Improving Arabic-English Translation for Humanitarian Response Efforts via Open LLMs and In-Context Learning

Ying Tang[1], **Ezzeldeen Al Sheikh Khalil**[2], **Samia Ismael Skaik**[3]

[1]SAIL Initiatives, an unfunded volunteer-led collective
[2]Faculty of Engineering and Intelligent Systems, University College of Applied Sciences
[3]Ministry of Education & Higher Education, Palestine

Abstract

A key step in humanitarian relief in low-resource settings is making translation tools easily accessible. Once validated, these tools enable improved access to educational materials, healthcare information, and other essential resources. Many existing Arabic-to-English translation services require Internet access or paid subscriptions, while non-proprietary approaches typically require powerful computers, which are often infeasible for communities with limited or unreliable connectivity and electricity. This study advocates a non-proprietary approach based on open-weight large language models and *in-context learning*, a strategy that enables these models to learn from a few examples without expensive retraining of models. We tested various open-weight models, including Meta's LLaMA3.3, Google's Gemma2, and Alibaba's Qwen2.5, to evaluate their Arabic-to-English translation performance. According to various quantitative metrics, our experimental results show that using 3 to 15 examples progressively enhanced translations accuracy, ***and*** that using the entire training corpus to fine-tune commonly used models did not yield performance gains. Additional subjective evaluations by native speakers revealed limitations that may be addressed by including examples of idiomatic expressions and other colloquial data. By identifying effective and lightweight translation tools, this work contributes to the development of digital tools that can support long-term recovery and resilience-building efforts in Gaza. To enable community replication and ongoing verification, our analysis scripts will be continuously updated and made available at `https://anonymous.4open.science/r/openllm4SPEAK`.

1 Introduction & Background

Despite the ceasefire announced in October 2025, the humanitarian situation in Gaza remains dire; the collapse of healthcare, education, and civil infrastructure has inflicted lasting harm including profound lifelong and inter-generational trauma (Jamaluddine et al., 2025; Mahamid & Bdier, 2025). While the international community worked toward conflict resolution (Hermawanto et al., 2025), grassroots networks and professional actors raised urgent calls for action, including appeals by more than 40 prominent experts (London et al., 2024), 900 medical professionals from Australia (Devi, 2024; Ge, 2025), and legal experts from the United Kingdom (Meagher, 2023). Collectively, these voices underscore the urgent need for complementary approaches that can support local humanitarian efforts remotely, particularly those capable of transcending geographic and political barriers via development of mobile applications (Yusoff et al., 2024).

Motivated by this need, our group of unfunded volunteers united with Palestinian partners to co-develop solutions, including the development of new assistive technologies for remote learning (Shraim & Crompton, 2020) and mental health care (Sweidan et al., 2024).

39th Conference on Neural Information Processing Systems 2025Muslims in MLWorkshop (peer reviewed).

A foundational requirement for such tools is reliable, automated Arabic-to-English translation, particularly systems adapted to the Palestinian dialect. Language barriers remain a significant challenge in humanitarian response, limiting access to critical information, educational resources, and mental health support. Although some studies report that ChatGPT outperforms Google Translate for Arabic (Cahyaningrum, 2024; Mohsen, 2024), many organizations have turned to DeepSeek because delivers comparable functionality at roughly 50-fold lower cost (Grey, 2025).

However, most existing Arabic-to-English translation services either require stable Internet access or paid subscriptions, while many open-source approaches depend on high-performance computing resources, both of which remain inaccessible for communities experiencing power outages and infrastructure collapse.

To bridge these gaps, this study proposes a non-proprietary solution built on open-weight large language models (LLMs) and an in-context learning strategy that enables high-quality translation from only a handful of examples, eliminating the need for expensive model retraining. Our results show that LLMs augmented through in-context learning deliver translation performance comparable to, and often not inferior to, models fine-tuned on far larger datasets. This points to a practical path forward for humanitarian technologies that require minimal download sizes and can function reliably despite Gaza's ongoing Internet connectivity constraints, even after the October 2025 ceasefire.

2 Methods

2.1 Datasets

We draw on three publicly available data sources in our analysis. 1) the **ATHAR dataset** includes materials from the domains of culture, philosophy, and science (Khalil & Sabry, 2024) and *has undergone manual verification by its creators*; 2) a subset of the **HuggingFace dataset** (Mor-Lan, 2024) includes samples written in the Palestinian dialect; 3) the **ArBanking dataset** (Jarrar et al., 2023) that centers on banking and financial services.

As noted by the authors (Khalil & Sabry, 2024), ATHAR focuses on classical Arabic, which significantly differs from Modern Standard Arabic (MSA) in both syntax and semantics. The HuggingFace subset includes 1,375 colloquial Palestinian examples, which were translated into English using a proprietary language model (Claude Sonnet 3.5). Using a set of manually engineered heuristics, the translations were subsequently *corrected* by an independent researcher external to our team. Lastly, the ArBanking dataset provides a consistent, domain-specific benchmark for evaluating model robustness on transactional Arabic, that is, short, task-oriented service requests typical of customer-service interactions.

2.2 Models

As ChatGPT is proprietary to OpenAI where costs are determined by usage, no-cost alternatives are needed. Further, recent benchmarks demonstrate that GPT4o and LLaMA3 are comparatively accurate, with no statistical difference in performance measures (Li et al., 2025). Accordingly, in addition to DeepSeek-R1 (Guo et al., 2025), we shortlisted the following open-weight models in our evaluation benchmark based on their performance in recent benchmarks: Phi4:14B (by Microsoft), Marco-o1:7B (by Alibaba International Digital Commerce), Qwen2.5 (by Alibaba Cloud), Mannix/Gemma2-9B (Google), and LLaMA3 (LLaMA3.1:8B and LLaMA3.3:70B by Meta).

Model weights were downloaded through Ollama (`https://ollama.com/`). Inferences were executed by deploying Ollama server locally in a clustered computing environment.

Study	Model	METEOR	chrF	H-BLEU	S-BLEU	S-chrF	ROUGE-L
Ours	Phi4:14B	0.298	0.286	0.074	0.093	2.931	0.337
	Qwen2.5:7B	0.336	0.372	0.103	0.145	3.078	0.368
	DeepSeek-R1:70B	0.345	0.378	0.118	0.087	3.502	0.394
	LLaMA3.3:70B	0.414	0.434	0.161	0.137	2.970	0.449
Ref.	LLaMA3.3:70B*	0.342	-	-	0.130	-	0.413
	GPT-4o*	0.357	-	-	0.147	-	0.441

Table 1: Evaluation on the ***entire*** test set of ATHAR involving modern standard Arabic (n=1,000). For reference, we included results from a prior study by (Khalil & Sabry, 2024) (we could not replicate due to constraints on time and cost to purchase subscriptions). We used the exact same test set and implementations of the evaluation metrics.

Model	METEOR	chrF	S-chrF	H-BLEU	S-BLEU	ROUGE-L	BERT-Score
DeepSeek-R1:8B	0.33±0.24	32.84±20.28	0.11±0.08	0.06±0.16	0.01±0.00	0.35±0.24	0.89±0.05
Meta's LLaMA3.1:8B	0.46±0.26	41.60±22.94	0.05±0.07	0.13±0.23	0.60±0.00	0.47±0.26	0.91±0.05
Microsoft's Phi4:14B	0.48±0.27	44.01±24.27	0.12±0.07	0.16±0.27	0.08±0.00	0.48±0.27	0.92±0.05
Alibaba's Qwen2.5:7B	0.56±0.25	50.43±23.53	0.05±0.07	0.19±027	0.11±0.00	0.59±0.25	0.93±0.04
Google's Gemma2-9B	0.61±0.26	54.68±24.58	0.07±0.08	0.25±0.31	0.06±0.00	0.62±0.25	0.94±0.04
Meta's LLaMA3.3:70B	0.61±0.26	55.86±24.84	0.09±0.08	0.27±0.32	0.08±0.00	0.62±0.25	0.94±0.04

Table 2: Evaluation on the Huggingface dataset (Mor-Lan, 2024) that consists of sentences written in the Palestinian dialect (n=1,325).

2.3 In-context learning

In-Context Learning (ICL) refers to the strategy of providing labeled examples within a prompt to guide an LLM in producing desired outputs (Dong et al., 2024). In this work, we employ a small subset of the training set of the ATHAR dataset (Khalil & Sabry, 2024) to serve as in-context demonstrations. To ensure evaluation integrity and prevent data leakage, we excluded these examples from the test set. An example prompt is provided in Appendix A. All analysis scripts were prepared and executed in Python 3.10.

3 Results

3.1 Quantitative evaluation

Performance evaluation using the ATHAR dataset is presented in Table 1 where LLaMA3.3:70B is shown with the top performance, even with as few as five translation examples, according to the METEOR metric (definition in Appendix B). For reference only (as both studies used the same test set and metric calculators but different compute environments), we included results compiled by Khalil & Sabry (2024) who suggested that LLaMA3.3:70B with 5 translation examples may yield results superior to those generated by GPT-4o.

Results of performance evaluation using the second dataset is presented in Table 2 where LLaMA3.3:70B is also shown to be the top performing model. More comparative analysis on the performance of fine-tuned models and ablation studies on sample-selection methods are provided in Appendix's Table 3 and Table 4, respectively.

3.2 Subjective evaluation

Since our group relies solely on volunteers' hours, we do not have the capacity to employ professional evaluators for large-scale translation assessment. Instead, we conducted five evaluation sessions with four native Palestinians, three of whom hold Doctor of Philosophy degrees. Examples of annotated analysis can be found in Appendix D. In summary, we identified the following types of recurring errors.

First, there was a consistent issue with the translation of polite requests being rendered as direct commands, which can significantly alter the tone of the original message. Such shifts can have important pragmatic consequences, especially in culturally sensitive or service-oriented contexts where indirectness is expected.

Second, domain-specific terminology was often mistranslated; an illustrative example involves the best model's inability to discriminate between "disposable" and "throw away," a distinction not fully captured even by the human evaluator. This highlights the challenges of conveying subtle semantic distinctions in specialized registers.

Finally, there was a recurring loss of precision in translating prepositional phrases, particularly in distinguishing between ownership and association, a subtle yet critical distinction in linguistic semantics. These relational ambiguities can significantly affect interpretation, particularly in legal, academic, or instructional texts where clarity about responsibility or possession is essential.

4 Discussions

Our findings reinforce recent observations that ICL provides a practical alternative to fine-tuning models, particularly in resource-constrained and crisis-response scenarios. Conversely, DeepSeek-R1 was previously praised for reasoning tasks (Guo et al., 2025), our results show it gave performance

comparable to that of GPT-4o for MSA. However, it suffered a performance gap in its translation capability for the Palestinian dialect that contains culturally nuanced and idiomatic expressions. Based on our experiences in a high performance compute environment, it also struggled to scale efficiently, requiring longer inference times and demonstrating input truncation under high-shot prompting schemes. These challenges limit its current practicality in real-time or mobile deployment.

In contrast, Google's Gemma2-9B balanced speed, accuracy, and dialect robustness. Its strong performance under few-shot settings (including Palestinian Arabic) makes it a viable candidate for embedded applications in humanitarian contexts. Compared to LLaMA-3.3-70B, which achieved the highest scores but requires downloading large model weight files that is often impractical in settings with limited or intermittent internet connectivity, Gemma-2-9B offers a more accessible alternative.

Finally, our qualitative results involving text written in the Palestinian dialects show that models vary significantly in their handling of idiomatic expressions and tone. Gemma2-9B occasionally outperformed larger models on short conversational utterances, while LLaMA3.3:70B better preserved formal structure.

Based on human review and detailed annotations of the errors made by the best model, we conjecture that some of the errors may be resolved in future development work. More specifically, in many languages, especially those with rich case systems or flexible syntactic structures, prepositions or their equivalents serve to encode a wide range of relationships between entities. As a result, when translating into a target language with less grammatical transparency can become blurred (such as English, where "of" may ambiguously denote either possession or association—these distinctions). Consequently, we are exploring how ICL can be leveraged further to reduce these errors, particularly by incorporating domain-specific prompts and disambiguated training data that foreground the ownership–association distinction. Additionally, more robust evaluation methods may be developed to assess how well LLMs preserve relational semantics across different language pairs and genres.

5 Conclusion

This preliminary study demonstrates that open-weight LLMs, particularly Qwen2.5:7B and LLaMA3.3:70B, can produce high-quality Arabic-English translations, including for dialectal and culturally embedded expressions. Compared to LLaMA3.3:70B that had the highest memory requirements in our benchmark, Gemma9:2B and Qwen2.5:7B showed promising results in handling translations of the Palestinian dialect. These lighter-weight models demonstrate greater potential for practical deployment in domain-specific humanitarian mobile applications, where speed, reliability, and cultural sensitivity are critical.

We also found that variations in the setup of in-context learning (e.g. prompt wording and number of examples shown) can affect translation accuracy in some models, although this challenge can be mitigated through prompt reconfigurations. Additionally, when appropriately prompted, this approach can provide interpretable reasoning traces that may support subsequent analysis (e.g. for translators to train human and/or LLMs).

Overall, our results show that in-context learning is a cost-effective alternative to fine-tuning models, making it especially valuable in low-resource settings. We truly hope future researchers can build on our work by exploring adaptive prompting, hybrid ICL fine-tuning strategies, and task-specific dialect modeling. In doing so, the research community can meaningfully contribute to equitable technological development and support communities facing linguistic, infrastructural, and geopolitical marginalization.

ACKNOWLEDGMENTS: Dr. Ying expresses sincere gratitude to Prof. Ali Hindi, Alifia Nur Azzizah, Amanda M. P., Amanda Resmana, Dr. Asmaa Abusamra, Annisa N., Argya Zahra Rapul Hanisi, Arief Naufal Pramudito, Mr. Begbie, Chung Wai Tang, Diah Paramitha, Dzaki Dwitama, Fikha A., Dr. Gerrit Krueper, Dr. Hanene Y., Dr. Inria Astari Zahra, Irfani Aura Salsabila, Prof. Ismail Khater, Prof. Mila Zuo, Nagham Ahmed Skaik, Prof. Rania Qassrawi, Rizqia V., Sarah Aisha, Savira Aristi, Dr. Sieun L., Tia Dwi Setiani, Wafaa Hijazi, and Yoonha J. for their support that helped sustained this work. Ezzeldeen would like to thank his parents and Profs. Alaa Al-Qazzaz and Ismail Khater for their continuous support. Dr. Samia expresses sincere gratitude to all members at the June workshops organized by LabTek Indie and SAIL Initiatives for offering her opportunities to share.

References

Ika Oktaria Cahyaningrum. Chat GPT vs Google Translate for Translation. In *The 3 International Symposium on The Practice of Coexistence in Islamic Culture*, pp. 450, 2024.

Sharmila Devi. Calls for Gaza ceasefire to tackle poliovirus. *The Lancet*, 404(10455):837, 2024.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, 2024.

Yipeng Ge. Palestine is freeing us all before palestine is free; comment on "the rhetoric of decolonizing global health fails to address the reality of settler colonialism: Gaza as a case in point". *International Journal of Health Policy and Management*, 14(1):1–3, 2025.

Grey. Experts Weigh In on DeepSeek AI Translation Quality. https://slator.com/experts-weigh-in-on-deepseek-ai-translation-quality/, 2025. Accessed: 2025-02-17.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Ariesani Hermawanto, Sucahyo Heriningsih, et al. Veto and the UN Security Council's Failure to Resolve the Israeli-Palestinian Conflict. In *RSF Conference Series: Business, Management and Social Sciences*, volume 5, pp. 146. Research Synergy Foundation, 2025.

Zeina Jamaluddine, Hanan Abukmail, Sarah Aly, Oona MR Campbell, and Francesco Checchi. Traumatic injury mortality in the gaza strip from oct 7, 2023, to june 30, 2024: a capture–recapture analysis. *The Lancet*, 2025.

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. Arbanking77: Intent detection neural model and a new dataset in modern and dialectical arabic. *arXiv preprint arXiv:2310.19034*, 2023.

Mohammed Khalil and Mohammed Sabry. Athar: A high-quality and diverse dataset for classical arabic to english translation. *arXiv preprint arXiv:2407.19835*, 2024.

Abdullah Salem Khered, Youcef Benkhedda, and Riza Theresa Batista-Navarro. Dial2msa-verified: A multi-dialect arabic social media dataset for neural machine translation to modern standard arabic. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pp. 50–62, 2025.

David Li, Kartik Gupta, Mousumi Bhaduri, Paul Sathiadoss, Sahir Bhatnagar, and Jaron Chong. Comparative diagnostic accuracy of GPT-4o and LLaMA 3-70b: Proprietary vs. open-source large language models in radiology. *Clinical Imaging*, 118:110382, 2025.

Leslie London, Andrew Watterson, Donna Mergler, Maria Albin, Federico Andrade-Rivas, Agostino Di Ciaula, Pietro Comba, Fernanda Giannasi, Rima R Habib, Alastair Hay, et al. A call from 40 public health scientists for an end to the continuing humanitarian and environmental catastrophe in gaza. *Environmental health*, 23(1):59, 2024.

Fayez Mahamid and Dana Bdier. Intergenerational Transmission of Traumatic Experiences among Palestinian Refugees. In *Intergenerational Trauma in Refugee Communities*, pp. 40–52. Routledge, 2025.

Kate Meagher. Protest, propaganda and politics: media coverage of the london ceasefire marches. *International Development*, 2023.

Mohammed Mohsen. Artificial Intelligence in Academic Translation: A Comparative Study of Large Language Models and Google Translate. *Psycholinguistics*, 35(2):134–156, 2024.

Guy Mor-Lan. Levanti: A Levantine Arabic Dataset, 2024. URL https://huggingface.co/datasets/guymorlan/levanti.

Aziz Mohammed Abdo Saeed. Machine translation evaluation between arabic and english during 2020 to 2024: A review study. *Arts for Linguistic & Literary Studies*, 7(2), 2025.

Farzan Saeedi, Ghaniya Al Hinai, Khoula Al Kharusi, and Abdulrahman AAl Abdulsalam. Effect of Context and Tokenization on Machine Translation of Arabic Conversations on Social Media. *Procedia Computer Science*, 258:1757–1763, 2025.

Khitam Shraim and Helen Crompton. The use of technology to continue learning in Palestine disrupted with COVID-19. *Asian Journal of Distance Education*, 15(2):1–20, 2020.

Saadeh Z Sweidan, Shyam K Almawajdeh, Ayah M Khawaldeh, and Khalid A Darabkh. MOLHEM: An innovative android application with an interactive avatar-based chatbot for Arab children with ASD. *Education and Information Technologies*, pp. 1–35, 2024.

Mohd Fitri Yusoff, Juliana Aida Abu Bakar, and Ruzinoor Che Mat. Guidelines for developers when developing an Islamic mobile app: a conceptual framework. *Multidisciplinary Science Journal*, 6 (11):2024162–2024162, 2024.

A Example prompt

A prompt may be structured as follows:

> Translate the given input text from Arabic dialect to English. Examples are listed below, each paired with its correct translations. Mark your translations with "Output". Do not explain your answer. Example: ...

B Evaluation metrics

To evaluate translation quality, we employed the following metrics:

1. BLEU (Bilingual Evaluation Understudy) measures word n-gram overlap between the translation and reference text, with penalties for length differences;
2. chrF (Character n-gram F-score) measures the F1-score using character n-grams instead of words, and thus handles morphologically rich languages and partial word matches better than BLEU (and less sensitive to word order than BLEU);
3. METEOR (Metric for Evaluation of Translation with Explicit ORdering) considers word order and uses synonymy, stemming, and paraphrase-matching. Because METEOR considers synonyms, it correlates better with human judgment than BLEU;
4. ROUGE-L measures the length of the longest common subsequence, which is some times used for evaluating the fluency of the text;
5. BERT-score computes the semantic similarity between text pairs using contextual embeddings from a pre-trained language model called Roberta-large.

B.1 BLEU

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where:

- p_n is the modified n-gram precision,
- w_n is the weight for the n-gram (typically $w_n = \frac{1}{N}$),
- N is the maximum n-gram length (often 4),
- BP is the brevity penalty.

The brevity penalty (BP) is defined as:

$$BP = \{\, 1 \; if c > r \exp\left(1 - \frac{r}{c}\right) if c \leq r$$

Where:

- c is the length of the candidate translation,
- r is the effective reference length (e.g., closest or average).

B.2 chrF

$$chrF_\beta = \left(1 + \beta^2\right) \cdot \frac{P_n \cdot R_n}{\beta^2 \cdot P_n + R_n}$$

Where:

- $P_n = \frac{\sum_{k=1}^{n} p_k}{n}$ is the average character n-gram precision up to order n,
- $R_n = \frac{\sum_{k=1}^{n} r_k}{n}$ is the average character n-gram recall up to order n,
- $p_k = \frac{\#of-matching-characterk-grams}{ofcharacterk-gramsincandidate}$,
- $r_k = \frac{\#ofmatching-characterk-grams}{\#ofcharacterk-gramsinreference}$,
- β is a weighting factor (commonly set to $\beta = 2$ to emphasize recall).

B.3 METEOR

$$METEOR = F_{mean} \cdot (1 - Penalty)$$

Where:

$$F_{mean} = \frac{10 \cdot P \cdot R}{R + 9P}$$

- $P = \frac{m}{w_t}$ is the unigram precision,
- $R = \frac{m}{w_r}$ is the unigram recall,
- m is the number of matched unigrams,
- w_t is the number of unigrams in the candidate,
- w_r is the number of unigrams in the reference.

The penalty is computed as:

$$Penalty = \gamma \left(\frac{ch}{m} \right)^{\beta}$$

- ch is the number of chunks (consecutive matched unigrams),
- γ and β are tunable parameters (commonly $\gamma = 0.5$, $\beta = 3.0$).

B.4 ROUGE-L

$$ROUGE - L = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{Recall + \beta^2 \cdot Precision}$$

Where:

- LCS is the length of the longest common subsequence between candidate and reference,
- $Precision = \frac{LCS}{lengthofcandidate}$,
- $Recall = \frac{LCS}{lengthofreference}$,
- β is usually set to $\beta = \frac{Precision}{Recall}$ to emphasize recall.

In summary, the BERT-score captures word meaning by aligning tokens based on cosine similarity in the embedding space. Because words are not compared literally but semantically, BERT-score often correlates better with human judgment. We kept the other evaluation metrics for cross-reference with previous benchmarks Khalil & Sabry (2024).

For completeness, we employed two implementations of the chrF and BLEU measures using the SacreBLEU and HuggingFace libraries.

C Ablation studies

C.1 Performance of fine-tuned models

We fine-tuned two commonly chosen models (`opus-mt-ar-e` Saeed (2025) and 2M100-418MKhered et al. (2025)) using the training set of the ATHAR dataset. Notably, as shown in Table 3, the fine-tuned models required 10k examples to reach competitive performance, while a 15-shot ICL configuration for two light-weight models achieved similar outcomes.

Model	l	METEOR	
`opus-mt-ar-en` Saeedi et al. (2025)	0	0.50 (0.42-0.63)	
	10	0.50 (0.34-0.59)	
	100	0.50 (0.38-0.60)	
	10,000	0.58 (0.45-0.78)	Link to notebook
2M100-418M Khered et al. (2025)	10,000	0.66 (0.49-0.82)	Link to notebook
`marco-o1:7b-q4_K_M`	15	0.66 (0.54-0.79)	
`gemma3:1b-it-q4_K_M`	15	0.53 (0.38-0.65)	

Table 3: Performance of fine-tuned models: l denotes the number of examples shown in the prompt. Fine-tuning with 100 examples or fewer yielded poor METEOR score. Fine-tuning with all available samples in the ATHAR training set gave the METEOR score of 0.58.

C.2 Comparative analysis of sample-selection scheme

Selection scheme	METEOR	H-BLEU	Similarity	BERT-Score
Lexical	57.6 (43.8-71.9)	55.6 (44.3-68.1)	83.2 (73.7-91.9)	94.4 (92.4-96.2)
Length	54.0 (35.3-69.1)	53.3 (38.4-67.1)	82.1 (69.7-90.7)	94.0 (91.1-95.8)
Random	48.8 (22.8-65.1)	47.9 (28.3-60.2)	76.2 (59.0-87.0)	93.2 (89.3-95.2)

Table 4: Comparison of sample-selection schemes based on lexical diversity, sentence length, and random sampling. Selection based on maximal lexical diversity generally yielded top performance scores.

We explored three sample-selection scheme: choosing samples with the highest lexical diversity, those with the longest sentence lengths, or choosing at random. As reported in Table 4, selection based on maximal lexical diversity generally yielded top performance scores.

D Subjective evaluation

One of our Palestinian partners who holds a Doctor of Philosophy degree annotated the translations generated for a subset of the ArBanking dataset by one of the top-performing LLMs. Examples of the annotations are shown in Table 5.

#1	LLM: ~~Can~~ you tell me where my refunded money is? I requested the refund a few days ago but it's still not ~~showing up~~. Did you receive the money, or maybe **it's not in my account yet**? Let me know when it's all sorted.
	Corrected: **Could** you tell me where my refunded money is? I requested the refund a few days ago but it's still not **received**. Did you receive the money, or maybe **the problem is not just in my account**? Let me know when **all is well**.
#2	LLM: Can I ~~top up~~ my Apple Pay balance?
	Corrected: Can I **fill** my balance **using** my Apple Pay?
#3	LLM: I want to know what's the reason that makes you not allow my ~~beneficiary~~?
	Corrected: I want to know what's the reason that makes you not allow my **recipient**?
#4	LLM: A currency exchange ~~booth~~
	Corrected: A currency exchange **rate**.
#5	LLM: How much ~~can you~~ take to cover the payment?
	Corrected: How much **does it** take to cover the payment?
#6	LLM: I tried to transfer an ~~amount~~ to someone but it didn't ~~reach them~~ in full. Who can I contact to solve the problem?
	Corrected: I tried to transfer **money** to someone but it did not **reach him**. Who can I contact to solve the problem? **money**

Table 5: Examples of human corrections of LLM-generated translations.