Retrieval-based Adaptation For Machine Translation Applications Using Large Language Models And In-Context Learning

João Fonseca
joao.r.fonseca@tecnico.ulisboa.pt

Abstract

Large Language Models (LLMs) are one of the applications of artificial intelligence with the most world-wide impact across a variety of different areas. In this work we further the existent research of their application to Machine Translation (MT), through the fields of MT evaluation, terminology-constrained MT and Automatic Post-Editing (APE). For this, we leverage the technique that allows the excellence of LLMs in many different areas of research without being explicitly trained to do so, *in-context learning*. We show that, despite their recent attention, these types of models can have very competitive performance against the state-of-the-art results in some fields, with little effort and cost compared to the industry standards.

1 Introduction

In the past few years, MT has been adopted in several real-world applications. However, the evolution of the technologies behind MT systems is not slowing down, and training and developing these models is becoming more expensive, with constantly growing models requiring an expanding amount of data to be trained. LLMs are one example of this scaling, appearing as a consequence of the ever-growing performance of Pre-trained Language Models (PLMs) on downstream tasks, the higher the parameter size and/or amount of training data (Kaplan et al., 2020).

One of the reasons that reinforces the rise of LLMs and distinguishes them from PLMs is the appearance of emergent abilities, which can be defined as abilities that are not present in small models but arise in larger ones (Wei et al., 2022). In other words this means that, after scaling up a model up from a certain point, performance in many different tasks rise substantially above random guessing. *In-context learning*, *instruction tuning* and *step-by-step reasoning* are the three main abilities that LLMs show.

In-context learning was introduced by Generative Pre-trained Transformer 3 (GPT-3) (Brown et al., 2020), one of the first LLM to be introduced, and it consists on the ability that LLMs possess of achieving surprising performance on downstream tasks by providing a few input-label demonstrations related to specific tasks.

These models have had a clear impact on society (Movva et al., 2023), with the proportion of research papers having a large growth recently as well as many companies world-wide adopting a LLM-based focus towards their business [1]. LLMs potential, recent growth, and impact in society were all reasons that motivated this work on retrieval-based machine translation application using LLMs through *in-context learning*.

2 Related Work

LLMs have been recently shown to be able to perform numerous varied tasks, such as machine translation, MT evaluation, and APE (Kocmi and Federmann, 2023; Dinu et al., 2019; Raunak et al., 2023), with high quality, despite not being fine-tuned for these purposes.

(Kocmi and Federmann, 2023) introduces zero-shot MT evaluation experiments using LLMs. The paper introduces `GEMBA`, a GPT-based metric for assessment of translation quality. They investigate nine versions of GPT models through zero-shot prompting, both with and without the use of references. The main conclusions are that the metric works well on document-level but lacks segment-wise.

Previous approaches for terminology-constrained MT, such as (Dinu et al., 2019), train a MT model to handle terminology con-

[1] Based on a recent survey by the Cutter Consortium (`https://www.cutter.com/article/generative-ai-enterprise-status-practices-trends`), approximately one third of organisations plan to integrate LLMs into their own applications.

straints during inference. This builds on previous work that focused on constrained decoding, an approximate search algorithm capable of enforcing any constraints over resulting output sequences. This introduces substantial computational overhead in the decoding phase during inference and shows inflexibility and stiffness when including terminology.

APE refers to the task of proposing improvements over a given translation, *T*, and generating the translation with the proposed improvements T^{+}. (Raunak et al., 2023) introduced the usage of GPT LLMs for the task of APE, on a zero-shot scenario. The paper focuses on the nature of the post-edited translation, general quality improvements, edits on human annotated error spans and fidelity of proposed edits.

3 Experimental Setup

For all the performed experiments, embeddings used as keys of the datastores for the *in-context learning* experiments were computed using the Language-Agnostic BERT Sentence Embedding (LaBSE) model (Feng et al., 2022) and retrieved using the euclidean distance. To create the datastores and perform the nearest neighbour search, the FAISS library (Johnson et al., 2017) was used. The used LLMs were `gpt-3.5-turbo` for all experiments and `gpt-4` for some APE experiments, both with a cutoff date of June 2023, *i.e.*, the model has not received any updates since the referred date.

The metrics used to evaluate our experiments were pair-wise accuracy (Kocmi et al., 2021) for system-level and Kendall's Tau-b (Freitag et al., 2022) for segment-level correlations, in the MT evaluation section and Bilingual Evaluation Understudy (BLEU), Crosslingual Optimized Metric for Evaluation of Translation (COMET) and CHaRacter-level F-score (chrF) for general quality measures of the remaining experiments.

4 Machine Translation Evaluation using LLMs

In this section we build on (Kocmi and Federmann, 2023) zero-shot MT evaluation experiments using various LLMs, by introducing few-shot in-context examples.

4.1 Datasets

The used test set is the Multi-dimensional Quality Metrics (MQM) 2022 human judgements for the English to German and Chinese into English language pairs. It contains a total of 54 machine translation systems, most of them participants of the Workshop on Machine Translation (WMT) 2022 general MT shared task (Kocmi et al., 2022). The gold standard is human MQM ratings annotated by professionals who mark translation errors in each segment according to (Freitag et al., 2021).

We consider experiments with and without reference based on the `GEMBA-DA` framework prompts used by (Kocmi and Federmann, 2023) which are scored from 0 to 100. All scores reported in the WMT-22 Metrics shared task findings paper were reproduced using the official script (Freitag et al., 2022).

4.2 Few-Shot Scenario

The examples for the few-shot learning experiments were retrieved from a pool of MQM annotated segments from WMT 2019, 2020 and 2021. The examples are appended to a slightly modified version of the `GEMBA-DA` prompt used by (Kocmi and Federmann, 2023), to accomodate the use of few-shot examples. Three different of querying methods were experimented for each k-shot experiment ($k \in \{0, 1, 2, 3\}$), described as follows:

- **A)** k most similar sources and then choose a translation from a random system;
- **B)** k most similar concatenation of source and machine translation (if we have ten different systems, for each source segment we will have ten different concatenations of source and machine translation);
- **C)** most similar source and choose k most similar translations associated with that source.

Reference and reference-free system-level results are shown in Table 1. This table contains the best performing few-shot experiments, made by us, the most relevant zero-shot experiments made by (Kocmi and Federmann, 2023) and some experiments involving some of the state-of-the-art neural metrics (COMET-22, COMET-QE-22, BLEURT-20) (Rei et al., 2022a; Sellam et al., 2020), for the English to German and Chinese to English language pairs.

In Table 2 are the segment-level correlations for both language pairs of the few-shot experiments, state-of-the-art neural metrics and zeor-shot experiments made by (Kocmi and Federmann, 2023).

Table 1: Results of system level accuracy for English to German and Chinese to English language pairs, using the GEMBA-DA framework. In yellow are the zero-shot experiments made by (Kocmi and Federmann, 2023).

Model	Setup	Reference	Query Method	Accuracy (en-de)(%)	Accuracy (zh-en)(%)	Accuracy (%)
GPT-3.5	2 shot	No	A	**0.949**	**0.890**	**0.917**
GPT-3.5	5 shot	No	A	0.910	0.879	0.894
Davinci-003	0 shot	Yes	-	0.923	0.868	0.893
GPT-3.5	2 shot	No	B	0.936	0.853	0.890
GPT-3.5	2 shot	No	C	0.923	0.861	0.889
GPT-3.5	1 shot	No	B	0.935	0.851	0.888
GPT-4	0 shot	Yes	-	0.897	0.868	0.882
Davinci-002	0 shot	Yes	-	0.872	0.835	0.852
GPT-3.5	2 shot	Yes	C	0.877	0.830	0.852
GPT-4	0 shot	No	-	0.846	0.857	0.852
Davinci-003	0 shot	No	-	0.872	0.824	0.846
COMET-22	-	Yes	-	0.769	0.868	0.822
GPT-3.5	0 shot	No	-	0.782	0.857	0.822
BLEURT-20	-	Yes	-	0.769	0.846	0.822
GPT-3.5	0 shot	Yes	-	0.795	0.780	0.787
COMET-QE-22	-	No	-	0.718	0.813	0.769

4.3 Results

4.3.1 System-Level

LLMs are state-of-the-art system-level evaluators of machine translation. Moreover, by augmenting GPT-3.5 zero-shot prompt with relevant few-shot examples, its performance increases largely, achieving a new state-of-the-art result in system-level accuracy for these two language pairs.

Among the few-shot experiments, the query method seems to have little to no effect on the performance of the model, despite the last query method consistently yielding the worst results. On the other hand, the 2 shot experiments seem to consistently outperform the 1 shot ones (most of the 1 shot experiments are not in Table 1 because they yielded worse results and trend similarly to the 2 shot experiments), and also the 5 shot scenario, which would indicate that the model takes advantage of the few-shot examples, but when providing too many it starts deteriorating the results.

Another interesting trend is the fact that when using few-shot experiments the inclusion of the reference seems to deteriorate results which is a contrary trend to the one obtained in the zero-shot experiments.

4.3.2 Segment-Level

The segment-level results do not yield the same success. At the top of the table, with the best performance, are the neural state-of-the-art metrics, even ones that did not perform well in the system-level (and therefore were not included in Table 1), such as UniTE (Wan et al., 2022), COMET, BLEURT-20 and MetricX-XXL, which completetly outclass the few-shot experiments using GPT-3.5. The few-shot experiments improve slightly on the reference-free zero-shot scenario using the same model (although the results are very low), and are very outclassed by the reference-based version of the model. GPT-4 is, by far, the best performing LLM, both with and without the use of reference, managing to compete with the state-of-the-art neural metrics, which begs the question of whether few-shot learning on top of GPT-4 would improve the results even further.

Another interesting take is that, on the segment-level, we observe a contrary trend to what was observed at the system-level, which is that reference-based models perform better than reference-free ones, which is the norm among neural metrics.

4.4 Reliability Considerations

Another relevant aspect to analyse is the reliability of the answers provided by the LLM on the various experiments. We perform the exact same method implemented by (Kocmi and Federmann, 2023) of increasing temperature upon invalid answer, however it is interesting to note that with the model `gpt-3.5-turbo` they obtained 565 and 935 invalid answers for the reference-based and reference-free experiments, while with few-shot we obtained 0 and 3 respectively. This indicates that providing few-shot examples largely improves the reliability of the provided answers regarding the task-specific rules.

The distribution of the outputted scores was also mentioned in the paper (Kocmi and Federmann, 2023). On the zero-shot experiments, the authors theorised that one of the factors behind the low segment-level scores could be the fact that the models outputs mostly scores multiples of five, and over

Table 2: Segment-level correlations for Chinese to English using the GEMBA-DA framework (Kocmi and Federmann, 2023). In yellow are the zero-shot experiments reported by the paper.

Model	Setup	Reference	Query Method	Correlations (en-de)(%)	Correlations (zh-en)(%)
UniTE	-	Yes	-	**0.362**	0.351
COMET-22	-	Yes	-	0.361	0.420
MetricX-XXL	-	Yes	-	0.356	**0.421**
GPT-4	0 shot	Yes	-	0.347	0.370
BLEURT-20	-	Yes	-	0.338	0.352
GPT-4	0 shot	No	-	0.337	0.394
Davinci-003	0 shot	Yes	-	0.301	0.360
GPT-3.5	0 shot	Yes	-	0.299	0.344
COMET-QE	-	No	-	0.277	0.356
GPT-3.5	2 shot	No	A	0.266	0.248
GPT-3.5	2 shot	Yes	A	0.253	0.271
GPT-3.5	2 shot	No	B	0.241	0.236
GPT-3.5	1 shot	Yes	A	0.238	0.253
GPT-3.5	1 shot	Yes	C	0.233	0.243
GPT-3.5	2 shot	No	C	0.233	0.249
GPT-3.5	2 shot	Yes	B	0.232	0.266
GPT-3.5	1 shot	Yes	B	0.231	0.245
GPT-3.5	1 shot	No	A	0.230	0.244
GPT-3.5	2 shot	Yes	C	0.230	0.269
Davinci-002	0 shot	Yes	-	0.228	0.294
GPT-3.5	1 shot	No	C	0.228	0.238
GPT-3.5	1 shot	No	B	0.226	0.232
GPT-3.5	0 shot	No	-	0.225	0.352
GPT-3.5	5 shot	No	A	0.217	0.238
Davinci-002	0 shot	No	-	0.203	0.270
Davinci-003	0 shot	No	-	0.176	0.275

three quarters of the answers are either 80, 95 or 100. For the reference-free experiments 60.5% of scores outputted were 95. When adding few-shot examples the scores became much more distributed across the spectrum, which could be a reason for the low segment-level scores obtained. The model is correctly ranking systems among each other, but is using a much more distributed score board which causes a substantial decrease in segment-level correlations.

5 Terminology-Constrained Machine Translation

In this section we will evaluate and analyse recent widely used LLMs for terminology-constrained MT, through the use of in-context learning ability, using zero and few-shot learning approaches fetching examples from a local datastore.

5.1 Datasets

The datasets used are portions of the English to German publicly available terminology databases, Wiktionary and IATE[2], which is constitutes around 727 sentences for the Wiktionary test set and 414 sentences for the IATE one. Furthermore, terminology entries that occur in the English top 500 most frequent words, or that are single character were removed, as well as the term bases were divided in two different sets, training and test, making sure there is no overlap on the source side, just as was done by (Dinu et al., 2019).

5.2 Few-Shot Scenario

The zero-shot scenario is constituted by a task-specific introductory sentence, followed by the sentence to translate and respective glossary, shown in Table 3. The few-shot prompt builds on the zero-shot one by appending the examples in the same format as depicted in Table 3 before the final sentence to translate.

The training sets (one for each of the Wiktionary and IATE datasets) described earlier is used as the pool of examples to retrieve from in order to create the few-shot prompt. These constitute only 168 sentences for the IATE experiments and 248 sentences for the Wiktionary experiments).

[2]More information in https://iate.europa.eu and https://www.wiktionary.org/

Table 3: Base version of the zero-shot scenario prompt for the task of terminology-constrained MT. SL stands for source language, TL stands for target language.

Prompt
Translate the following sentence from English to German without providing any explanation and using the provided glossary. Glossary: {term$_1$ in SL}={term$_1$ in TL} ; ... ; {term$_k$ in SL}={term$_k$ in TL} {SL} source: {source_sentence}. Your {TL} translation:

Table 4: Term percentage and quality scores (BLEU, COMET and chrF) of terminology-constrained MT for the English to German language pair using the base prompt.

Wiktionary **Base prompt, English-German**				
	Term Percentage (%)	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	87.75	55.46	87.50	63.85
0 shot	95.80	56.39 (+0.93)	87.72 (+0.22)	64.67 (+0.82)
1 shot	96.15	56.17 (+0.71)	87.77 (+0.27)	64.45 (+0.60)
2 shot	96.03	56.53 (+1.07)	87.88 (+0.38)	64.76 (+0.91)
3 shot	96.50	56.53 (+1.07)	**87.90 (+0.40)**	64.70 (+0.85)
4 shot	**96.58**	**56.74 (+1.28)**	**87.90 (+0.40)**	**64.79 (+0.94)**

IATE **Base prompt, English-German**				
	Term Percentage (%)	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	84.27	54.85	87.05	62.96
0 shot	95.80	55.66 (+0.81)	87.40 (+0.35)	63.85 (+0.89)
1 shot	95.73	55.66 (+0.81)	87.43 (+0.38)	63.94 (+0.98)
2 shot	**95.96**	55.41 (+0.56)	87.44 (+0.39)	63.86 (+0.90)
3 shot	95.28	55.88 (+1.03)	87.52 (+0.47)	64.13 (+1.17)
4 shot	95.93	**55.97 (+1.12)**	**87.54 (+0.49)**	**64.23 (+1.27)**

5.3 Results

5.3.1 Base Prompt

The results are shown in Table 4. Terminology percentages, which are the percentage of times the term translation was generated in the output out of the total number of term annotations, improves significantly (around 8 to 9 percent points) when introducing the glossary (which corresponds to $0-4$ shot scenarios). In fact, the 3 to 5 % of terms where the LLM is not able to output the correct terminology corresponds to 29 to 34 sentences in the Wiktionary dataset and 18 to 21 sentences in the IATE dataset (depending on the experiment). Due to the small-sized test set and the high terminology percentages we can manually analyse the instances where the model failed to output the correct terminology.

The first situation in which the model fails is by using synonyms, or even different words, of the ones in the presented terminology. Some examples are when the model uses the word "Abend" instead of "Nacht" on the 74^{th} sentence of the IATE dataset, or in the 146^{th} sentence of the same dataset, where the model uses its own expression "beiden Seiten" instead of the requested "beidseitig", or in the 182^{nd} of the Wiktionary dataset where the model substitutes the word "league" ("Liga") with the synonym "class" ("Klasse"). Secondly we have situations in which the model ignores the glossary completely, such as in the 146^{th} sentence of the IATE dataset where the model uses the full word "Weltmeisterschaft" instead of the requested abbreviation "WM". Finally we have situations in which the models uses an inflected version of the requested word, or with a different casing than the requested as in the examples shown in the 132^{nd} sentence of the IATE dataset, in which the model uses "Die republikanischen" instead of the requested "Die Republikaner", in sentence 219 of the Wiktionary dataset in which the model uses "statt" instead of the terminology "stattfinden".

Regarding quality metrics, the 0-shot scenario improved over the baseline across all metrics. The fact the lexical-based metrics such as chrF and BLEU also improved, higher than the values reported by (Dinu et al., 2019), is an indication that the model is able to not only insert the correct terminology but also adapt the surrounding words to the word inserted, contrary to previous methods like constrained decoding (Chatterjee et al., 2017; Hokamp and Liu, 2017; Hasler et al., 2018).

Our method obtained higher terminology percentages and quality improvements than the method proposed by (Dinu et al., 2019).

5.3.2 Simpler Prompt

To test the effect of the prompt on the results, a simpler version of the base prompt (shown in Table 3) was tested. The simpler version corresponds to the base version without the task-introductory sentence. The simpler version of the prompt did not affect the terminology percentage metric, which stayed roughly the same in every scenario (slightly lower in the 0 shot experiment), however, the quality metrics have a different pattern. In the 0 shot experiments, all quality metrics decrease considerably over the base prompt experiments, even becoming

lower than the baseline. From the 1-4 shot experiments, quality metrics go up again, surpassing the ones of the base prompt experiment across all quality metrics. This indicates that, without either a task-introductory sentence or few-shot examples, the model is not able to perform well. Moreover, the inclusion of both seems to consistently yield worse results than with just the inclusion of few-shot examples.

5.3.3 Datastore with Large Amounts of Parallel Data

The extremely small size of the datastore in the previous experiment (168 sentences for the IATE experiments and 248 sentences for the Wiktionary experiments) is a bottleneck. Since the retrieval is done by source similarity, having a larger sized datastore with similar-domain content, allows to retrieve better examples. Following this, a different style of datastore was experimented using the training data of the WMT 2017 news translation task (which contains around 6M news-related sentences). Through this experiment we can analyse the trade-off between providing better source-translation examples but without the in-context terminology component or less similar source-translation examples but with the terminology component (experiments presented in the previous section).

By doing this, in all metrics except chrF, the best results were obtained in the 0 or 1 shot scenarios. This indicates that by providing more out-of-context examples does not help the model to include terminology and also deteriorates the quality results. The terminology inclusion errors in this setup are of the same nature as the ones already analysed above.

6 Automatic Post-Editing using Large Language Models

Recent work has been done on investigating the ability of LLMs (such as GPT 3 and 4) to handle the task of automatically post-editing of machine translations, due to their versatility and recent popularity, on a zero-shot scenario (Raunak et al., 2023). This chapter proposes to further research this topic, including an analysis of the capability of these LLMs to automatically choose which machine translations benefit from a post-edit step as well as extending all existent research to a few-shot scenario.

Table 5: Base version of the zero-shot scenario prompt for the task of APE. `SL` stands for source language, `TL` stands for target language.

Prompt
You're going to improve a given sentence which is a machine translation in {TL} from a source sentence in {SL}, without providing explanations. {SL} source: {source_sentence}. {TL} machine translation: {translation} Your improved translation (in {TL}):

6.1 Datasets

The datasets used are the WMT-22 general machine translation task (Kocmi et al., 2022) and WMT-21 news translation task annotated with MQM errors (Freitag et al., 2021), namely the German to English and the Ukrainian to Czech language pairs, with the first one being a high resource language pair and the second one being a low resource language pair.

In the German to English experiments the `Lan-Bridge` and `PROMT` systems were used (best performing and worst performing systems), while in the Ukrainian to Czech experiments the `ALMAnaCH-Inria` system (worst) was used.

6.2 Few-Shot Scenario

Table 5 shows the zero-shot prompt for the experiments. The few-shot scenario prompt builds on the zero-shot one by including the examples after the initial task-introductory sentence. These examples are retrieved from a pool of with the WMT-22 training set for each used language pair.

6.3 Results

6.3.1 German to English Scenario

The first experiment was to use `gpt-3.5-turbo` and apply the base prompt shown in Table 5 to the entire WMT-22 test set for the winning system, `Lan-Bridge`. The results can be seen in table 6. The baseline is the performance of the MT system without any post-editing step.

This approach yielded terrible results, due to the LLM not being able to distinguish good from bad translations. The system had mainly good translations, which caused the model to change most of

Table 6: Results of automatic post-editing experiments for the Lan-Bridge system and the German to English language pair using gpt-3.5-turbo.

Base prompt, German-English, gpt-3.5-turbo Sys: Lan-Bridge			
	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	**56.89**	**85.63**	**59.64**
0 shot	51.81 (-5.05)	61.91 (-23.72)	55.51 (-4.98)
1 shot	51.83 (-5.03)	62.09 (-23.54)	55.79 (-3.85)
2 shot	51.52 (-5.37)	62.06 (-23.57)	55.63 (-4.01)
3 shot	51.13 (-5.76)	62.06 (-23.57)	55.29 (-4.35)
4 shot	50.96 (-5.93)	62.05 (-23.58)	55.46 (-4.18)

Table 7: Results of automatic post-editing experiments for the German to English language pair using the base prompt, considering segments with COMET-QE score inferior to 85, for the Lan-Bridge system.

Base prompt, German-English, gpt-3.5-turbo Sys: Lan-Bridge, Only COMET< 85			
	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	**49.63**	78.97	**52.21**
0 shot	47.92 (-1.71)	80.25 (+1.28)	51.61 (-0.60)
1 shot	48.20 (-1.43)	80.40 (+1.43)	52.09 (-0.12)
2 shot	48.07 (-1.56)	80.43 (+1.46)	51.98 (-0.23)
3 shot	47.84 (-1.79)	80.40 (+1.43)	51.90 (-0.31)
4 shot	47.72 (-1.91)	**80.59 (+1.62)**	51.65 (-0.56)

these good translations to bad ones.

Consequently, a different version of the experiment was made, where we filter out all translations bellow a certain quality threshold using COMET-QE (wmt22-cometkiwi-da model) (Rei et al., 2022b). The results are shown in Tables 7, 8 and 9.

These experiments confirm the fact that gpt-3.5-turbo is not able to deduce on its own which translations would benefit from a post-editing step and which are not able to do so. Every quality score increases strictly and considerably from each table. As a general trend for the three experiments, adding few-shot in-context learning improves all quality scores over the zero-shot scenario. However, when we compare the results against the baseline only the COMET scores increase consistently. In Table 7 only COMET scores improve, in Table 8 both COMET and chrF improve and in Table 9 all quality scores improve over the baseline.

Table 8: Results of automatic post-editing experiments for the German to English language pair using the base prompt, considering segments with COMET-QE score inferior to 80, for the Lan-Bridge system.

Base prompt, German-English, gpt-3.5-turbo Sys: Lan-Bridge, Only COMET-QE < 80			
	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	**46.06**	73.90	48.28
0 shot	45.11 (-0.95)	76.59 (+2.69)	48.63 (+0.35)
1 shot	46.03 (-0.03)	**76.89 (+2.99)**	**49.39 (+1.11)**
2 shot	45.68 (-0.38)	76.70 (+2.80)	48.97 (+0.69)
3 shot	45.29 (-0.77)	76.87 (+2.97)	48.92 (+0.66)
4 shot	45.36 (-0.70)	76.83 (+2.93)	48.80 (+0.52)

Table 9: Results of automatic post-editing experiments for the German to English language pair using the base prompt, considering segments with COMET-QE score inferior to 65, for the Lan-Bridge system.

Base prompt, German-English, gpt-3.5-turbo Sys: Lan-Bridge, Only COMET-QE < 65			
	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	25.07	54.12	28.43
0 shot	33.06 (+7.99)	64.75 (+10.63)	34.75 (+6.32)
1 shot	**33.31 (+8.24)**	**67.37 (+13.25)**	**36.06 (+7.63)**
2 shot	32.45 (+7.38)	67.19 (+13.07)	35.04 (+6.61)
3 shot	32.05 (+6.98)	63.55 (+9.43)	34.47 (+6.04)
4 shot	31.21 (6.14)	65.57 (+11.45)	34.38 (+5.95)

Table 10: Results of automatic post-editing experiments for the German to English language pair using a modification of the base prompt that includes a quality indication.

Base prompt, German-English, gpt-3.5-turbo Sys: Lan-Bridge, QI on prompt			
	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	**56.89**	**85.63**	**59.64**
0 shot	52.61 (-4.28)	85.29 (-0.34)	56.34 (-3.30)
1 shot	52.90 (-3.99)	85.31 (-0.32)	56.62 (-3.02)
2 shot	52.33 (-4.56)	85.26 (-0.37)	56.13 (-3.51)
3 shot	52.12 (-4.77)	85.20 (-0.43)	56.01 (-3.63)
4 shot	51.92 (-4.97)	85.22 (-0.41)	55.78 (-3.86)

6.3.2 Quality Indication (QI) On The Prompt

In the previous section it was established that LLMs are not able to select which machine translations are worthy of performing post-editing. In order to further research on this statement, an alternative approach was experimented, with the objective of assessing the model's selection capabilities by changing the wording on the prompt to include a quality indication and instruct the model to only make this selection if necessary. The provided quality indication is obtained by computing a quality score using the reference-free model wmt22-cometkiwi-da, and distributing the scores over the tags {bad, ok, excellent}, using manually calculated thresholds (around the baseline's mean the QI is ok, above it it is excellent, and bellow it is bad). The results are shown in Table 10.

Table 11: Results of automatic post-editing experiments for the German to English experiments using the base prompt on the worst performing system of WMT-22 for this language pair, PROMT.

Base prompt, German-English, gpt-3.5-turbo Sys: PROMT			
	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	**53.90**	**83.84**	57.04
0 shot	53.00 (-0.90)	65.52 (-18.32)	57.01 (-0.03)
1 shot	53.20 (-0.70)	65.57 (-18.27)	**57.11 (+0.07)**
2 shot	53.08 (-0.82)	65.55 (-18.29)	57.08 (+0.04)
3 shot	52.67 (-1.23)	65.53 (-18.31)	56.89 (-0.15)
4 shot	52.43 (-1.47)	65.50 (-18.34)	56.66 (-0.38)

Analysing the results, we observe that the model is able to adapt really well to the new information of a quality indication indicated in the prompt with high quality. The scores are greatly superior to the vanilla case with the base prompt, despite still not being able to surpass the baseline. *i.e.*, the model's performance without an post-editing step. This indicates that the model's weakness of not being able to distinguish poor from great translations can be substantially mitigated through a change in prompt. In these experiments, adding more than one few-shot example deteriorates the results, since the best setup occurs mostly in the $1-2$ shot scenario for all metrics.

6.3.3 Worse Machine Translation System

Earlier in this section it was stated one of the reasons gpt-3.5-turbo struggled in the task of APE was because the experiments were done on the winning system of the German to English WMT-22 general machine translation task, Lan-Bridge, which produced a substantial amount of translations that do not benefit from a post-editing step. The next natural step is to consider the worst performing system on this translation task, which is the PROMT system. The results are shown in Table 11.

The results are positive compared to the initial experiments with the winning system of WMT-22 (in Table 6). The system is able to actually obtain substantially higher values of COMET, BLEU and chrF scores across all experiments, and is able to outperform the baseline in the chrF metric. These improvements indicate that there are less situations with a good translation hypothesis which is where the model struggles. In fact if we analyse the outputs, we observe that the core problem persists.

Table 12: Results of automatic post-editing German to English experiments for the base prompt using gpt-4.

Base prompt, German-English, gpt-4 Sys: Lan-Bridge			
	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	**56.89**	**85.63**	**59.64**
0 shot	49.89 (-7.00)	85.56 (-0.07)	54.66 (-4.98)
1 shot	47.19 (-9.70)	84.91 (-0.72)	52.22 (-7.42)
2 shot	48.12 (-8.77)	84.97 (-0.66)	52.84 (-6.80)
3 shot	48.15 (-8.79)	84.98 (-0.65)	52.79 (-6.85)
4 shot	48.26 (-8.63)	85.00 (-0.63)	52.86 (-6.78)

6.3.4 Experiments Using GPT-4

In this subsection we analyse the performance of a better LLM model which is the gpt-4[3], a direct improvement over the gpt-3.5-turbo model. Mainly it is relevant to analyse the ability for gpt-4 to detect which translations are already of high quality and do not benefit from a post-editing step. The results are shown in Table 12.

Comparing Table 6 with Table 12, GPT-4 significantly outperformsgpt-3.5-turbo when considering COMET scores, which demonstrates its superior ability in distinguishing good translations from poor ones, despite still not being able to surpass the baseline. However, it is important to note that the lexical-based quality metrics, BLEU and chrF, are significantly lower than the ones obtained through gpt-3.5-turbo. On the other hand, the inclusion of few-shot examples for the in-context learning task seems to considerably deteriorate all quality scores when comparing to the zero-shot scenario.

6.3.5 Low Resource Language Pair

So far only high-resource language pairs were considered. In this subsection we test the ability of the higher performant APE system, GPT-4, in the same scenario as the previous experiments but in a low- resource language pair, which is Ukrainian to Czech. The results are shown in Table 13.

GPT-4 is able to perform very well in this low-resource language pair. All quality metrics increase over the baseline by a large amount, and also, the more few-shot examples are provided in the prompt, the higher the quality scores for the most cases, although the biggest increases occurs from the 0 to 1 shot and from 1 to 2 shot experiments. The best results out of the three tables were obtained in the datastore with the pool of 34M

[3] https://platform.openai.com/docs/models/gpt-4.

Table 13: Results of automatic post-editing experiments for the Ukrainian to Czech language pair using the base prompt and using the `gpt-4` model, for the system `ALMAnaCH-Inria`.

Datastore size: 34M sentences Base prompt, Ukranian-Czech, gpt-4 Sys: `ALMAnaCH-Inria`			
	BLEU (Δ)	COMET (Δ)	chrF (Δ)
baseline	43.61	82.30	49.80
0 shot	47.04 (+3.43)	85.23 (+2.93)	53.13 (+3.33)
1 shot	50.27 (+6.66)	89.58 (+7.28)	55.76 (+5.96)
2 shot	51.91 (+8.30)	91.52 (+9.22)	58.19 (+8.39)
3 shot	52.36 (+8.85)	91.79 (+9.49)	**58.74 (+8.94)**
4 shot	**52.48 (+8.87)**	**92.13 (+9.83)**	58.67 (+8.87)

sentences for BLEU and COMET, while chrF was higher in the datastore with 17M sentences (by a slight amount).

7 Conclusions

LLMs have achieved ground-breaking development with the introduction of dozens of new models in the past few years. The power to extend and diversify their range of skills with the increase of model size and training data, known as the emergent skills (Wei et al., 2022), is one of their main reasons for interest. This work shows the versatility of LLMs by testing the technique of *in-context learning* on different MT applications: MT evaluation, terminology-constrained MT and APE of machine translations.

Regarding MT evaluation, Although the segment-level results obtained by GPT LLMs are still not comparable to current state-of-the-art neural metrics (such as COMET), the document-level results achieve state-of-the-art results, further increasing with the inclusion of few-shot examples.

In terminology-constrained MT GPT-3.5 showed great terminology inclusion percentage as well as improvements in the quality metrics. It was clear that the LLM sometimes tends to ignore the instruction given completely, using synonyms instead of the required temrinology for example, which raises reliability issues. It was also shown that for this task, it is better to include worse but task-specific examples from a lower pool of sentences then better examples of general machine translations, without the glossary portion.

Lastly, in APE, the LLM models show poor capabilities of distinguishing poor translations from great ones that do not benefit from a post-editing scenario, although GPT-4 outperforms GPT-3.5. However, when provided with only poor translations, which is the standard scenario in a real-world APE pipeline, the model is able to perform the correct modifications needed to improve the quality scores. A big advantage of LLMs is their ability to adapt to different scenarios, which was demonstrated through different experiments that varied the prompt template according to certain assumptions, and on different language pairs, without any further training.

As a final note, this work implicitly reinforces the amazing versatility that LLMs have. It was shown that a single LLM could perform every step of a real-world MT pipeline, ranging from machine translation, MT evaluation and automatic post-editing, without any further training.

References

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

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference*

on Machine Translation (WMT), pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2023. Large language models shape and are shaped by society: A survey of arxiv publication patterns.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

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

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.