show significant gains in BLEU. Finally, according to the COMET scores, Tower+ narrowly outperforms with all categories, and only EuroLLM shows a significant improvement within the 5-shot FM setting.

4.1 Few-shot Examples and MT Output Analysis

We use the Levenshtein distance to compare the English-Romanian, as well as the English-German, examples (i.e. pairs of source and target segments) retrieved with FM and EM to ensure that the retrieved examples are different. The Levenshtein distance quantifies the difference between two sequences by computing the minimum cost of transforming one into the other based on characters (a lower score denotes similar sequences ↓).

For English-Romanian, the identical 1-shot examples between FM and EM (0 Levenshtein distance) represent 24.15%, while the identical 5-shot examples between FM and EM (0 Levenshtein distance) represent 0.85%. For English-German, the identical 1-shot examples between FM and EM (0 Levenshtein distance) represent 26.35%. The identical 5-shot examples between FM and EM (0 Levenshtein distance) represent 0.75%.

Table 3 shows the average Levenshtein distance across 1-shot and 5-shot examples for English-Romanian. For the 1-shot, the examples retrieved for the EM and FM configurations show a high character-level overlap (29.85). In contrast, the 5-shot examples show a lower character-level overlap (190.743) across them.

	1-shot FM	1-shot EM	5-shot FM	5-shot EM
1-shot FM	-	28.979	737.223	731.889
1-shot EM		-	742.268	729.221
5-shot FM			-	190.743
5-shot EM				-

Table 3: Average Levenshtein distance↓ across 1-shot and 5-shot examples for English-Romanian.

Table 4 shows the average Levenshtein distance across 1-shot and 5-shot examples for English-German. The 1-shot examples have a high character-level overlap, whereas the 5-shot introduces variability, resulting in a lower overlap between the retrieved examples.

We also use the Levenshtein distance to compare the MT outputs (monolingual target language segments) for Tower+ for both Romanian and German. For English-Romanian, the amount of identical (0 Levenshtein distance) MT outputs obtained with 0-shot, 1-shot, and 5-shot prompting is 4.7%. For English-German, the amount of identical (0 Levenshtein distance) MT outputs is 8.95%.

	1-shot FM	1-shot EM	5-shot FM	5-shot EM
1-shot FM	-	29.85	760.038	755.193
1-shot EM		-	765.352	752.955
5-shot FM			-	197.658
5-shot EM				-

Table 4: Average Levenshtein distance↓ across few-shot examples for English-German.

5 Manual Evaluation

We selected Tower+ for further evaluation because of its competitive 0-shot baseline. This allows us for a more controlled comparison of the performance differences from the few-shot configurations. We report on the outcome of a manual evaluation task conducted by two Romanian native speakers with translation experience, but no specialist medical knowledge (working with English-Romanian collaboratively (Esperança-Rodier et al., 2019)) and one German native speaker (working with English-German) who also has translation experience, but not in the medical domain. All three collaborated before and during the evaluation to increase agreement and minimise personal preference. The evaluators worked on two samples of 100 segments (one for each language pair) which contained a balanced number of short segments (up to 100 characters), medium segments (101-250), and long segments (over 251 characters). For each segment, the evaluators were presented with the source, a reference translation, and five anonymised and randomised MT system outputs (the 0-shot output, the two 1-shot outputs with FM and EM respectively, and the two 5-shot outputs with FM and EM respectively). The evaluators were asked to rate the MT outputs on two tasks, namely system ranking (Vela and van Genabith, 2015) and quality. The quality scoring is informed by the accuracy and fluency errors encountered in the MT output and their severity, as

well as the amount of overall perceived editing effort required to correct the MT output fully. We did not ask for specific errors to be annotated by category with a typology such as the MQM (Lommel et al., 2014), but our quality scoring method is inspired by the Error Span Annotation (ESA) (Kocmi et al., 2024). To orient our evaluators away from strong individual preferences, we use a 0–100 scale broken down into discrete spans, as follows:

- 0% no meaning
- $\geq$ 33% some meaning preserved and major editing needed
- $\geq$ 65% most meaning preserved and some editing needed
- 100% perfect.

To further guide other evaluators in the process, we offer guidelines containing examples of what might constitute minor and major errors, together with suggestions regarding point deductions. We include our guidelines in Appendix A.

For the system ranking, each system output is ranked from best to worst (1 to 5), with ties allowed. When these occur, the next level rank is used for the remaining segments. For example, in a situation with three MT hypotheses ranked as best, followed by the remaining two in descending order, the resulting ranking is 1,1,1,2,3.

Unlike the monolingual Direct Assessment (DA) (Graham et al., 2015) used within WMT, we consider it crucial for evaluators to also understand and have access to the source segments. This is due not only to the professional duty to acknowledge that a translation-related task should be bilingual, but also to the variable quality of the reference translations included in testing and training data sets. Gaona et al. (2023) identify and list omissions, additions, and mistranslations as important issues related to the reference translations provided in the Medline corpus used in that study, and we also noticed several examples of misalignments in our 2,000-segment test set extracted from the EMEA corpus. Because of this, when ranking and scoring the MT outputs, our evaluators were instructed to not only use the reference translations provided and the guidelines, but also their professional translation experience, as well as further research in relevant external resources.

5.1 Quality Scores and Ranking

The average quality scores and ranking in Table 5 for English-Romanian put the 5-shot FM prompting top, and the 0-shot prompting bottom (Table 5). Paired t-tests reach significance between 5-shot FM and 0-shot for both ranking ($p =$ **0.003**) and quality ($p =$ **0.0013**). One-way ANOVA p-values for all five prompting configurations reach significance for ranking ($p =$ **0.027**) and no significance for quality ($p = 0.144$).

Tower+	Quality↑	Ranking↓
0-shot	86.83	2.05
1-shot FM	89.16	1.84
1-shot EM	88.30	1.85
5-shot FM	**91.34**	**1.69**
5-shot EM	90.03	1.71

Table 5: Average quality scores and ranking for English-Romanian from Tower+.

Tower+	Quality↑	Ranking↓
0-shot	84.13	1.90
1-shot FM	93.88	1.40
1-shot EM	91.99	1.49
5-shot FM	95.40	1.30
5-shot EM	**95.45**	**1.23**

Table 6: Average quality scores and ranking for English-German from Tower+.

Table 6 shows the average quality scores and ranking for English-German. The 5-shot EM prompting has the highest quality score, and 0-shot prompting the lowest ranking. Paired t-tests reach significance for ranking: 5-shot EM with 1-shot EM ($p =$ **0.001**), 5-shot EM with 1-shot FM ($p =$ **0.03**), and 0-shot across all prompting ($p < 0.05$). For quality the paired t-tests show significance with 0-shot across all prompting ($p < 0.05$). One-way ANOVA p-values for all five prompting configurations reach significance for ranking ($p =$ **5.5e-11**) and quality ($p =$ **5.8e-07**).

6 Discussion

The manual evaluation tasks put the 5-shot FM prompting top, and the 0-shot prompting bottom for English-Romanian. For English-German, 5-shot EM was best and 0-shot was worst. The averages listed in Tables 5 and 6 support the automatic evaluation scores from Tables 1 and 2, suggesting that

Example: few-shot	Source English	Hypothesis Romanian	Issue
1: 5-shot EM	can be submitted	trebuie depuse	obligation *trebuie (must)* instead of possibility *pot fi*
2: 5-shot FM 5-shot EM	ACR 20 response	răspuns ACR 20 (ameliorare clinică cu 20% în funcţie de criteriile Colegiului American de Reumatologie)	explicitation of the acronym added
3: 5-shot EM 5-shot FM	lactation	alăptare	the term *alăptare* is used for humans instead of *lactaţie*, which is used for animals
4: 1-shot FM	effects of linagliptin	efecte ale linagliptinului	wrong masculine agreement instead of feminine agreement *efectele linagliptinei*
5: 5-shot EM	effects of linagliptin	efectele linagliptin	incorrect agreement
6: 1-shot EM 5-shot FM	talk to your doctor	adresați-vă medicului dumneavoastră	superfluous possessive adjective *dumneavoastră*

Table 7: Examples of translations errors in few-shot MT outputs English-Romanian.

prompting with examples brings an improvement in every few-shot prompting configuration when compared to 0-shot for both English-Romanian and English-German. In several instances, this improvement is also statistically significant.

Notwithstanding the improvements recorded in both the automatic and manual evaluation scores, various issues still remain in the few-shot MT outputs. Using previous studies, Muñoz-Miquel (2025) lists aspects which make the use of MT in medical translation challenging. We use these to structure the discussion of issues identified in the few-shot MT outputs during our manual evaluation and provide specific examples in Tables 7 and 8.

High specialisation of texts The use of modal verbs is particular in medical texts, and their inaccurate translation can lead to serious consequences. In the MT output (Table 7 **example 1**) a modal verb is erroneously translated with a stronger form. In the 5-shot EM the wrong translation of *can* with *trebuie* (*must*), denoting an obligation, is likely a consequence of *trebuie* being used in four of the five fuzzy examples included in the prompt for that segment.

Given that medical texts require a high degree of precision, even meaning changes that are not as severe as the one in the previous example can have an impact. For English-German (Table 8 **example 1**), the target text is inappropriately more specific than the source text. The overtranslation of *of this* as *dieser Befunde* (*of this evidence*) likely occurred due to the wrong translation being present in the example used for the 1-shot EM prompt. Similarly, the addition of *weitere (further)*, which was again present in the 1-shot EM, results in a slight meaning change that implies that the patient has received medical treatment in the past (Table 8 **example 2**).

Need for terminological precision In a highly specialised domain, the use of correct terminology is crucial. Tower+ generally performed well by learning from the examples provided in the prompts. An example of this can be found in English-German when comparing the output of the 0-shot to the output of all other conditions (1-shot FM, 5-shot FM, 1-shot EM, 5-shot EM). The term *potassium-sparing diuretics* was translated correctly as *kaliumsparende Diuretika* in all conditions except for 0-shot, where it was translated as *kaliumspendende Diuretika* (*potassium-providing diuretics*). Since the term was translated correctly in all examples, Tower+ was able to maintain the correct term.

However, when the examples were faulty, Tower+ failed to maintain the correct terminology. A term used for humans *alăptare* (*breastfeeding*) was erroneously employed in a segment about animals in the 5-shot EM and FM hypotheses instead of the correct *lactaţie* (*lactation*) (Table 7 **example 3**), as the correct terms did not appear in any examples used in the 5-shot EM prompting and only in two for the 5-shot FM.

Similarly, in the 1-shot EM prompt in **example 3** (Table 8) the term *monohydrate* was omitted from *lactose monohydrate,* which resulted in erroneous output. Given that *lactose* and *lactose monohydrate* have different water contents, melting points and physical forms, this is an imprecise use of terminology.

Wide variety of audiences Some medical texts will address medical experts, while others will be

Example: few-shot	Source English	Hypothesis German	Issue
1: 1-shot EM	The clinical significance of this is unknown.	Die klinische Bedeutung dieser Befunde ist unbekannt.	The translation of *of this* as *dieser Befunde* is an over-translation and a potential mistranslation depending on the context
2: 1-shot EM	you may need urgent medical treatment	Sie könnten dringend weitere medizinische Behandlungen benötigen	Addition of *weitere* implies that the patient has received treatment in the past
3: 1-shot EM	Also contains lactose monohydrate.	Enthält Lactose.	Omission of *monohydrate*
4: 1-shot EM 1-shot FM 5-shot FM	talk to your doctor or nurse	sprechen Sie mit Ihrem Arzt oder Ihrer Krankenschwester	*nurse* was translated in the feminine form even though the gender-neutral form would be preferable
5: 5-shot FM	NovoRapid FlexTouch	Levemir FlexTouch	Wrong drug name
6: 1-shot EM	in the authorised use	in den genehmigten Anwendungsbereichen	*genehmigt* is a valid translation of *authorised*, but in this context, it is weaker than *zugelassen*
7: 1-shot FM	Your doctor will monitor you for this.	Ihr Arzt wird Sie darauf überwachen.	*darauf überwachen* is a literal translation from English and too strong for this context; a more neutral construction with *beobachten*/*im Auge behalten* would be preferable

Table 8: Examples of translations errors in few-shot MT outputs English-German.

aimed at patients or the general public. The EMEA corpus reflects this diversity. For the same EN segment, the examples to be included in prompts can each address a different target audience, and therefore display different translation strategies. For example, one translation can use a direct form of address, 2nd person plural, formal *Evitați administrarea la animalele deshidratate* (*Avoid giving to dehydrated animals*), while another has a neutral form of address using the verb in the infinitive *A se evita administrarea la animalele deshidratate* (*Not to be given to dehydrated animals*). Moreover, translations for the segment *Patients should not lie down* can include specialised terms if aimed at medical practitioners *Pacientele nu trebuie să stea în clinostatism* (*Patients should not be in the supine position (clinostatism)*), or lay terminology if aimed at non-specialist patients *Pacienții nu trebuie să stea la orizontală* (*Patients must not lie down*).

All variants are viable depending on the intended style and audience, but their mixed presence in the prompting examples will lead to inconsistent choices made by the LLM. Conversely, providing the LLM with enough examples that address the intended target audience may lead to more appropriate results. For example, in a case where the 1-shot and 5-shot (both FM and EM) prompts translated *dysphagia* as *Dysphagie*, the output contained the specialised term. In the 0-shot condition, however, it was translated as *Schluckbeschwerden* (*difficulty swallowing*), thus addressing a lay audience.

Cultural asymmetries These can appear from differences in health systems or in how one society uses language in a domain. Gender-inclusive language is rarely used in Romanian and therefore the EMEA corpus has limited examples of this. However, gender-inclusive language did appear in some MT outputs. When *partenerului/-partenerei dumneavoastră* (*"your partner" in English which does not have grammatical gender, like Romanian and German do*) appears in three out of the five prompt examples, the 5-shot FM offers this gender-inclusive translation for *your partner*. The 5-shot EM offers only the standard masculine form *partenerului dumneavoastră*, even if the gender-inclusive form appears in two out of its five prompt examples.

In German, different strategies for using gender-inclusive language exist and are becoming increasingly widespread. In recent years, terms like *nurse* are often translated into a gender-neutral form such as *medizinisches Fachpersonal* (*medi-*

cal professionals). That said, the EMEA corpus likely contains numerous examples that use the term *Krankenschwester*, which exclusively refers to female nurses. This bias also appears in MT output, as shown in Table 8 (**example 4**). In this example, the 1-shot (EM and FM) and 5-shot prompts (FM) did little to alleviate this bias. In the 1-shot EM and 1-shot FM, *doctor or pharmacist* was used instead of *doctor or nurse*. In the 5-shot FM, the term *nurse* was translated into the gender-neutral form in two out of five of the examples (while the other three mentioned *pharmacist* instead of *nurse*). Still, Tower+ was unable to adapt to the gender-neutral version.

Poor wording of some original texts While the corpus contains plenty of source segments which are incomplete due to poor segmentation, we made sure that the 100 source segments selected for manual evaluation were correct. However, the biggest problem we faced was the quality of reference translations. These were in many cases faulty, longer or shorter than the source segment, or completely misaligned. Sometimes, the outputs offered by the MT systems for a specific phrase (*efficacious in cattle*) are better (*eficace la bovine*) than the reference (*eficace la viţei (efficatious in calves)*). In addition, the quality of the translations included in the corpus and then used as examples in the prompts was also not perfect. For example, agreement of the name of medicines was frequently problematic. As exemplified in Table 7 (**examples 4 and 5**), the right translation using the indefinite article female agreement (*Efectele linagliptinei*) was never offered, and instead either the wrong indefinite article masculine agreement (also present in the example for the 1-shot FM), or a non-inflected form (also present in three out of the five examples from the 5-shot EM) were given.

Similar problems in relation to the examples used in the prompts can be observed in English-German. In **example 5** (Table 8), each example of the 5-shot FM prompt contained a different product name (*Levemir FlexTouch*; *NovoRapid*; *NovoRapid Penfill*; *NovoRapid InnoLet*; *NovoRapid FlexPen*) which matched part of the name in the source segment to be translated (*NovoRapid FlexTouch*). Given that *Levemir FlexTouch* is a different brand of insulin that has a different acting time and administering instructions, the severity of this error in the output is critical.

Polysemy, synonymy, and register matches An example of this issue is the treatment of acronyms. The term *PSUR* (*Periodic Safety Update Report*) present as such in the source was at times translated correctly as *RPAS*, or given its full form *raport periodic actualizat privind siguranța*, which is offered by the two 5-shot MT outputs as it appears in three out of the five examples provided in the respective prompts. The same for ACR 20, which was only sometimes explicitated (Table 7 **example 2**). This variation in translation approaches present in the prompting examples leads to inconsistencies in and unnecessary editing of the MT output.

A further example from English-German is the use of synonyms that have different connotations, at times being too weak or too strong for the particular context. In **example 6** (Table 8), *genehmigt* (*approved*) was used to translate *authorised.* In a different context, this would be a viable translation; however, the stronger and more accurate term *zugelassen* (*authorised*) is preferable. The corresponding noun *Zulassung* refers to the marketing authorisation process, which has to be completed before medicines can be marketed and made available to patients. As in previous examples, the particular word choice in the output can likely be attributed to the example used in the 1-shot EM prompt.

Strong influence of English As the international language of medicine, English has an influence on the lexis, morphosyntax, and typography of the target language. Most prevalent are literal translations (Table 7 **example 6** and Table 8 **example 7**). **Example 7** (Table 8) shows that *monitor* was translated as *überwachen* (*surveil*), which is more appropriate in a policing context and might elicit negative emotions among patients. A more neutral verb such as *beobachten* (*observe*) or *im Auge behalten* (*keep an eye on*) might be a more idiomatic and less intimidating choice. Interestingly, *beobachten* was used in the 1-shot FM prompt, which again points to the variable quality of the translations in the EMEA corpus.

Overall, the addition of 1- and 5-shot examples improve the results compared to the 0-shot baseline for both our language pairs and in all few-shot configurations. This information can support translators and the language services industry in general when exploring suitable local solutions. Moreover, when using the FM configuration, the BLEU scores across some LLMs produce statistically significant

($p < 0.05$) improvements compared to the EM setting. For English-Romanian, the 5-shot FM configuration leads to the best results. For English-German, the 5-shot EM configuration leads to the best results.

7 Conclusions and Future Work

In this study, we compared the impact of fuzzy matching against semantic sentence embeddings for few-shot prompted LLMs on translation performance within a specialised domain. We used the EMEA medical corpus for the English-Romanian and English-German language pairs, and we evaluated translation quality with automatic metrics and manual evaluation. Our findings show that 1-shot and 5-shot prompting significantly outperforms the 0-shot baselines for quality in automatic evaluations for both language pairs, and in manual evaluation for English-German. For the English-Romanian pair, the average scores of the manual evaluation for both quality and ranking follow the same trend, but statistical significance is not consistently reached for all few-shot prompting configurations.

Due to the different computational resource costs and technical knowledge implications associated with obtaining FM examples (low) versus EM examples (high), as well as our different results for the two language pairs (for English-Romanian 5-shot FM is best, for English-German 5-shot EM is best), further research is needed with different domain corpora and languages. At the same time, evaluators who are domain experts will be involved. In addition, future work will investigate the impact of combining Retrieval-Augmented Generation with ICL on the quality of LLM MT output.

Acknowledgements

We want to thank the reviewers for their time and valuable comments, and our pharmacist colleague for providing input on specialised terminology.

References

Artetxe, Mikel and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610. Place: Cambridge, MA.

Bawden, Rachel and François Yvon. 2023. Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland, June. European Association for Machine Translation.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners, July. arXiv:2005.14165 [cs].

Duquenne, Paul-Ambroise, Holger Schwenk, and Benoît Sagot. 2023. SONAR: Sentence-Level Multimodal and Language-Agnostic Representations, August. arXiv:2308.11466 [cs].

ELG, ELG. 2020. ELG - Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020).

Esperança-Rodier, Emmanuelle, Francis Brunet-Manquat, and Sophia Eady. 2019. ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. In *Translating and the computer 41*, Londres, United Kingdom, November.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May. Association for Computational Linguistics.

Gaona, Miguel Angel Rios, Raluca-Maria Chereji, Alina Secara, and Dragos Ciobanu. 2023. Quality Analysis of Multilingual Neural Machine Translation Systems and Reference Test Translations for the English-Romanian language pair in the Medical Domain. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 355–364, Tampere, Finland, June. European Association for Machine Translation.

Graham, Yvette, Timothy Baldwin, and Nitika Mathur. 2015. Accurate Evaluation of Segment-level Machine Translation Metrics. In Mihalcea, Rada, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado, May. Association for Computational Linguistics.

Kocmi, Tom, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA, November. Association for Computational Linguistics.

Liu, Alexander H., Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. 2026. Ministral 3, January. arXiv:2601.08584 [cs].

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

Martins, Pedro Henrique, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM: Multilingual Language Models for Europe. *Procedia Computer Science*, 255:53–62, January.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland, June. European Association for Machine Translation.

Muñoz-Miquel, Ana. 2025. Approaching machine translation in the medical and health field: An exploratory study. *The Journal of Specialised Translation*, (44):22–43, July.

Nieminen, Tommi, Jörg Tiedemann, and Sami Virpioja. 2025. Incorporating Target Fuzzy Matches into Neural Fuzzy Repair. In Johansson, Richard and Sara Stymne, editors, *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 408–418, Tallinn, Estonia, March. University of Tartu Library.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for

MT Evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Rei, Ricardo, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging Generality and Translation Specialization in Multilingual LLMs, June. arXiv:2506.17080 [cs].

Vela, Mihaela and Josef van Genabith. 2015. Reassessing the WMT2013 Human Evaluation with Professional Translators Trainees. In El-Kahlout, İlknur Durgar, Mehmed Özkan, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Fred Hollowood, and Andy Way, editors, *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 161–168, Antalya, Turkey, May.

Vilar, David, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada, July. Association for Computational Linguistics.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July. arXiv:1910.03771 [cs].

Xu, Jitao, Josep Crego, and Jean Senellart. 2020. Boosting Neural Machine Translation with Similar Translations. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online, July. Association for Computational Linguistics.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report, May. arXiv:2505.09388 [cs].

Zebaze, Armel Randy, Benoît Sagot, and Rachel Bawden. 2025. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico, April. Association for Computational Linguistics.

A Evaluation Guidelines

You are shown 100 segments, each with a source and its reference translation, followed by five candidate translations.

Your first task is to rank the candidate translations from best to worst (1-5). Ties are allowed. As reference, in a segment with three candidate translations ranked as best followed by the remaining two, the ranking is 1,1,1,2,3.

Your second task is to score (0-100) each candidate translation based on the errors encountered (meaning) and their severity (minor or major), as well as the amount of overall editing (effort) you consider a translator would need to spend in order to correct the translation. Score taking in mind the following sub-categories:

- 0% no meaning
- $\geq$ 33% some meaning preserved and major editing needed
- $\geq$ 65% most meaning preserved and some editing needed
- 100% perfect.

Error types and severity (as an indication, please use your overall judgement):

- Lower caps when full or mixed caps needed - 5 points
- General formatting issues (bullets) - 5 points
- Inadequate informal or formal use - 5/10 points
- Wrong grammatical agreements (lack of agreement in the names of medicines such as *darunavir* instead of *darunavirului*) - 5/10 points
- Missing punctuation in localisation (10000 for 10.000) - 5/10 points

- Superfluous additions - 5/10 points
- Incorrect diacritics (*puteţi apăsă* instead of *puteţi apăsa*) - 5/10 points
- Using adjective superlative form rather than the comparative one - 10 points
- Hallucinations (*substancelor activă(e)* instead of *substanței(lor) activă(e)*) - 10 points
- Subtle spelling errors in the names of the medicines (*didanosină* instead of *didanozină*) - 10 points
- Unnecessary use of English words (*clearance* used in Romanian instead of its translation) - 5/10 points
- Wrong specialised term - 10 points
- Editing changes relating to fluency - 5/10 points

Downgrade to the next sub-category if:

- Change in the meaning (positive becomes negative); also if done using one of the items above (such as adjective form)
- A significant modifier missed or introduced (*never*, *exclusively*, or *only* is dropped)
- Wrong key specialised term
- Incorrect numbers or punctuation changing the meaning (63 instead of 200, or 10,000 instead of 10.000)