Evaluation of Qwen3 for English to Ukrainian Translation

Cristian Grozea[1] * Oleg Verbitsky[2]
[1]Fraunhofer Institute FOKUS, Berlin, Germany
[2]Pidstryhach Institute for Applied Problems of Mechanics and Mathematics
Ukrainian Academy of Sciences, Lviv, Ukraine

Abstract

We report the results of evaluating Qwen3 for the English-to-Ukrainian language pair of the general MT task of WMT 2025.

In addition to the quantitative evaluation, we performed a qualitative evaluation, in collaboration with a native Ukrainian speaker - therefore we present an example-heavy analysis of the typical failures the LLMs still do when translating natural language, particularly into Ukrainian.

We report also on the practicalities of using LLMs, such as on the difficulties of making them follow instructions, on ways to exploit the increased "smartness" of the reasoning models while simultaneously avoiding the reasoning part improperly interfering with the chain in which the LLM is just one element.

1 Introduction

This submission to the general MT task of WMT 2025 from Fraunhofer FOKUS continues our series of baselines prepared for the WMT biomedical translation task, as this system was intended for preparing baselines in case the task had been organized this year. Those baselines evolved with the field: starting with training sequence-to-sequence models on biomedical corpora (Bawden et al., 2019), continuing with transformers trained by us on biomedical corpora (Bawden et al., 2020), starting using generic pretrained models like T5 – not necessarily targeting the biomedical field (Yeganova et al., 2021; Neves et al., 2022), then using one of the first widely deployed LLMs, ChatGPT 3.5 (Neves et al., 2023) and eventually using the emerging more capable locally-hosted open-model LLMs, LLama 3.1 70B (Neves et al., 2024).

Using locally hosted models keeps data away from cloud providers, leveraging the competitiveness of the locally-hosted large LLMs such as the Qwen3 (Yang et al., 2025) we used here. This is particularly important to ensure unfiltered access to the potential hiding in these models, as online services may further censor or otherwise limit the cloud-served models.

2 System Description

2.1 Hardware

The hardware this system ran on is part of the Fraunhofer FOKUS GPU cluster. It is a server with 8 Nvidia H100 SXM GPUs, each with 80 GB Video RAM (VRAM), 3 TB RAM, 192 CPU cores (384 threads). The power consumption of this server in idle mode is about 1800 W. Only 7 of the 8 GPUs were used to perform the translation task, at an average of 140 W/GPU, adding approximately 1000 W to the power consumption during the experiment.

2.2 Software

The GPU cluster is managed with Kubernetes[1]. To run the selected LLM, we have used the ollama system[2], ran as a Helm[3] deployment and limited to 7 of the 8 GPUs.

The translation itself has been controlled with a Python script that used the ollama Python API for calling the LLM, once per each paragraph to translate.

Translating the texts took approximately 12 hours.

2.3 LLM

The LLM we have employed is Qwen3, as available in the ollama library under the tag qwen3:235b-a22b-q4_K_M. It is a mixture of experts model, with 235.1e9 parameters, context length 262144, embedding length 4096. It has been used with temperature 0.6 (randomness level in decoding),

*cristian.grozea@fokus.fraunhofer.de

[1] https://kubernetes.io/

[2] https://ollama.com/

[3] https://helm.sh/

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 594–598
November 8–9, 2025. ©2025 Association for Computational Linguistics

top_k=20, top_p=0.95 and repeat_penalty=1. This is a quantized model, with Q4_K_M quantization, leading in average to the use of 4.84 bits for every parameter of the original not-quantized model.

2.4 Prompt

This is the prompt we used uniformly, regardless of the context (be it chat, or video captioning or another source):

“You are a helpful assistant specialized in translation. You will be provided with a text in English, and your task is to translate it into Ukrainian. Keep the formatting as close as possible to the source. Preserve meaning, tone, emotions, and nuances and target the cultural context of Ukrainian. For easier selection, mark the translated text with <translation-begin> and <translation-end>”.

3 Results and Analysis

3.1 Quantitative Evaluation

The automated evaluation performed by the organizers(Kocmi et al., 2025) computed several metrics characterizing the performance of our system: a trained quality estimation (CometKiwi-XL) score of 0.597 (versus 0.65 for the top system), an LLM evaluation score with the Commander-A model as judge of 75.7 (versus 84.1 for the top system), and of 78.1 with GPT4.1 as judge (versus 85.3 for the top system). A metric based on the reference translation (xCOMET-XL) was computed as well; score 0.513 (versus 0.662 for the top system). These metrics resulted in an AutoRank score of 8.7.

3.2 Qualitative Evaluation

We evaluated, grouped, and classified a randomly selected subset of errors on the very test set of the WMT 2025 general MT task, hoping that the recency of this dataset prevented the LLMs from being trained on these texts. We attempted to understand how the error came to be, given the statistical nature of the LLMs.

Types and sources of errors (**be warned, Example 3 could be shocking to sensitive readers, who might want to skip reading it**; the misspelled words in the source of translation are reproduced verbatim):

- The absence of idiomatic analogs in Ukrainian:

  Example 1

  EN: Rmember, theres alot of fish in the sea!

  UKR: Пам’ятай, у морі багато риби!

  This is a word-for-word translation. Ukrainians do not say it this way. However, in internet there are a lot of translations like this, which indicates this might become idiomatic in Ukrainian as well!

  Example 2

  EN: wejt out of their way

  UKR: вийшов зі свого шляху

  This is again an incorrect word-for-word translation. Correct would be: полишив свої справи.

- Unexplainable typos:

  Example 3

  EN: boil on my ass

  UKR: присом на сракі

  Correct: прищем на сраці

- (questionable) preservation of the English acronym:

  Example 4

  EN: WTF

  UKR: WTF (not translated)

- Play of words leading to the loss of sense:

  Example 5

  EN: Basically if you didn’t know what that is,it’s basically an example of two sentences. One says “Lets eat granny” making it seem like someone’s gonna eat their nan.However,the other example says “Lets eat,granny” implying a different meaning to the sentence. Pretty creative way to teach kids to use commas ig.

  UKR: По суті, якщо ти не знав(-ла), що це таке, це просто приклад двох речень. Одне каже: «Їдьмо, бабуся», через що здається, ніби хтось збирається з’їсти свою бабусю. А інший приклад: «Їдьмо, бабусю», — що змінює зміст речення. Досить творчий спосіб навчити дітей ставити коми, припускаю.

  Mistake: “Їдьмо” is not “Let’s eat” (though this incorrect guess is still somehow plausible/natural, due to the closeness of the stems of "to go" and "to eat" in Slavic languages). The actual meaning of “Їдьмо” is "let's go/drive", but even then "to drive someone" does not exist in Ukrainian. Thus, the sense is completely lost. Correct: "Давайте їсти бабусю" and "Давайте їсти, бабусю". This word-for-word translation would perfectly preserve the general meaning of the original.

Example 6

EN: One word I simply can't say properly is water..... It actually hurts my brain trying to add the T and I sound american if I do or still say wut-ta, IDK why it's so hard for me to say tbh. Just wondering what ppl thought are when you meet people who drop H's and have glottal stops.

UKR: Одне слово, яке я просто не можу правильно вимовити – це вода... Навіть бiльно у головi, коли намагаюся додати T, i виглядає, нiби я американець, якщо так роблю, або все ще кажу "вутта". Не знаю, чому менi так важко це вимовити, щиро кажучи. Цiкаво, що люди думають, коли зустрiчають тих, у кого випадає "г" i є гортаннi зупинки.

The sense is lost as the word "вода" (translated from "water") does not contain the letter T at all.

- Translating with a blend of Ukrainian and Russian, including made-up words:

  Example 7

  EN: innit

  UKR: а?

  Correct: га? Or even better: чи не так? правда ж?

  Example 8

  EN: BLOW

  UKR: ЗIРВЕТЬСЯ

  Correct: ВИБУХНЕ.

  ЗIРВЕТЬСЯ here seems a blend of Ukrainian and Russian stemming from the Russian word ВЗОРВЁТСЯ. In fact, ЗIРВЕТЬСЯ exists in Ukrainian but has a completely different meaning (to fall off).

- corrupted words (in the translation)

  Example 9

  EN: RN

  UKR: ПРЯМОСЬОГО ТЕПЕР

  Correct: ПРЯМО ТЕПЕР.

  The word ПРЯМОСЬОГО doesn't exist in Ukrainian, it sounds slightly like ПРЯМIСIНЬКО (infrequent, more stringent variant of ПРЯМО).

  Yet another case appears in the translation from Example 6 above, "бiльно" does not exist in Ukrainian, although it is formed in a plausible way: бiль+но. Correct would be боляче (painful).

- missing slang equivalents

  EN: BRUH

  UKR: БРУХ

  This is a transliteration, such a word does not exist in Ukrainian.

4 Discussion

When selecting Qwen3 235B, we evaluated it briefly and informally against another truly large LLM that can be run locally on powerful machines, DeepSeek-R1 671B (DeepSeek-AI, 2025), also quantized, with the same type of quantization. Somewhat surprisingly, the Qwen3 model, which is nearly three times smaller, seemed to outperform the largest DeepSeek-R1 model on English to Romanian and English to German tasks. In the introductory blog entry[4] Qwen3 has been presented as supporting 119 languages, including Ukrainian and English, whereas the training of DeepSeek-R1 focuses on English and Chinese (DeepSeek-AI, 2025). All this made us employ Qwen3 instead of DeepSeek-R1, as the more efficient Qwen3 was also faster. In retrospect, this seems to have been a poor decision, as the organizers report better results with DeepSeek-R1 for the task of English-to-Ukrainian translation. We should have contrasted the performance of those two models on the intended language pair.

Fine-tuning a general LLM to a task like this language pair translation was an option with the smaller models, but it became much more difficult as the models grew towards the limits of the available hardware resources, especially VRAM. The hope is not to have to specialize the models, but to get good results from generalist models instead. As the system we employed was intended as a baseline,

[4] https://qwenlm.github.io/blog/qwen3/

we refrained from attempting to improve upon the standard pre-trained Qwen3 model, as published for everyone.

Despite the instructions shown in Section 2.4, the LLM we employed did not always preserve the newline structure of the source, where double newlines served as paragraph delimiter. Therefore, and to make the task easier for the LLM, although with the risk of providing too little context for the translation, we split explicitly in the Python script the text into paragraphs, translated each paragraph and recomposed the resulting text by joining the translations with the expected delimiter. Even with this procedure, 50 out of 1251 texts had to be retranslated, because the LLM introduced spurious double newlines inside the translation of single paragraphs, disrupting the correspondence of the input and the output paragraphs.

Qwen3 is a so-called "reasoning" model, meaning that before outputting the desired translation it produces "reasoning" text describing its approach and its doubts about the task. This part is delimited by the tags <think> and </think>. We deleted this part of the output by removing everything up to and including the closing </think> tag.

The LLM inconsistently used the required tags to mark the translation and separate it from various other comments it produced in addition to it. It produced randomly such alternative closing tags for the translation section: </end-translation> and </begin-translation>. Our Python script was designed to detect and accept also these alternative closing tags.

In the end, two of the outputs still contained traces indicating that something went wrong with the automatic extraction of the translated text, that is they still contained the <think> tag. We decided not to fix those, in order to get a realistic evaluation of what to expect when using a "reasoning" LLM for translation.

We expect the need for explicit well-controlled postprocessing – as described here – to remain, as none of the models we interacted with have complete adherence to the instructions.

Concerning the errors those models still produce in machine translation, it might be that successfully finetuning on a well-curated parallel corpora might eliminate some of them, but probably not all.

5 Conclusion

We have evaluated the use of one of the largest local LLMs for automatic translation of English to Ukrainian. Beyond the automated quantitative analysis performed by the task organizers, we have performed a qualitative analysis of several translation errors observed, and explained also the engineering issues one has to deal with when relying on LLMs for machine translation, such as the incomplete adherence to instructions and the subsequent need for postprocessing, especially when using a "reasoning" LLM. We conclude that, although neural machine translation of occasionally challenging texts in natural language has advanced significantly, the LLMs, as the other neural models before them, continue to be characterized by two aspects: smooth and polished output, convincing thanks to the good form, paired with instances where meaning is sometimes missed – or even reversed. Still, when a human with knowledge of the target and of the source languages is proofreading the outcome, the translation process can be significantly accelerated using such a local LLM.

6 Thanks

We thank the anonymous reviewers, whose suggestions led to improving the quality of this article.

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, et al. 2025. Preliminary ranking of wmt25 general machine translation systems. *arXiv preprint arXiv:2508.14909*.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-lopez, Eulalia Farre-maduel, Martin Krallinger, Cristian Grozea, and Aurelie Neveol. 2022. Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 694–723, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. Association for Computational Linguistics.