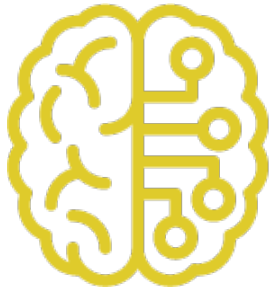

Is AI the new “Human Evaluator”?
Aneta Sapeta, MT and AI Specialist
vistatec

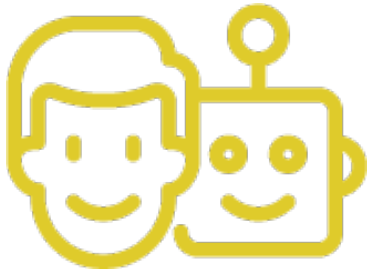
Key role of the MT evaluations

vistatec



Assessing the quality of MT engines for implementation in translation workflows

To show the client tangible data on how good/bad the MT output is



To see if baseline engines QUALITY IS sufficient or is MT training required

Use the data for estimating MTPE discounts

Think Global

MT Evaluation Steps for measuring MT quality

Current steps for evaluating the MT quality:

Step 1

METEOR

TER

BLEU

Automated Evaluation Metrics

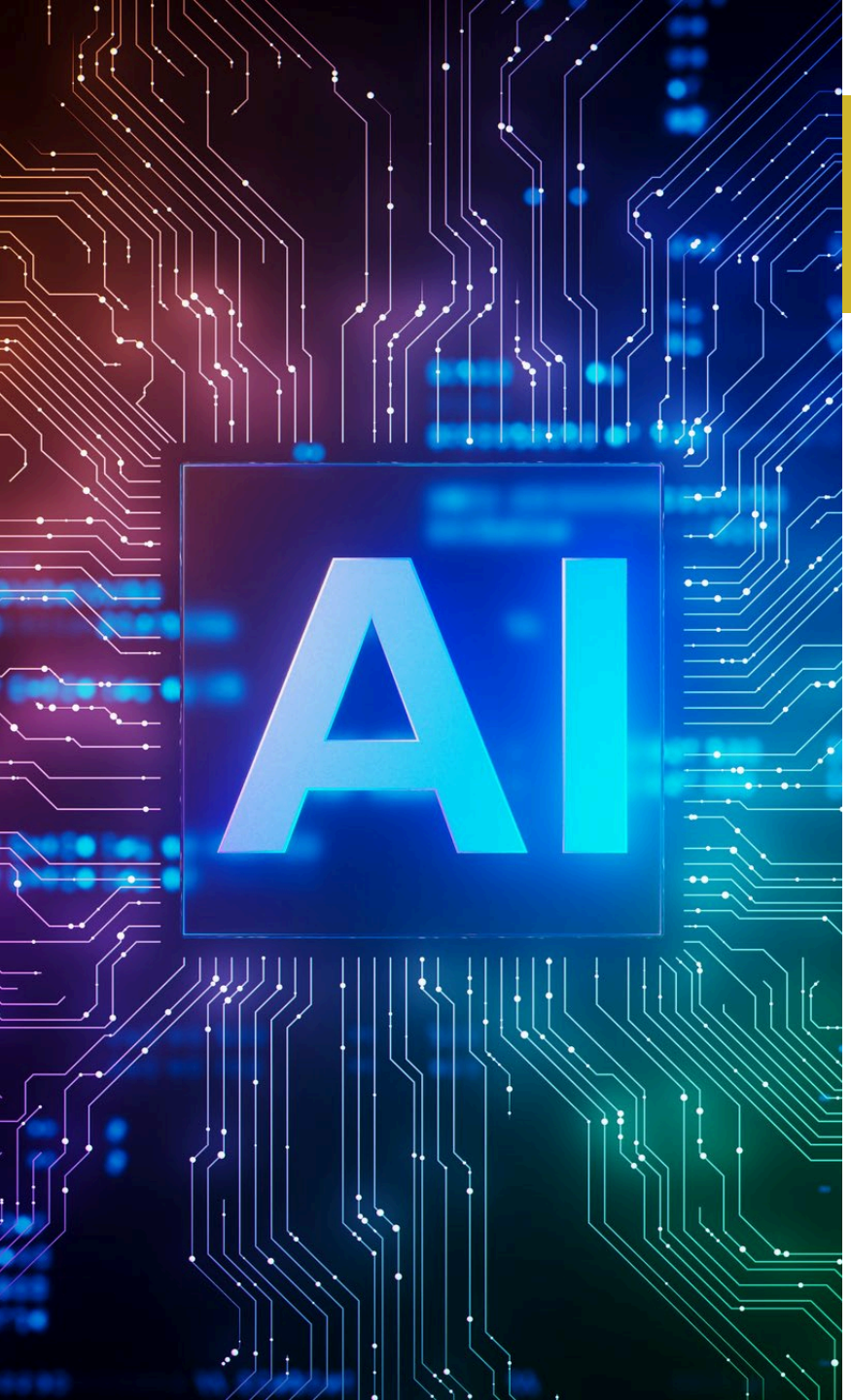
Indicator for MT quality

Step 2

Determinator for MT quality

MT Evaluation Steps for measuring MT quality

- Human Evaluators are evaluating the MT quality by:
  - Labeling the error type(s) using the MQM error typology for Accuracy, Fluency, Grammar etc.
  - Providing scoring from 1 - 4 on the evaluated segments.
  - Providing comments and feedback.
  - Quality indicator: more than 70% of segments to have score of 3 and 4 (segments with few minor issues and usable or no error segments).

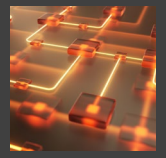


Using AI for automatic MT Evaluation

Automating MT evaluation with AI we wanted to:

- See if we can rely on AI evaluation to reduce the cost for human evaluation
  (usually, minimum two independent linguists are needed for human evaluation.)
- Shorten the time to get an indication of MT output quality by providing correct
  and reliable evaluation to the clients.
- By that to shorten the time for integration of MT solutions in the production
  (usual time for performing evaluation on a sample is minimum 1h.)

Using AI for automatic MT Evaluation



GEMBA framework for MT quality evaluation.

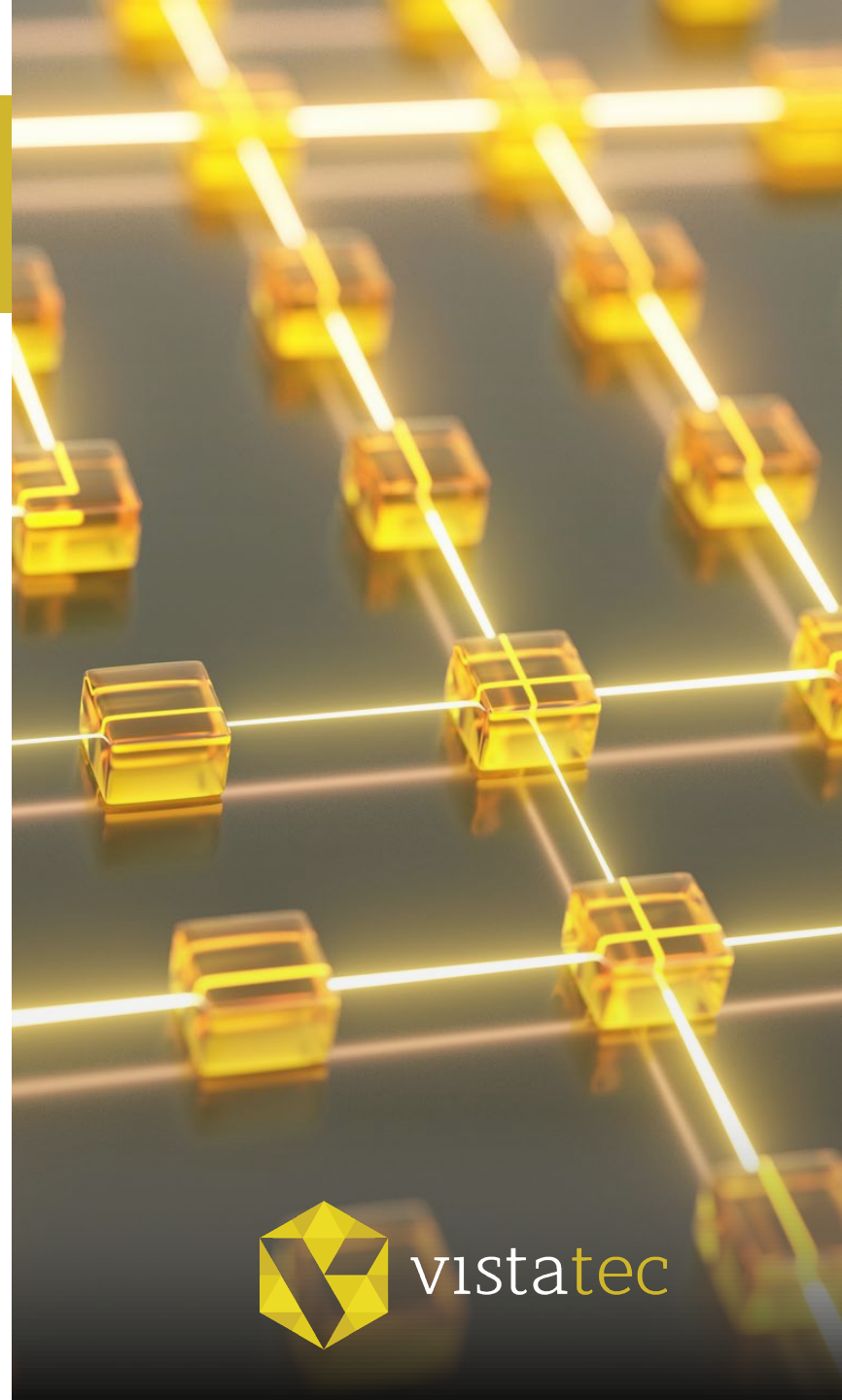
We checked the provided prompt models and the one that fits our purpose was the rating prompt. We modified the prompts by adding scores instead of stars and added error types for evaluation and the severity.

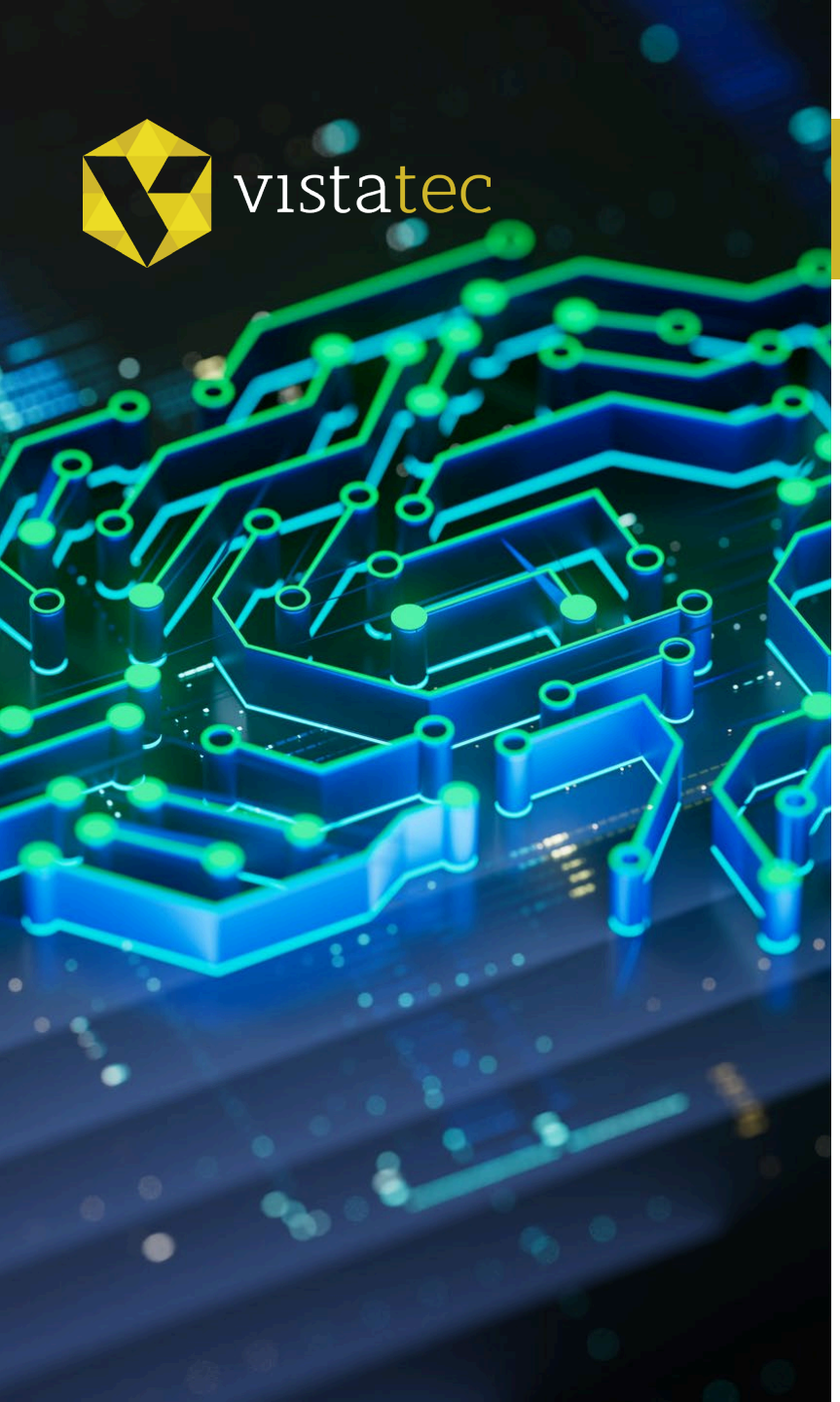
We used GPT3.5-Turbo and GPT4 for evaluating translations with provided reference (Human Translation) and without.

Think Global

Using AI for automatic MT Evaluation

- We conducted tests across multiple languages, including Spanish, German, Russian, Japanese, Chinese, Polish, Brazilian Portuguese, Indonesian, Italian, and French. From these, we selected French and Polish as illustrative examples.
- In comparing the scores, we aimed to determine the extent of overlap with the scores from human evaluations. On average, we found that the overlap in scoring varied by language, typically ranging between 30-60%. E.g:
  - PL – 55.45% of score overlap with the human scores
  - FR – 48.19% of score overlap with the human scores





Findings

- AI did not follow the instructions 100% of the time.
- Unreliable scoring – scoring the same issue with different score.
- Missing obvious errors that were visible especially the punctutation issues for example in French language, decilination missing in Polish MT output etc.
- Missing to identify untranslated segments and scoring them with highest score.
- Seemed to focus on one error in segment only and missed secondary errors.
- Scoring too enthusiastic.

Examples

Score definition:

1. Meaning many major issues, not comprehensible nor fluent not usable at all.
2. Meaning partially comprehensible and fluent with few major and minor issues, partly usable.
3. Meaning one or two minor errors, mostly comprehensible and usable.
4. Meaning no errors, perfect translation.

Think Global

Examples

Source	MT output	Reference	AI score	AI comment	AI scoring correct?
On the contrary, the elegance in the execution of the tango's leg hooks contrasted with the slum scenery in the theatre	Wręcz przeciwnie, elegancja w wykonaniu haczyków na nogi w tangu kontrastowała ze slumsową scenerią w teatrze	Wręcz przeciwnie, elegancja wykonania haczyka w tangu kontrastowała ze slumsową scenerią w teatrze	3 – 1-2 minor errors, mostly comprehensible and usable	The translation is mostly comprehensible and fluent with only minor issues. The addition of "na nogi" (on the legs) is not necessary as "haczyki" in the context of tango already implies leg hooks. This could be considered a minor addition error. However, it does not significantly alter the meaning and the sentence remains clear and understandable. There are no major errors, and the translation is mostly usable	NO
Blockchain technologies are used to manage public digital ledgers	Technologie Blockchain służą do zarządzania publicznymi księgami cyfrowymi	Technologie blockchain są używane do zarządzania publicznymi cyfrowymi księgami rachunkowymi	3 – 1-2 minor errors, mostly comprehensible and usable	Accuracy - Mistranslation . The word "rachunkowymi" (accounting) is omitted, which slightly changes the meaning. "Księgi cyfrowe" is less specific than "księgi rachunkowe"	YES
These digital ledgers record transactions across many globally distributed computers	Te cyfrowe księgi rejestrują transakcje na wielu komputerach rozproszonych na całym świecie	W tych cyfrowych księgach rachunkowych transakcje są zapisywane na wielu globalnie rozproszonych komputerach	4 – no errors, perfect translation	No errors found	NO

Source	MT output	Reference	AI score	AI comment	AI scoring correct?
According to the text, it is not always necessary to have dinner after the "apericena"	Zgodnie z tekstem, nie zawsze konieczne jest zjedzenie kolacji po "apericenie"	Według tekstu po „apericena” nie zawsze trzeba jeść kolacji	4 – no errors, perfect translation	The Polish translation "Zgodnie z tekstem, nie zawsze konieczne jest zjedzenie kolacji po 'apericenie'." is fully comprehensible and fluent. It accurately conveys the meaning of the English source text without any major, less major, or minor errors. The terminology is correct, and the style is consistent with the source. There are no issues with punctuation, grammar, spelling, capitalization, formatting, rendering, or length. Therefore, it is entirely usable and deserves the highest score	NO
We went to the tango festival to see the "Grupo Piazzolla"	Poszliśmy na festiwal tanga, aby zobaczyć "Grupo Piazzolla"	Poszliśmy na festiwal tanga, żeby zobaczyć „Grupo Piazzolla”	4 – no errors, perfect translation	The Polish translation "Poszliśmy na festiwal tanga, aby zobaczyć "Grupo Piazzolla"." is fully comprehensible and fluent. There are no major errors, no less major errors, and no minor errors. The use of "aby" instead of "żeby" is a stylistic choice and does not constitute an error. The quotation marks used are not the Polish style („”), but this is a minor formatting difference and does not impede understanding. The translation is entirely usable and accurately conveys the meaning of the English source	NO

Source	MT output	Reference	AI score	AI comment	
Agata says "{1}", which means she is a doctor.	Agata dit "{1}", ce qui veut dire qu'elle est médecin	Agata dit « {1} », ce qui signifie qu’elle est médecin	4 – no errors, perfect translation	No errors found	NO
When I went to Buenos Aires I bought several Carlos Gardel records for my dad	Quand je suis allé à Buenos Aires, j'ai acheté plusieurs disques de Carlos Gardel pour mon père	Quand je suis allé à Buenos Aires, j’ai acheté plusieurs albums de Carlos Gardel pour mon père	4 – no errors, perfect translation	No errors found	NO
for the Goods	for the Goods	N/A	4 – no errors, perfect translation	No errors found	NO

Think Global

Conclusion

- The unreliability of AI results is a potential risk of missing the real issues in production environment.
- AI has a potential but it needs more time to develop.
- Needs more focus.
- It must have a Human in the loop = additional cost.
- It misses contextual understanding on the target language style.

Is AI the new “Human Evaluator”?

vistatec

THANK YOU