MTUncertainty: Assessing the Need for Post-editing of Machine Translation Outputs by Fine-tuning OpenAI LLMs

Serge Gladkoff[2], **Lifeng Han**[1], **Gleb Erofeev**[2], **Irina Sorokina**[2], and **Goran Nenadic**[1]

[1] The University of Manchester, UK

[2] Logrus Global, Translation & Localization

lifeng.han, g.nenadic @ manchester.ac.uk

gleberof, irina.sorokina, serge.gladkoff @ logrusglobal.com

Abstract

Translation Quality Evaluation (TQE) is an essential step of the modern translation production process. TQE is critical in assessing both machine translation (MT) and human translation (HT) quality without reference translations. The ability to evaluate or even simply estimate the quality of translation automatically may open significant efficiency gains through process optimisation. This work examines whether the state-of-the-art large language models (LLMs) can be used for this uncertainty estimation of MT output quality. We take OpenAI models as an example technology and approach TQE as a binary classification task. On **eight language pairs** including English to Italian, German, French, Japanese, Dutch, Portuguese, Turkish, and Chinese, our experimental results show that fine-tuned *GPT3.5* can demonstrate good performance on translation quality prediction tasks, i.e. *whether the translation needs to be edited*. Another finding is that simply increasing the sizes of LLMs does not lead to apparent better performances on this task by comparing the performance of three different versions of OpenAI models: *Curie*, *Davinci*, and *GPT3.5* with 13B, 175B, and 175B parameters, respectively.

1 Introduction

Most modern translation projects include post-editing (PE) of machine-translation (MT) output (Han and Gladkoff, 2022; Gladkoff et al., 2022). Instead of translating from scratch, the MT+PE process increases productivity and allows to speed up global content delivery (Gladkoff and Han, 2022; Han et al., 2013). However, in regulated industries and many other scenarios raw MT output is not suitable for final publication due to the inevitable errors caused by the inherently stochastic nature of neural MT (NMT) (Han, 2022a; Freitag et al., 2021; Hong et al., 2024). Hallucinations, incorrect terminology, factual and accuracy errors, small and large, as well as many other types of mistakes are inevitable to varying degrees of extent, and therefore for premium quality publication human revision is required. MT output serves as input for a professional human translator, who reviews and revises the MT proposals to eliminate factual errors and ensure that the quality of translated material conforms to the customer specifications. At the same time even with those languages that are not handled well by MT, there is a significant portion of segments that are not changed after human review. This portion varies from 10% to 70% in some cases [1], and the question arises, "Is it possible to use machine learning (ML) methods to mark these segments and save time for human reviser and make them focus on those segments that need attention instead"? In other words, *Is it possible to capture editing distance patterns from data of prior editing of this material, which already has been made*? This could further speed up the translation process and decrease the costs while preserving the premium quality of the translated product.

This problem is also closely related to the traditional MT quality estimation (QE) shared task that has been held with the Workshop of MT (WMT) series since 2012 (Callison-Burch et al., 2012; Koehn et al., 2022; Zerva et al., 2022; Han et al., 2013; Han, 2022b), where both token-level and segment-level QE were carried out.

---

[1] logrusglobal.com statistics

From practical application and industrial usage, we formulate the problem into a single classification task, i.e. we are trying to solve a classification task to answer if the translated segment (sentence) needs to be edited, or not.

With the development of current large language models (LLMs), we choose OpenAI models as state-of-the-art LLMs to examine their capabilities for this task. In this work, our first experimental investigation is on "**Predictive Data Analytics with AI: assessing the need for post-editing of MT output by fine-tuning OpenAI LLMs**". We also follow up with an experiment that explores "**if the size of sample or LLM matters in such a task**" by experimenting with three OpenAI models: *curie*, *davinci*, and *gpt3.5*, with parameter sizes varying from 13B to 175B.

The rest of this paper is designed as below. Section 2 introduces related work to ours including MT-QE-related shared task and challenge events, Section 3 presents our methodology design and pilot study using two language pairs, Section 4 extends the experimental investigation with six more language pairs, section 5 discusses experiment on English-Japanese news content with the increasing sizes of training and testing corpus and explores two more OpenAI LLMs with varying model sizes, and Section 6 concludes this paper with future work and research perspectives.

2 Related Work

The Quality Evaluation (QE) of MT output has always been a critical topic for MT development due to its critical role in assessing quality in the process of training. In many cases, evaluation has to be done without seeing the reference translations. In many practical situations, reference translations are not available or even impossible to acquire, i.e. it is not practical to "manufacture" them for evaluation. The earliest QE shared task with the annual WMT conference started in 2012 when word level QE was introduced by (Callison-Burch et al., 2012) to estimate if the translated tokens need to be edited or not, such as deletion, substitution, or keeping it as it is. In the later development of QE, a sentence-level task was introduced to predict the overall segment translation scores, which are to be correlated with human judgement scores, such as using Direct Assessment (Graham et al., 2015). In WMT-2022, a new task on binary sentence-level classification was also introduced to predict
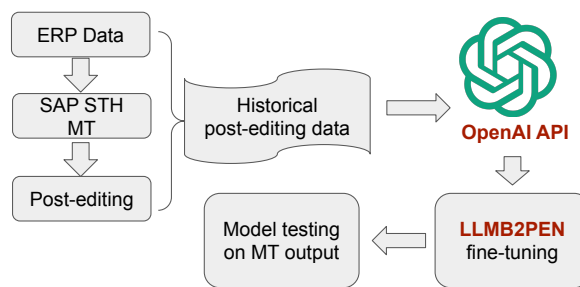

Figure 1: LLMB2PEN Methodology Design on Fine-tuning LLMs for Binary Prediction of Post-editing Need on Translations.

if a translated output has critical errors to be fixed on English-German and Portuguses-English language pairs (Zerva et al., 2022).

The recent methods used for such QE tasks included prompt-based learning using XLM-R by KU X Upstage (Korea University, Korea & Upstage) from Eo et al. (2022), Direct Assessment and MQM features integration into fine-tuning on XLM-R and INFOXLM (Chi et al., 2021) by the Alibaba team (Bao et al., 2022), and incorporating a word-level sentence tagger and explanation extractor on top of the COMET framework by Rei et al. (2022), in addition to historical statistical methods such as support vector machine (SVM), Naive Bayes classifier (NB), and Conditional Random Fields (CRFs) by Han et al. (2013).

However, *to the best of our knowledge*, this work is the first to investigate the OpenAI LLMs with varying sizes on such MT error prediction tasks with positive outcomes.

3 Methodology and Experiments

As shown in the system diagram in Figure 1, we first collect the historical post-editing data from our past projects on eight languages of Enterprise Resource Planning (ERP) content translation on English→German, French, Italian, Japanese, Dutch, Portuguese, Turkish, and Chinese (DE, FR, IT, JA, NL, PT, TR, ZH). This project was completed by using an MT engine to automatically translate the source into the eight languages, followed by post-editing by professional linguists. Two examples of MT and PE in English-Italian and English-German languages as Pilot Experiments are shown in Figure 2 and 3. Regarding MT system selection, since the content was from the ERP domain, we used the SAP STH as our MT

EN-USA	EN-IT MT	MT PE
Have affected personnel in the immediate area of the hazardous work been notified of the work to take place?	Il personale interessato nell'area immediatamente interessata dal lavoro pericoloso è stato informato dei lavori da svolgere?	Il personale interessato dai lavori nell'area vicina all'intervento pericoloso è stato informato sui lavori da svolgere?
Have all Energy Sources been restored per restart procedures and isolations removed?	Sono state ripristinate tutte le fonti energetiche per ogni ripresa e rimozione degli isolamenti?	Sono state ripristinate tutte le sorgenti di energia secondo le procedure di riavvio e sono stati rimossi gli isolamenti?
Have all equipment and safeguards been completely restored	Tutte le attrezzature e le misure di sicurezza sono state completamente ripristinate?	Attrezzature e salvaguardie interamente ripristinate?
Have all equipment and safeguards been completely restored?	Sono state completamente ripristinate tutte le attrezzature e le salvaguardie?	Attrezzature e salvaguardie interamente ripristinate?
Have all safeguards been completely restored?	Tutte le salvaguardie sono state completamente ripristinate?	Salvaguardie interamente ripristinate?
Have conditions changed?	Le condizioni sono state modificate?	Sono cambiate le condizioni?
Hazardous Energy Control (HEC) Procedural Format	Formato procedurale per controllo energetico pericoloso (HEC)	Format procedurale per controllo energetico pericoloso (CEP)
Hazardous Energy Control Permit	Permesso di controllo dell'energia pericolosa	Permesso di controllo energia pericolosa (HEC)
Hazardous Energy Control Permit completed as planned	Permesso di controllo dell'energia pericoloso completato come previsto	Permesso CEP (HEC) completato come pianificato
Hazardous Energy Control Permit Template	Modello di permesso di controllo dell'energia pericolosa	Modello di permesso di controllo energia pericolosa (HEC)
Hazardous operations within 35ft (11m) of work area are shut down	Le operazioni pericolose all'interno di 35 ft (11 m) dell'area di lavoro vengono chiuse	Le operazioni pericolose entro 35 ft (11 m) dell'area di lavoro vengono chiuse

Figure 2: EN-IT Examples on MT and Post-Editing

EN-USA	EN-DE MT	MT PE
Chemical/Liquid Protective Boots	Chemische/flüssige Schutzboote	Schutzstiefel für Chemikalien und Flüssigkeiten
Child Location	Unterlokation	Untergeordneter Standort
Child Permit	Kindgenehmigung	Untergeordnete Genehmigung
Child State	Untergeordneter Bundesstaat	Untergeordneter Zustand
Child Workflow	Untergeordneter Workflow	Untergeordneter Arbeitsablauf
Child Workflow State	Untergeordneter Workflow-Status	Untergeordneter Arbeitsablaufstatus
Chilled Water - specify Pressure	Kühlwasser - Geben Sie den Druck an.	Kühlwasser - Druck angeben
Choose Workflows to Import	Zu importierende Workflows auswählen	Zu importierende Arbeitsabläufe auswählen
The file size exceeds the limit allowed	Die Dateigröße überschreitet das zulässige Limit	Die Dateigröße überschreitet die zulässige Grenze
Qualified Person returned service to full operation and notified all affected workers?	Qualifizierte Person hat den Dienst in Betrieb genommen und alle betroffenen Mitarbeiter benachrichtigt?	Hat die qualifizierte Person den Betrieb wieder vollständig hergestellt und alle betroffenen Mitarbeiter benachrichtigt?
Qualified persons are First Aid/CPR Certified	Qualifizierte Personen sind Erste-Hilfe-Zertifizierung/CPR-zertifiziert	Qualifizierte Personen verfügen über Erste-Hilfe-Zertifizierung/CPR-Zertifizierung
Record results every 30 minutes for length of work.	Erfassen Sie die Ergebnisse alle 30 Minuten für die Dauer der Arbeit.	Ergebnisse alle 30 Minuten für die Dauer der Arbeit erfassen.
Record Worst Case Reading of any meter at any level in the space.	Erfassen Sie den Worst Case Reading eines beliebigen Zählers auf einer beliebigen Ebene des Bereichs.	Aufzeichnung des Worst-Case-Messwerts eines beliebigen Zählers auf einer beliebigen Ebene im Raum.
The symbol displayed at the top left of the application, used as the home button.	Das oben links in der Anwendung angezeigte Symbol, das als Home-Drucktaste verwendet wird.	Das oben links in der Anwendung angezeigte Symbol, das als Startseiten-Schaltfläche verwendet wird.

Figure 3: EN-DE Examples on MT and Post-Editing

engine. [2]

With this data from a real-world translation project, we used API to fine-tune the OpenAI *curie* model for our classification task. The input is the triple set (English source, MT outputs, post-edited "gold standard") we prepared in Phase 1. The goal of this step is to optimise the weights of the model parameters for our classification task. The custom fine-tuned model produced as a result of LLMB2PEN (LLM for Binary Prediction of Post-editing Need) method is created in our private space on the OpenAI account.

We *did not* apply "prompt engineering" for this task by doing zero-shot, one-shot, or few-shot training; we *did a full-scale fine-tuning* of OpenAI LLMs via API. It is important to note that we did not simply train the LLM for edit distance either; instead, the model was trained to learn whether the strings were edited or not taking into account the full content of the string and the entire context of the training data. One of the reasons that we did not use prompting is that "Prompt Engineering" of ChatGPT-3 is limited by 3,000 tokens, and

with ChatGPT-4 the context has been increased to 25,000 tokens, but still very significant limitation remains. OpenAI documentation states that 100 tokens = 75 words, meaning that the average sentence is 20 tokens, therefore 3000 tokens is only 150 sentences, or 75 translation units of bilingual text, or 50 segment triples of source, target and reference. The context of 25,000 sentences is only about 150 segment triples.

Also, fine-tuning is a deeper process of adjusting the model's weights, and not just an in-context learning. That's why we chose fine-tuning method, which is not constrained by such limitations.

For our classification experiment we took about 4000 lines of bilingual data in triples of source, target, and reference, and split it into train (large) and test (smaller) sets with a ratio of 9:1.

There were no specific selection criteria for the data because we took the entire project dataset after project completion. (Please, note that since we used the entire data from the actual project, and split the data set as 9:1, the sizes of test sets are not round and slightly different for different languages.)

---

[2] https://www.sap.com/ SAP is an enterprise resource planning, automation and business software company.

We also combined source sentences in groups of length, so that the test data set has the same distribution of sentences by their length as the training dataset.

Since the average sentence size is about 17 words, the training dataset contained about 35000 words of source data, 35000 words of MT output, and 35000 words of post-edited human reference.

It is also important to note what the model learns in this case - in such an experiment it learns not to translate, but to spot MT translation errors that were made by the specific MT engine in a specific language pair on particular content.

3.1 Outputs on EN-DE/IT

As a first step, we trained the *curie* LLM model using our data for two language pairs: English-Italian and English-German. To illustrate the results of prediction with our LLMB2PEN method, we draw the confusion matrix for both language pairs in Figures 4 and 5.

In the Confusion Matrix, from the top left corner in a clockwise direction, the 1st quadrant means True Negative (**TN**): segment is predicted as not requiring editing and it does not indeed require post-editing. The 2nd quadrant is False Positive (**FP**): segments which are predicted as requiring editing, but in reality, they do not, that is **FP** means that the segment is correct but wrongly flagged for post-editing. The 3rd quadrant is True Positive (**TP**) - reflecting the situation when a segment is correctly flagged as requiring post-editing. The fourth quadrant is False Negative (**FN**): segment is predicted as correct, while in reality, it does require post-editing. So the first and third are successful classifications, and the other two are incorrect classifications.

It is worth mentioning that if the segment is incorrectly predicted as requiring post-editing, this only leads to a small increase in post-editing cost, while False Negative predictions represent the consumer's risk of seeing substandard segments as not corrected in the final product. So in the context of our task, we are much more concerned with the share of False Negatives in the test classification dataset.

In the Italian situation shown in Figure 4, you can see that the model predicts correctly that many more translated sentences need to be edited (TP=503) than sentences that do not need to be edited (TN=191). In incorrectly predicted cate-
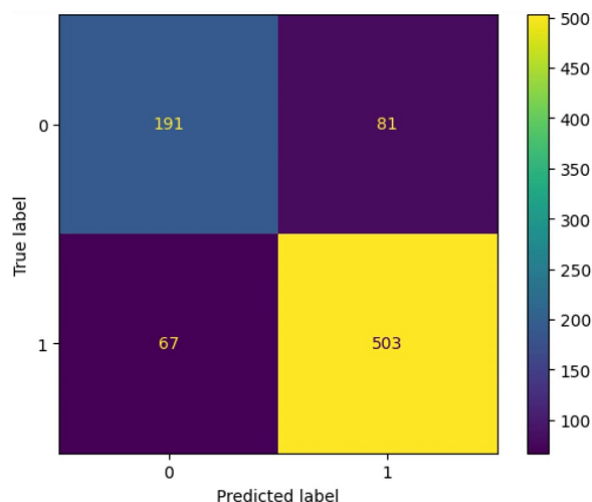

Figure 4: EN-IT Confusion Matrix of LLMB2PEN, *curie* model: Clockwise from top-left corner (TN, FP, TP, FN)

gories, 67 sentences need to be edited but predicted as good, and 81 translated sentences do not need to be edited, but the prediction says they have to be reviewed.

In the English-German set from Figure 5, the situation is the opposite: there are more translated sentences that do not need to be edited (442) than prescribed for review (256) in the correct predictions. In the wrong prediction categories, such numbers are 90 and 46 respectively.

The prediction **accuracy** of the LLMB2PEN model on our designed task is **(TP+TN)/Total** = (503+191)/842 = 82.42% for English-Italian MT, and (442+256)/834 = 83.69% for English-German MT. Overall, our LLMB2PEN method shows that the English-German output is clearly better than the English-Italian.

However, if we only count the Type II errors (incorrect prediction that the segments should NOT be edited), then the corresponding error rates will be 67/842 = 8% for Italian and 90/834 = 10% for German.

3.2 Discussion

The first and foremost finding is that the fine-tuned model learned enough information to make a very significant prediction of whether the segment has to be edited or not. It should be noted that such successful classification holds the promise of a viable method to significantly reduce the volume of post-editing efforts and therefore time and costs. There is, however, a problem: while it is OK to
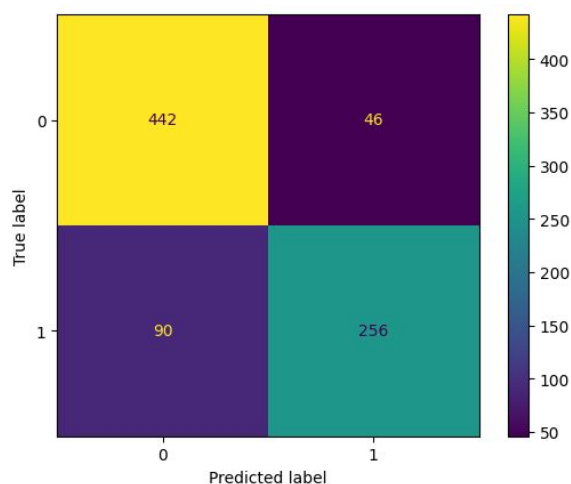
Figure 5: EN-DE Confusion Matrix of LLMB2PEN, *curie* model: Clockwise from top-left corner (TN, FP, TP, FN)

present the editor with segments that are predicted as required for editing, but in reality do not require editing (the fourth quadrant, FP), real consumer risk comes from the segments that have been predicted as not requiring editing and made their way to the final predict, but in reality, they contain errors (the fourth quadrant, FN).

Such segments represent a significant portion of segments predicted as not requiring post-editing: **FN/(TN+FN)** = 67/(191+67) = 67/258 = 26% of "leave as is" (let's call them "LAI") segments for Italian, and 90/(442+90) = 90/532 = 16.9% for German.

It is possible that for specific language pairs and MT engines the portion of the LAI segments will decrease with the size increase of the training data and further fine-tuning, but it is unlikely to become zero, since with neural models the error rate is never zero.

Two strategies can be considered for implementing such prediction in production:

1. The LAI segments are excluded from the human loop and go into publication unvetted, but not straight away as they advance through the workflow along with all the other segments. In this scenario, the potential error rate ceiling for final content will be **FN/Total** = FN/(TP+FN+TN+FP) = 8% for Italian, i.e. 67/(81 + 67 + 191 + 503) = 81/842 and 10.8% = 90 / (90 + 46 + 442 + 256) = 90/834 for German.

It is not impossible to predict what would be the actual error rate in those 8% and 10.8% segments that will not be reviewed or the severity of errors in them. It is, obviously, the decision of the customer to decide whether this is an acceptable level of consumer risk for their situation (domain, type of content, audience, etc.). Additional risk assessment may be required to be carried out.

The savings on post-editing volume in this scenario would be (TN+FN)/Total = (191+67)/842 = 30.1% for Italian and (442+90)/834 = 63.8% for German.

2. All LAI segments are marked as "100% MT matches" in a CAT tool. With this approach, translators are requested to review them, but at a lower per-word rate, using the traditional approach which is well familiar to translation providers. In this scenario the reduction of the total time, effort, and cost can be estimated as follows: without this approach, translators working on the Edit Distance Calculation (EDC) model will get lower payment (which can vary from 10% to 40% with different payment models) for not changed segments. In this scenario, translators may be asked to review such LAI segments but paid only a small part of the full rate for the review of such segments.

Simple proportion allows us to calculate the savings in the second scenario: if we take the full payment for all the segments for 100% of post-editing costs, and assume that 10% pay reflects adequate pay for the review of LAI segments that are marked as such, the volume of post-editing decreases 27.6% for Italian and 57.4% for German with zero error rate of the final product (no producer's or consumer's risk).

This estimate of a potential economy with a guarantee of zero error rate begs for further research and implementation of this method.

4 Extended Experiments On Six More Language Pairs

We hereby also present extended experimental results using six more language pairs obtained with LMB2PEN method for translation editing distance prediction. These language pairs include English-to-French, Japanese, Dutch, Portuguese, Turkish, and Chinese (EN→FR/JA/NL/PT/TR/ZH), whose

results are listed in Figure 6, 7, 8, 9, 10, and 11 respectively.

From the results presented in the figures, in general, the ratio of correct prediction (TP+TN) is much higher than the one from mis-prediction (FN+FP) across all these language pairs, as for English-Italian and English-German in the pilot studies. On one hand, the following language pairs have more True Positive than True Negative predicted segments **than for English-German/Italian**: English-Japanese, English-Portuguese, and English-Chinese. On the other hand, the rest of the language pairs have more TN than TP: English-French, and English-Dutch, except for English-Turkish which has a comparable number of segments between TP (347) and TN (353) labels. This finding also indicates that such language pairs with a high number of TN labels are still much more challenging for MT system development to produce more correct outputs, i.e., English to French, Dutch, and Turkish. Earlier research findings from Gladkoff et al. (2022) on TQE conclude that 200+ segments can be enough amount of data to reflect the MT system quality.
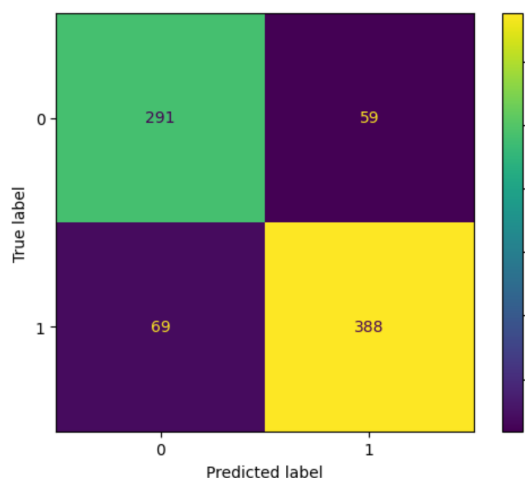

Figure 6: EN-FR Confusion Matrix of LLMB2PEN, *curie* model: Clockwise from top-left corner (TN, FP, TP, FN)

5 Different LLMs on EN-JA News Domain

In the subsequent experiment on data, we used different news items translation corpus from different projects, translated from English to Japanese.
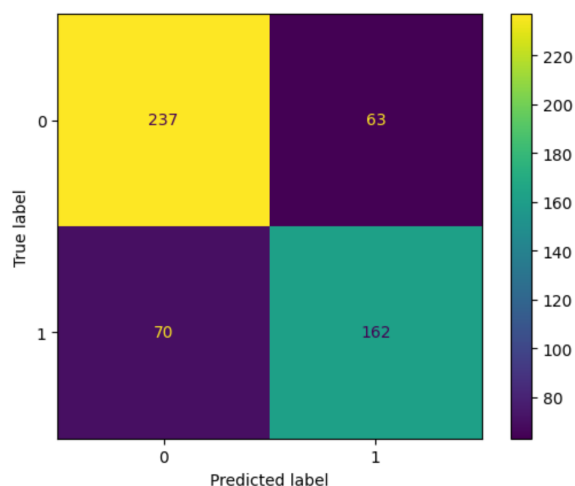

Figure 7: EN-JA Confusion Matrix of LLMB2PEN, *curie* model: Clockwise from left-up corner (TN, FP, TP, FN)
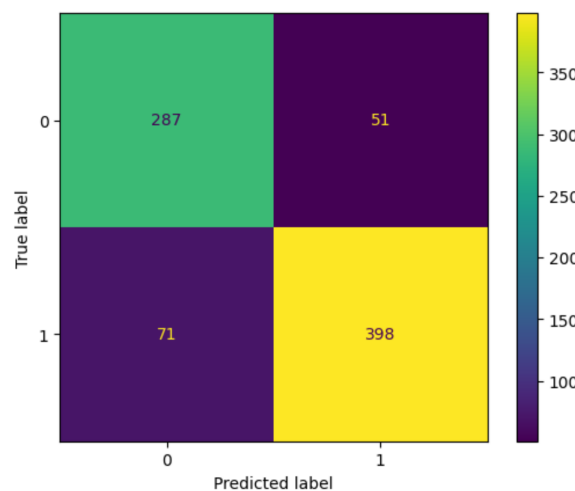

Figure 8: EN-NL Confusion Matrix of LLMB2PEN, *curie* model: Clockwise from top-left corner (TN, FP, TP, FN)

5.1 Using OpenAI GPT3.5turbo

In this experiment, we have repeated experiments of fine-tuning the OpenAI **gpt3.5turbo** model on datasets of different sizes: 2000 pairs, 4000 pairs, and 6000 pairs.

Figure 12 shows the confusion matrix for the training set of 6000 bilingual EN-JA translation pairs in the news domain.

We ran several experiments with varying training set sizes, with results shown in Figure 13.

These results are interesting because although False Positive prediction does not improve with the increase of training set, in the context of the
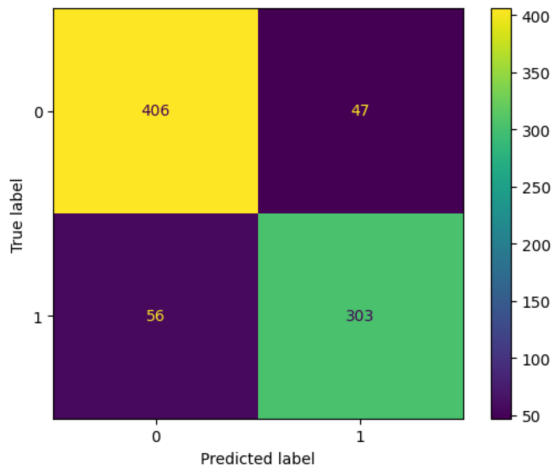
Figure 9: EN-PT Confusion Matrix of LLMB2PEN, *curie* model: Clockwise from left-up corner (TN, FP, TP, FN)
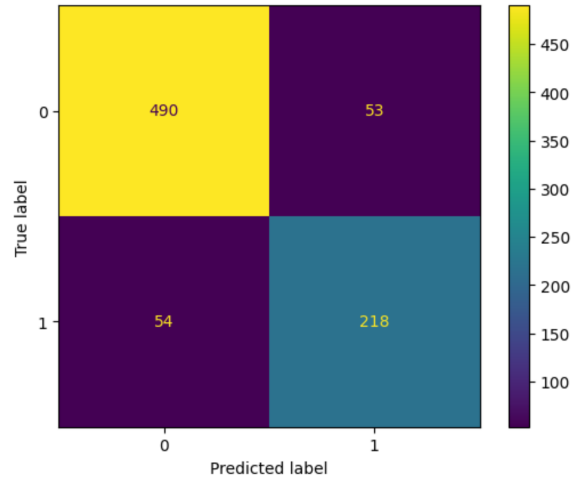

Figure 11: EN-ZH Confusion Matrix of LLMB2PEN, *curie* model: Clockwise from top-left corner (TN, FP, TP, FN)
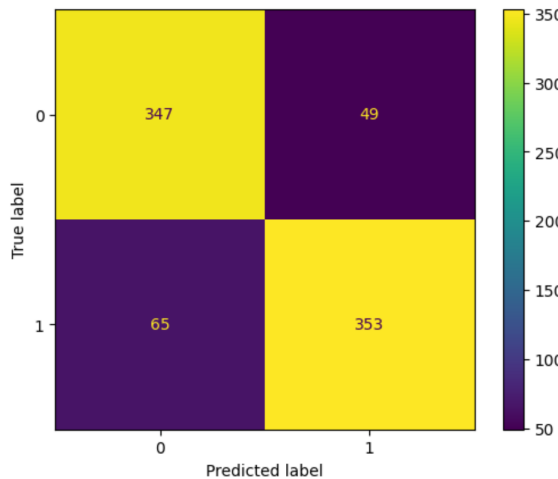

Figure 10: EN-TR Confusion Matrix of LLMB2PEN, *curie* model: Clockwise from top-left corner (TN, FP, TP, FN)
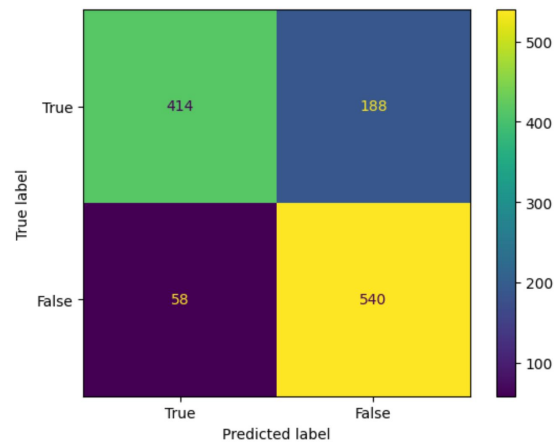

Figure 12: EN-JA news items Confusion Matrix of LLMB2PEN, *gpt3.5turbo* model: Clockwise from top-left corner (TP, FN, TN, FP)

Train	Test	Total	Legth groups	FN	FN, % total	FP	FP, % total
2000	400	2400	10	79	19.75%	17	4.25%
2000	400	2400	15	74	18.50%	23	5.75%
2000	400	2400	5	74	18.50%	18	4.50%
4000	800	4800	5	162	19.00%	16	2.00%
4000	800	4800	10	102	12.75%	75	9.38%
6000	1200	7200	10	188	15.67%	58	4.83%

Figure 13: EN-JA news items predictions with fine-tuning completed on different training dataset sizes, *gpt3.5turbo* model

need for post-editing the False Negative category is much more interesting, because we are interested in better prediction of those segments which do NOT require post-editing. And, as we see from the experimental data, the prediction of FN improves from almost 20% to 12%-15% with the increase of training set from 2000 bilingual segments to 6000 bilingual segments.

We, therefore, can recommend the training set in that range, since larger sizes of training set will be more expensive and will take significant time for models with the size of *gpt3.5turbo*.

5.2 Comparison of performance on different OpenAI models

It was also interesting to see how the extra-large LLMs (xLLMs) from OpenAI, the *davinci* and *gpt3.5turbo* models, perform on the same task in

comparison to *curie* model we used earlier. These three LLMs have parameter sizes around 13B, 175B, and 175B respectively.

So we used the same English-Italian data from our original experiment to compare performance on different models of the same EN-IT dataset.

Figure 14 shows the comparison of these three LLMs regarding their confusion matrix and parameter sets. Surprisingly, their performances on predicting MT errors are very close, i.e. the larger-sized *davinci* model and extra-large sized *gpt3.5turbo* did not demonstrate much improvement on model classification accuracy. Their correct labels (TP+TN) are (694, 699, 706) respectively out of 842 all labels, which results in the accuracy ratios 82.42%, 83.02%, and 83.85%. In comparison to the much smaller *curie* model with 12 layers of Transformer and 768 hidden units, the xLLM *gpt3.5turbo* only achieved 1.43 points (83.85%-82.42%) increase of accuracy score despite using 175 layers of Transformer and 4096 hidden units.

The explanation for this may probably be found because the fine-tuning loss on this classification task drops down very quickly.

Figure 15 shows the fine-tuning loss on the *gpt3.5turbo* model. As can be seen from this graph, only 100 steps are sufficient to bring the loss to almost zero, and then all other steps contribute very little to the classification quality improvement.

As we can see, there is no need to use larger models since results hardly improve as compared with *curie* model.

6 Conclusions and Future Work

In this work, to investigate the LLM's capability of predicting MT output errors, we fine-tuned GPT models via OpenAI API. We formulated the task as a classification challenge using prepared historical post-editing data on English-Italian and English-German for pilot studies. The experimental output using fine-tuned LLMB2PEN demonstrated promising results. We also analysed the possible solutions for addressing the error rates, i.e. whether prediction errors can be ignored and published without the review, or letting them be reviewed by the linguists at a lower rate, and how much saving can be achieved for the client who uses this process, in comparison to 100% post-editing without using LLMB2PEN method.

In the extended experiments, we added six more language pairs including English-to-French, Japanese, Dutch, Portuguese, Turkish, and Chinese, in total resulting in eight, and summarised our findings by classifying the language pairs. We also compared GPT models from different sizes and the experimental results surprisingly show that the larger LLMs (*davinci* and *gpt3.5turbo*) do not improve the accuracy performance of much smaller *curie* model with apparent margins but with much more cost.

In the future, we are going to work on response rate and training times to see whether the model can continue learning as *being fed with more consecutive chunks of data* for the same languages, to implement an ongoing learning of prediction. In addition, we plan to carry out the LLMB2PEN fine-tuning on other language pairs for which we have historical data. We intend to explore to what extent the model is capable of absorbing data for several languages, i.e. one fine-tuned multilingual model serving several language pairs.

To further extend this project, it will also be interesting to explore and check whether the LLMB2PEN method can help to identify human-introduced errors or translationese.

Limitations

In this work, we reported MT QE experiments using eight language data translated from English. The positive results produced from the OpenAI models can be further enhanced by more language pairs, as well as broader domains of the corpus.

The main limitation of the method is non-zero fine-tuning time. The fine-tuning takes about 20 minutes and therefore cannot be made continuous, which has to be done periodically, in batches. This hardly can be overcome, but deployment methods can be applied to quickly replace the older fine-tuned models with the newer ones.

Ethical Statement

This work has no ethical concerns since we did not disclose any identifiable private user data. All experiments were carried out in a secure computing environment.

Acknowledgements

		Curie			DaVinci			ChatGPT 3.5 Turbo	
	True Label								
FALSE	0	191	81		201	68		206	63
TRUE	1	67	503		68	505		80	493
	Predicted Label	0	1						
		Total	842			842			842

Parameter	Curie	DaVinci	GPT-3 (ChatGPT)
Architecture	Transformer	Transformer	Transformer
Layers	12	24	175
Hidden Units	768	1024	4096
Parameters	~13B	~175B	~175B
Training Data	Web Text	Web Text	Web Text
Use Case	Specialized tasks like translation, summarization	General-purpose, wide range of tasks	General-purpose, wide range of tasks
Released	2019	2020	2020

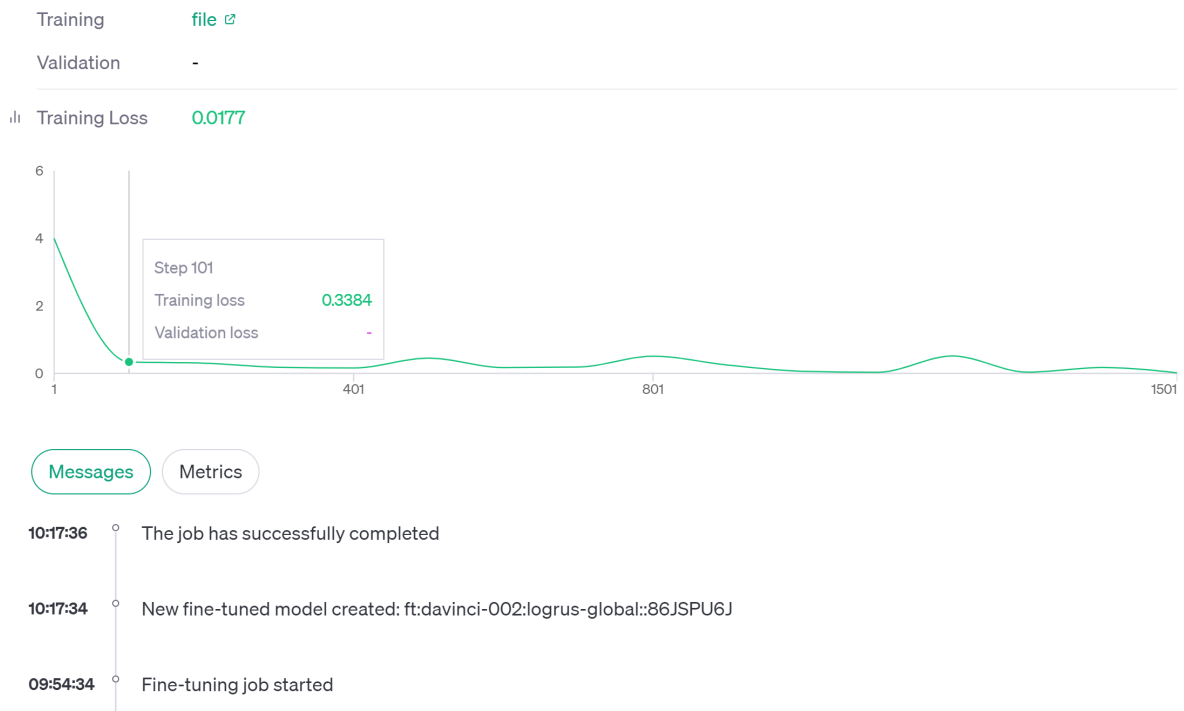Figure 14: Comparisons of three LLMs on Confusion Matrix and Parameters.

Training	file
Validation	-
Training Loss	0.0177

Step 101
Training loss   0.3384
Validation loss   -

(Messages)   (Metrics)

10:17:36	The job has successfully completed
10:17:34	New fine-tuned model created: ft:davinci-002:logrus-global::86JSPU6J
09:54:34	Fine-tuning job started

Figure 15: Fine-tuning progress on *gpt3.5turbo* model fine-tuning.

References

Keqin Bao, Yu Wan, Dayiheng Liu, Baosong Yang, Wenqiang Lei, Xiangnan He, Derek F. Wong, and Jun Xie. 2022. Alibaba-translate China's submission for WMT 2022 quality estimation shared task. In *Proceedings of the Seventh Conference on*

Machine Translation (WMT), pages 597–605, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, editors. 2012. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2022. KU X upstage's submission for the WMT22 quality estimation: Critical error detection shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 606–614, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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. *arXiv e-prints*, page arXiv:2104.14478.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

Serge Gladkoff, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. 2022. Measuring uncertainty in translation quality evaluation (TQE). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1454–1461, Marseille, France. European Language Resources Association.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1183–1191.

Aaron Li-Feng Han, Yi Lu, Derek F. Wong, Lidia S. Chao, Liangye He, and Junwen Xing. 2013. Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 365–372, Sofia,

Bulgaria. Association for Computational Linguistics.

Lifeng Han. 2022a. *An investigation into multi-word expressions in machine translation*. Ph.D. thesis, Dublin City University.

Lifeng Han. 2022b. An overview on machine translation evaluation. *arXiv preprint arXiv:2202.11027*.

Lifeng Han and Serge Gladkoff. 2022. Meta-evaluation of translation evaluation methods: a systematic up-to-date overview. In *Tutorial at LREC2022*, Marseille, France.

Kung Yin Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024. Cantonmt: Cantonese to english nmt platform with fine-tuned models using synthetic back-translation data.

Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors. 2022. *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

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. 2022. 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.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.