American Journal of Interdisciplinary Research and Development
ISSN Online: 2771-8948
Website: www.ajird.journalspark.org
Volume 53, June - 2026

A THEORETICAL FRAMEWORK FOR COMPARING NEURAL MACHINE TRANSLATION AND LARGE LANGUAGE MODELS IN THE UZBEK-RUSSIAN LANGUAGE PAIR

Maqsuda K. Haitova
English Language Teacher, School No. 43, Bukhara, Uzbekistan

Abstract
Institutions in Uzbekistan now choose between dedicated neural machine translation engines and general-purpose large language models for Uzbek-to-Russian work, and they choose without evidence: no published study measures the two families against each other on this pair. This article does not supply the measurement. It supplies the theoretical framework the measurement would require, and argues that the framework cannot be borrowed unchanged from the existing literature. Three arguments are developed. The comparative evidence accumulated in shared-task evaluation establishes a high-resource advantage and a low-resource deficit for general-purpose models, but the Uzbek-to-Russian configuration, a low-resource agglutinative source with a high-resource fusional target, is not among the configurations tested, and the direction of the combined effect is underdetermined. The two families of system are optimised over different objects and therefore fail in different ways, so a comparison expressed only as a quality score cannot distinguish them in the respect that matters to a post-editor. The typological distance between Uzbek and Russian, in suffix chaining, in the absence of grammatical gender and in constituent order, makes the pair a distinct theoretical case rather than another instance of a known one. The article closes with what evaluation theory permits to be claimed and with the design constraints that follow.

Keywords: Machine translation theory; large language models; Uzbek; Russian; low-resource language pairs; agglutinative morphology; translation quality evaluation.

Introduction

Uzbek-to-Russian translation is produced in quantity in Uzbekistan every working day, in ministries and local administrations, in universities, in news agencies and in publishing. Where the work has been automated, it was until recently handled by dedicated neural machine translation engines. Since 2023 general-purpose large language models have been offered for the same task, often at lower cost and with the practical advantage that a single interface covers translation, summarisation and editing. Institutions are consequently

choosing between two families of system, and no published study measures the two families against each other on Uzbek-to-Russian text.

The natural response is to consult the comparative literature and transfer its conclusion. This article argues that the transfer is not licensed, and that the obstacle is theoretical rather than merely a shortage of data. The comparative literature was built on language configurations that differ from this one in the respects that determine the answer; the two families of system are optimised over different objects and therefore produce different kinds of error at the same measured quality; and the typological relation between Uzbek and Russian introduces a class of decisions that the source does not determine at all. Each of these is a reason why a measurement for another pair does not predict this one, and each imposes a constraint on how the measurement for this pair would have to be built.

The argument proceeds in four steps. Section 2 sets out what the accumulated evidence establishes and what it leaves open. Section 3 distinguishes the two families by their training objective and derives the error profiles that follow. Section 4 grounds the Uzbek-to-Russian case in the documented typology of the two languages, with glossed illustrations of the three decision points where the source underdetermines the target. Section 5 states what evaluation theory permits to be claimed and the constraints that follow for any study of this pair. No measurement is reported and none is estimated.

2. What the comparative evidence establishes and what it leaves open

The most authoritative recurring source is the WMT General Machine Translation shared task. Its 2023 edition reported that submissions built on large language models were competitive with dedicated systems for high-resource directions without uniformly surpassing them (Kocmi et al., 2023). Its 2024 edition, which collected translations from eight general-purpose models and four online providers across eleven language pairs, each tested on three to five domains, framed the finding as the arrival of the large-model era alongside the persistence of unsolved translation (Kocmi et al., 2024). Hendy et al. (2023), evaluating three GPT models across eighteen directions, reported very competitive quality for high-resource languages and limited capability for low-resource ones. This split is the most stable generalisation the field has produced.

A second result narrows the question further. Manakhimova et al. (2023), in the first fine-grained linguistic analysis of GPT-4 translation output, found performance comparable to the strongest systems for the German directions and a fall into a lower significance cluster for English–Russian. Because that analysis isolates the target language rather than the direction as a whole, it is the closest existing evidence to a Russian-target case, and it points against the general-purpose family.

A third result places Uzbek. Every large multilingual effort treats it as low-resource. The 200-language NLLB project reported an average gain of almost 7.3 spBLEU over the nearest prior system on the FLORES-200 benchmark, an improvement of 44 per cent, with human assessment on the XSTS scale (NLLB Team et al., 2024); the gain is an average,

and the low-resource tail remains the harder part of the distribution. The Turkic Interlingua corpus assembled roughly two million sentence pairs across 22 Turkic languages with baselines for 26 pairs and test sets in three domains (Mirzakhalov et al., 2021), but the effort is English- and Turkish-centred and reports no Uzbek-to-Russian system quality.

Three findings therefore converge on the pair without determining it. The low-resource deficit is established for the pairs in which it was observed; the Russian-target weakness is established for an English source; the low-resource status of Uzbek is established by resource counts. What none of them settles is how a low-resource source combines with a high-resource target. Two theoretical expectations pull in opposite directions. A strong Russian prior may repair a weak Uzbek reading, producing well-formed output from an imperfect analysis; or the same prior may generate fluent Russian that has drifted from a source the model read poorly, in which case fluency conceals the failure rather than compensating for it. The published evidence contains no case that discriminates between them, because the configuration has not been tested. The gap is thus not a missing number in an otherwise complete picture but an untested combination of factors whose interaction is the object of interest.

There is a further reason why the resource axis alone is the wrong instrument for placing this pair. Resource counts are reported per language, whereas translation quality is a property of a direction, and a direction has two ends that may sit at opposite points on that axis. The literature has largely tested directions whose ends are close together, either both high-resource or both comparatively thin, and where they are far apart the high-resource end has usually been the source. Uzbek-to-Russian inverts that arrangement. The quantity that matters is not the resource level of either language but the relation between the model's evidence about the source and its evidence about the target, and no published figure reports that relation for this pair.

3. Two architectures and two kinds of failure

Systems that are optimised differently fail differently, and the distinction is theoretically prior to any measurement of their quality. A dedicated engine is trained on parallel text for a direction and optimises a conditional distribution over target sentences given source sentences. Its characteristic failures are those of a system with too little of that parallel text: the wrong lexical choice in a thin domain, breakdown on rare morphology, and silent omission of material it cannot align. A general-purpose model optimises next-token prediction over an unbalanced multilingual corpus and is steered toward translation only afterwards, by instruction. Its characteristic failures are those of a fluent generator: plausible target text with no warrant in the source, drift of register, and elaboration where the source was terse.

Two consequences follow. The first is that a single quality score cannot separate the families in the respect that matters in practice. Two systems may reach the same score by

making opposite mistakes, and an institution deciding what it will have to post-edit needs to know which mistakes, not which score. Omission and addition are not symmetrical failures for a post-editor: a dropped clause is visible against the source, whereas an added one is invisible in the target and is caught only by comparison. A comparison that reports quality without reporting an error profile is therefore theoretically incomplete, whatever its statistical rigour.

The asymmetry runs deeper than a difference in error counts, because the two families dissociate adequacy from fluency in opposite directions. A dedicated engine with thin parallel data tends to produce output whose defects are visible on the surface of the target text: an ungrammatical form, a missing constituent, a term left untranslated. Fluency degrades with adequacy, and the reader is warned. A general-purpose model trained principally on monolingual text has a strong prior over well-formed target sentences and comparatively weaker evidence about the source, so its defects tend to be invisible on the surface: the sentence is well formed, idiomatic and wrong. Fluency is maintained while adequacy degrades, and the reader is not warned. This is the theoretical reason the parity literature insists on expert raters with access to the source (Freitag et al., 2021): a rater judging target text alone will systematically prefer the family whose errors are harder to detect, and will do so more strongly the more fluent that family becomes. A comparison of the two families is therefore also a test of the evaluation protocol, and a protocol that cannot detect this dissociation will report the wrong winner without any statistical fault.

The second consequence concerns the instruction itself. Where one of the compared systems is steered by a prompt, the prompt is an experimental factor and not a presentational detail. Zhang et al. (2023) showed that prompt format and exemplar quality move translation quality substantially and that variance across repeated samplings can be large, reporting the effect as a distribution rather than as a point. Vilar et al. (2023) found the quality of the exemplar pool to dominate the strategy by which exemplars are selected. Bawden and Yvon (2023) found format to matter particularly for low-resource pairs. Raunak et al. (2023) located the mechanism more precisely, finding that perturbing the target side of demonstrations degrades quality far more than perturbing the source side, which indicates that the output-text distribution carries the main in-context learning signal. If the spread across reasonable formulations is of the same order as the distance between systems, then a comparison that fixes one formulation has measured the formulation. The theoretical status of any such comparison is conditional, and the condition must be stated.

4. Why Uzbek–Russian is a distinct theoretical case

The typological facts are documented rather than assumed. The morpheme-alignment corpora of Zong et al. (2016) state that Uzbek is agglutinative, lacks grammatical gender, and adds suffixes in a fixed order encoding a high number of inflectional categories. The Uzbek Universal Dependencies treebank corroborates the morphological complexity and

the parsing burden it imposes, at inter-rater agreement above 0.90 (Matlatipov & Aripov, 2026). Russian is fusional, marks gender obligatorily on nouns, on adjectival agreement and on past-tense verbs, and orders constituents differently. Three mismatches follow, and each is a point at which the source underdetermines the target. The examples below are constructed to illustrate the mismatches; none is the output of any system.

The first mismatch concerns suffix chains. A single Uzbek word may carry possessive, case, derivational and number marking in sequence, where Russian bundles several features into one inflection or distributes them across several words.

(1) Uzbek: maktabimizdagilarga

Gloss: maktab-imiz-da-gi-lar-ga — school-1PL.POSS-LOC-ADJ-PL-DAT

Russian: тем, кто находится в нашей школе (tem, kto nakhoditsya v nashey shkole)

Back-translation: 'to those who are at our school'

One source token becomes a five-word Russian relative clause. No single Russian inflection carries the chain, so the system must restructure rather than translate item by item. The restructuring can fail morphologically, when the case or number of the Russian head is wrong, or syntactically, when the relative clause is absent and the content is compressed into an ill-formed noun phrase. A theory of the comparison must keep the two apart, because there is no reason to expect the two families to fail in the same proportion across them.

The second mismatch concerns gender, and it is the clearest instance of information that the target obligatorily encodes and the source does not mark at all.

(2) Uzbek: U shifokor. U kasalxonaga bordi.

Gloss: 3SG doctor. 3SG hospital-DAT go-PST-3SG

Russian, masculine reading: Он врач. Он пошёл в больницу (On vrach. On poshyol v bolnitsu)

Russian, feminine reading: Она врач. Она пошла в больницу (Ona vrach. Ona poshla v bolnitsu)

Back-translation: 'S/he is a doctor. S/he went to the hospital.'

Uzbek u is genderless, a property stated explicitly in the source description (Zong et al., 2016). Russian requires gender on the pronoun and on the past-tense verb, so the system must choose, and in an isolated sentence there is nothing in the input on which to base the choice. What a system does here is diagnostic of its architecture. Consistent defaulting produces a predictable error rate and a self-consistent text; a choice driven by the target-language prior may attach gender to the occupational noun and produce readings that conflict within a single passage. This is also the point at which document context stops being a refinement of evaluation design and becomes a determinant of the result, which is the theoretical reason the parity literature insists on it (Läubli et al., 2018; Läubli et al., 2020).

The third mismatch concerns constituent order and the unwinding of pre-nominal participial material, which is where administrative text is hardest.

(3) Uzbek: Vazirlik tomonidan tasdiqlangan nizomga muvofiq ariza uch ish kuni ichida ko‘rib chiqiladi.
Gloss: ministry by approve-PTCP regulation-DAT according.to application three work day within consider-PASS-PRS
Russian: В соответствии с положением, утверждённым Министерством, заявление рассматривается в течение трёх рабочих дней (V sootvetstvii s polozheniyem, utverzhdyonnym Ministerstvom, zayavleniye rassmatrivayetsya v techeniye tryokh rabochikh dney)
Back-translation: ‘In accordance with the regulation approved by the Ministry, the application is considered within three working days.’
The Uzbek participial phrase precedes its head and the finite verb closes the clause; Russian places the participial phrase after its head, sets it off with commas, and puts the verb medially. Administrative prose compounds the obligatory reordering with formulae whose Russian equivalents are conventional rather than compositional, so a system that translates the formula compositionally produces output that is comprehensible and wrong in register. That failure is terminological rather than syntactic, and a taxonomy that does not separate the two will record it in the wrong place. Domain is therefore not a sampling convenience in this pair but a theoretical variable, which is why pooled reporting is inadequate and per-domain reporting is the shared-task convention (Kocmi et al., 2024).
The three mismatches share a property that separates them from ordinary translation difficulty and that gives the pair its theoretical interest. In each of them the source does not carry the information the target obligatorily requires. This is underdetermination rather than difficulty, and it cannot be removed by more parallel data for the direction, because the missing information is absent from the source rather than rare in the training corpus. It can be resolved only by inference from context, by inference from world knowledge, or by a default. The comparison is therefore not simply a question of which family translates more accurately but of which strategy each family adopts where accuracy is unavailable, and of which strategy suits the text at hand. A default is auditable and correctable at scale: if a system resolves unmarked gender to the masculine, the behaviour is known, its error rate is estimable, and an institution can instruct post-editors accordingly. An inference is neither auditable nor uniform, and may still be right more often. Which is preferable depends on the genre and on whether the output is read as a finished text or as a draft, and this is a question about the use of translation rather than about its quality, which no single score can answer.
One further variable is specific to Uzbek and has no counterpart in the pairs on which the comparative literature rests. Uzbek is written in two active systems, a fact documented in the construction of the first monolingual Uzbek language model (Mansurov & Mansurov, 2021). Script is a property of the input rather than of the language, so mixed input would confound the comparison of systems with a difference in preprocessing. Any claim about this pair presupposes that script has been fixed and reported.

5. What evaluation theory permits to be claimed

Conclusions of this kind are fragile to method, and the fragility has been demonstrated rather than merely asserted. When human parity was claimed for Chinese–English news translation in 2018, two independent reassessments dissolved it: Läubli et al. (2018) showed that the claim did not survive document-level rating, and Toral et al. (2018) identified translationese source text and insufficient rater proficiency as confounds. The recommendations that followed established that perceived quality depends jointly on who rates, how much context they see and how the reference was produced (Läubli et al., 2020). Freitag et al. (2021), re-scoring top systems with professional annotators under Multidimensional Quality Metrics, obtained a substantially different ranking from the one crowd workers had produced, with a clear preference for human over machine output. A ranking is thus a joint product of the systems and the protocol, and reporting one without the other is uninterpretable.

Metric choice carries the same conditionality. BLEU (Papineni et al., 2002) is reportable only under the conditions set by Post (2018), since otherwise scores are not comparable across studies; character-level scoring is preferable where the target is morphologically rich, which is precisely the Russian case (Popović, 2017); neural metrics have displaced both as the primary signal (Rei et al., 2020). Marie et al. (2021), surveying 769 papers, documented the cost of reliance on BLEU alone, of omitted significance testing and of unreported metric versions. The WMT23 metrics findings add that references are not neutral instruments (Freitag et al., 2023), which bears directly on a pair for which no established reference set exists and every reference must be newly commissioned.

Four constraints follow for any study of Uzbek-to-Russian translation quality. Quality must be reported with an error profile, because the families differ in kind rather than merely in degree, and the taxonomy should map onto an established typology (Lommel et al., 2014) so that the counts are readable against the wider literature. Results must be reported by domain, because the mismatches of Section 4 are distributed unevenly across genres. Prompt formulation must be reported as a condition with its spread, because otherwise the model and the instruction are confounded. Metric versions, system versions and access dates must be recorded, because outputs change without notice. Beyond these, the scope of any claim is bounded by its instrument: a finite test set, a fixed set of systems and a fixed set of formulations support a claim about those systems on that data at that date, and not a claim about large language models as a class or about Uzbek-to-Russian translation in general.

A final constraint is epistemic rather than methodological, and it governs how the material assembled here may be used. The literature reviewed supplies findings; this article supplies hypotheses derived from them; the two must not be merged in the writing. Kocmi et al. (2023, 2024) report system rankings for the pairs they tested and offer no prediction about Uzbek. Manakhimova et al. (2023) speak to English–Russian, not to Uzbek–Russian, and their result becomes a claim about the present pair only through an inference that is this

author's and not theirs. The NLLB and Turkic Interlingua reports describe the quality of dedicated systems and make no comparison with general-purpose models. Crediting any of these authors with a conclusion about Uzbek-to-Russian translation would misrepresent them and would also disguise the point of the exercise, which is that the conclusion does not yet exist. The distinction between what is established, what is hypothesised and what is interpreted is therefore maintained explicitly throughout the argument above, and it is the first thing a study of this pair should preserve when it reports its results.

6. Conclusion

The question of whether dedicated engines or general-purpose models translate Uzbek into Russian better is open, and this article has argued that it is open for reasons internal to the theory rather than for want of an experiment. The comparative literature establishes a high-resource advantage and a low-resource deficit, but it has not tested a low-resource agglutinative source against a high-resource fusional target, and the two mechanisms by which a strong target prior could operate, repair and drift, predict opposite outcomes. The families are optimised over different objects and fail in different ways, so a comparison expressed as a single score is incomplete in the respect that matters to the institutions that would act on it. The typological distance between the two languages creates decision points, in suffix chaining, in obligatory gender and in constituent order, at which the source does not determine the target and the system's behaviour is diagnostic of its architecture rather than of its quality.

Until measurement exists, the recommendation that the present evidence supports is narrow. Published comparisons should not be treated as transferable to this pair. Some post-editing should be assumed whichever family is selected, since translation is not solved for pairs of this kind (Kocmi et al., 2024). Where a general-purpose model is used, the prompt should be fixed, documented and treated as part of the configuration rather than left to the individual user. Where evaluation is carried out locally, it should not rest on BLEU alone (Marie et al., 2021). The framework set out here defines what a study of this pair would have to hold constant, what it would have to report, and how far its conclusions could reach; the same framework transfers without modification to Kazakh–Russian, Kyrgyz–Russian and Tajik–Russian, for which the evidence base is equally thin.

References

1. Bawden, R., & Yvon, F. (2023). Investigating the translation performance of a large multilingual language model. The case of BLOOM. Proceedings of the 24th Annual Conference of the European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.16
2. COPE Council. (2024). Handling text recycling (also known as self-plagiarism) (Version 1). Committee on Publication Ethics. https://doi.org/10.24318/6eWbU5xd

3. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (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. https://doi.org/10.1162/tacl_a_00437
4. Freitag, M., Mathur, N., Lo, C., Avramidis, E., Rei, R., Thompson, B., Kocmi, T., Blain, F., Deutsch, D., Stewart, C., Zerva, C., Castilho, S., Lavie, A., & Foster, G. (2023). Results of WMT23 metrics shared task. Metrics might be guilty but references are not innocent. Proceedings of the Eighth Conference on Machine Translation. Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.51
5. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation [Preprint]. arXiv. https://arxiv.org/abs/2302.09210
6. Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., … Zouhar, V. (2024). Findings of the WMT24 general machine translation shared task. The LLM era is here but MT is not solved yet. Proceedings of the Ninth Conference on Machine Translation (pp. 1–46). Association for Computational Linguistics. https://aclanthology.org/2024.wmt-1.1
7. Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Koehn, P., Marie, B., Monz, C., Morishita, M., Murray, K., Nagata, M., Nakazawa, T., Popel, M., … Zouhar, V. (2023). Findings of the 2023 Conference on Machine Translation (WMT23). LLMs are here but not quite there yet. Proceedings of the Eighth Conference on Machine Translation (pp. 1–42). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.1
8. Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., & Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. Journal of Artificial Intelligence Research, 67, 653–672. https://doi.org/10.1613/jair.1.11371
9. Läubli, S., Sennrich, R., & Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4791–4796). Association for Computational Linguistics. https://aclanthology.org/D18-1512
10. Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM). A framework for declaring and describing translation quality metrics. Tradumàtica, 12, 455–463. https://doi.org/10.5565/rev/tradumatica.77
11. Manakhimova, S., Avramidis, E., Macketanz, V., Lapshinova-Koltunski, E., Bagdasarov, S., & Möller, S. (2023). Linguistically motivated evaluation of the 2023

state-of-the-art machine translation. Can ChatGPT outperform NMT? Proceedings of the Eighth Conference on Machine Translation (pp. 224–245). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.23

12. Mansurov, B., & Mansurov, A. (2021). UzBERT. Pretraining a BERT model for Uzbek [Preprint]. arXiv. https://arxiv.org/abs/2108.09814
13. Marie, B., Fujita, A., & Rubino, R. (2021). Scientific credibility of machine translation research. A meta-evaluation of 769 papers. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (pp. 7297–7306). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.566
14. Matlatipov, S., & Aripov, M. (2026). UzUDT. Uzbek Universal Dependencies treebank. Proceedings of the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (pp. 11642–11649). ELRA and International Committee on Computational Linguistics. https://doi.org/10.63317/2uedjqezjxn5
15. Mirzakhalov, J., Babu, A., Ataman, D., Kariev, S., Tyers, F., Abduraufov, O., Hajili, M., Ivanova, S., Khaytbaev, A., Laverghetta, A., Jr., Moydinboyev, B., Onal, E., Pulatova, S., Wahab, A., Firat, O., & Chellappan, S. (2021). A large-scale study of machine translation in the Turkic languages. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 5876–5890). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.475
16. NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejia Gonzalez, G., Hansanti, P., … Wang, J. (2024). Scaling neural machine translation to 200 languages. Nature, 630, 841–846. https://doi.org/10.1038/s41586-024-07335-x
17. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU. A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135
18. Popović, M. (2017). chrF++. Words helping character n-grams. Proceedings of the Second Conference on Machine Translation (pp. 612–618). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4770
19. Post, M. (2018). A call for clarity in reporting BLEU scores. Proceedings of the Third Conference on Machine Translation. Research Papers (pp. 186–191). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319
20. Raunak, V., Menezes, A., Post, M., & Awadalla, H. H. (2023). Dissecting in-context learning of translations in GPTs. Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.564

21. Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET. A neural framework for MT evaluation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213
22. Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. Proceedings of the Third Conference on Machine Translation. Research Papers (pp. 113–123). Association for Computational Linguistics. https://aclanthology.org/W18-6312
23. Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., & Foster, G. (2023). Prompting PaLM for translation. Assessing strategies and performance. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 15406–15427). Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.859
24. Zhang, B., Haddow, B., & Birch, A. (2023). Prompting large language model for machine translation. A case study. Proceedings of the 40th International Conference on Machine Learning (PMLR 202, pp. 41092–41110). https://proceedings.mlr.press/v202/
25. Zong, H., Xu, J., Grimes, S., Song, Z., Strassel, S., & Maxwell, M. (2016). Uzbek–English and Turkish–English morpheme alignment corpora. Proceedings of the Tenth International Conference on Language Resources and Evaluation (pp. 2925–2930). ELRA. https://aclanthology.org/L16-1467.