

# AI, technology and the spoken word: What are the risks and opportunities for public services and PSIs?



**CENTRE FOR  
TRANSLATION  
STUDIES**

UNIVERSITY OF SURREY

Sabine Braun & Diana Singureanu  
Centre for Translation Studies  
University of Surrey

17 April 2026

# Introduction

## » Setting the scene:

- AI is transforming spoken multilingual communication
- New tools
- New opportunities, new risks
- New questions about what constitutes quality

## » Aim of this presentation:

- Explore **interpreting quality in the age of AI**, with particular emphasis on is **evaluation**
- Review human, automated and hybrid approaches to evaluation
- Highlight key implications for public services and public service interpreters (PSIs)

## » Part 1: Framing of quality, approaches to evaluation

## » Part 2: Empirical study – Evaluating machine interpreting in court

# Why a focus on quality and its evaluation matters for public services

- » Public services rely on **accurate, effective spoken communication**;
- » AI has potential role, but in public service settings, communication failures can entail **serious risks**
- » **Understanding interpreting quality and agreement on how it should be evaluated** is central to deciding on viable language solutions (including but not limited to the use of AI)
  
- » Focus on quality and its evaluation
  - ensures professional standards and safety (e.g. healthcare, legal contexts)
  - supports training, certification, and service provision
  - maintains integrity of multilingual communication

# Interpreting quality is multidimensional

- » Not a single construct, but a complex, situation-dependent phenomenon (Grbić, 2008)
- » **Core dimensions**
  - Semantic fidelity | Linguistic quality | Pragmatic appropriateness | Communicative effectiveness
- » **Situational factors**
  - Purpose (training vs professional) | mode (SI/CI) | setting (e.g., conference vs healthcare)
  - Additional dimensions in PSI, e.g.: interactional demands
- » **Whose quality?** Stakeholder priorities differ:
  - Interpreters (fidelity) | Users (clarity) | Commissioners (efficiency) | Trainers (standards)
- **Bottom line:** quality = successful communication in context (Pöchhacker, 2001)
- **In connection with role of AI:** whose definition of “quality” is being applied?

# Human evaluation

## » Criterion-referenced (top-down)

- » Rating grids assessing multiple quality dimensions (accuracy, fluency, delivery, terminology)
- » Used widely in training, certification, and professional assessment (e.g. DPSI)
- » Holistic scoring (single global score) or analytic scoring (separate scores per dimension)
- ✓ multidimensional, captures overall performance
- ✗ resource-intensive, subjective

## » Category-based (bottom-up)

- » Foundational: Barik (1969): omissions, additions, substitutions
- » Bespoke taxonomies, e.g., remote interpreting in legal settings (Braun, 2013); AI interpreting in global health setting (WHO, 2025)
- » Multidimensional Quality Metrics (MQM) (Lommel et al., 2014):
  - Structured error taxonomy with severity levels
  - For translation — needs adaptation for interpreting
- » NTR Model (Romero-Fresco & Pöchhacker (2017))
  - Originally for AI-assisted live subtitling (respeaking)
  - Distinguishes ‘translation’ vs. speech recognition errors
  - Weights errors by severity: minor / major / critical
  - Increasingly applied to interpreting quality
- ✓ fine-grained, diagnostic; facilitates comparison across language pairs
- ✗ focus can be narrow (accuracy); also resource-intensive

# Human evaluation – criterion-referenced

## Example: DPSI

Skill	Expert	Competent	Minimum Expected	Below minimum		Communication	0%	Overall Total	0%	
Score	10	8	6	0		Technique	0%			
Description	<i>Demonstrates complete mastery of subject matter, interpreting skills and behaviours</i>	<i>Demonstrates competence of subject, skills and behaviours with maximum of three minor errors</i>	<i>Demonstrates passable subject knowledge and ability with no major errors</i>	<i>Makes Serious errors or omissions. Clearly does not have full understanding of subject</i>	Score	Comments				
Communication	Interprets exact meaning of the message	Interprets meaning of the message with only one or two minor errors	Adequate interpretation of message meaning	Partial or no interpretation of message meaning						
	Does not add to or leave anything out	One or two minor omissions or additions	No serious omissions or additions	One or more serious omissions or additions						
	Does not change the message	Remains impartial	Remains impartial	No impartiality						
	Remains impartial and interprets message in entirety (faithfully)	Message not 100% faithful but the overall meaning is not altered	Overall meaning of message is not altered	Meaning of message is altered						
Technique	Demonstrates perfect competence in both languages	Demonstrates good competence in both languages	Demonstrates adequate competence in both languages	Inadequate competence in either language						
	Reflects tone and intonation perfectly	Reflects tone and intonation most of the time	Tone and intonation does not have a detrimental effect on the interpreting	Tone and intonation is incorrect and inappropriate						
	Accurately reflects non-verbal communication	Reflects non-verbal communication most of the time	Attempts to reflect non-verbal communication most of the time	No attempt to reflect or inaccurate portrayal of non-verbal communication						
	Is totally professional and confident	Is mainly professional and confident	Professional and confident on occasions	Unprofessional or not confident						
	Controls the session and	Reasonable controls of the	Does not always intervene or asks	No attempt at						

# Human evaluation – category-based

## Example: WHO evaluation of machine interpreting

### Evaluation criteria

<b>Scope</b>	Speech-to-speech translation All 6 WHO official languages (Arabic, Chinese, English, French, Russian, Spanish)
<b>Test Material</b>	Recordings from 77th World Health Assembly (May 2024), selected for <b>accents, figures, acronyms, cultural references, speed, syntax</b>
<b>Evaluation Team</b>	WHO Interpreting & Translation staff + current/former University of Geneva interpreting teachers
<b>Quality Criteria</b>	Based on UN competitive exam standards; <b>content prioritised</b> , expression/delivery considered if meaning impacted; <b>reputational risk as a category</b>
<b>Grading</b>	<b>Segment-based</b> (1–2 sentences), 3-tier scoring (Good 1/Poor 0.5 /Unacceptable 0) 75% passing grade

# Human evaluation – category-based

## Example: WHO evaluation of machine interpreting

### Evaluation criteria

<b>Scope</b>	Speech-to-speech translation All 6 WHO official languages (Arabic, Chinese, English, French, Russian, Spanish)
<b>Test Material</b>	Recordings from 77th World Health Assembly (May 2024), selected for <b>accents, figures, acronyms, cultural references, speed, syntax</b>
<b>Evaluation Team</b>	WHO Interpreting & Translation staff + current/former University of Geneva interpreting teachers
<b>Quality Criteria</b>	Based on UN competitive exam standards; <b>content prioritised</b> , expression/delivery considered if meaning impacted; <b>reputational risk as a category</b>
<b>Grading</b>	Segment-based (1–2 sentences), 3-tier scoring (Good 1/Poor 0.5 /Unacceptable 0) 75% passing grade

### Key findings

- **Overall performance:** Experimental stage; **not fit for external stakeholder meetings**
- **Scores:** 5-83%, only 1/90 segments passed ( $\geq 75\%$ ), overall avrg. **46%**
- **Reputational risk:** Every interpretation contained risks; even one risk = eliminatory
- **Best/Worst languages:** English highest average (51%) but most reputational risks; French lowest (36%)
- **Major errors:**
  - **Latency and omissions:** Delays up to 32s; sentences missed
  - **Misgendering and proper nouns:** Errors creating reputational risk
  - **Language identification and code-switching:** Multilingual series averaged 40%, bilingual 48%
  - **Figures, technical terms:** Mistranslations, numbers misread
    - **Delivery and expression:** Monotonous, literal, pronunciation errors affected comprehension
- **Recommendation:** Only use **internally with staff fluent in languages present**; baseline established for future evaluation

# Automated metrics

- » Traditional (rule-based): e.g., BLEU
  - Estimate **surface-level similarity** (lexical/semantic overlap) between source and output
  - Widely used in machine translation and in some interpreting research
- » Neural (learnt): e.g., COMET
  - **Trained on human judgements**; reference-based and reference-free (quality estimation) modes
  - xCOMET: Adds MQM-style error localisation; highlights where errors occur
- » Key advantages: **scalability**, i.e. ability to evaluate large datasets quickly; **comparison** between systems
- » **General Limitations:**
  - Developed for translation, not interpreting
  - Typically require reference translations (resource-intensive)
  - Penalise legitimate reformulation, condensation, and other outcomes of strategic interpreter behaviour
- » **Interpreting-specific limitations:**
  - Sentence-level equivalence assumption is not suited to interpreting, where meaning is made at discourse level
  - Not captured: delivery features, interactional dynamics, turn-taking, pragmatic meaning, risk (clinical, legal, reputational etc.)
- Automated metrics measure **text similarity, not interpreting quality as a whole**
- High scores  $\neq$  validity for interpreting quality

# Automated metrics

*The witness said the suspect was seen leaving at 22:45 before the incident.*

System Output	Translation (FR)	BLEU (surface similarity)	COMET (meaning-based)	Key points
Reference translation	Le témoin a déclaré que le suspect avait été vu quittant les lieux à 22 h 45 avant l'incident.	—	—	Written gold standard for MT evaluation
Output A (time error)	Le témoin a déclaré que le suspect avait été vu quittant les lieux à 20 h 45 avant l'incident.	Moderately high score (strong overlap despite error)	Low score (detects incorrect time: 22:45 ≠ 20:45)	numerical/time distortion
Output B (role/entity shift)	Le policier a déclaré que le suspect avait été vu quittant les lieux à 22 h 45 avant l'incident.	Moderately high score (shared structure retained)	Low score (detects wrong speaker role: witness ≠ police officer)	entity substitution
Output C (paraphrase – correct meaning)	Le témoin a indiqué avoir aperçu le suspect quitter les lieux à 22 h 45 avant l'incident.	Lower score (penalised lexical variation)	High score (semantic equivalence preserved)	BLEU likely to penalise valid paraphrase
Interpreting D (formal spoken)	Le témoin a dit avoir vu le suspect quitter les lieux à 22 h 45 avant l'incident.	Slightly lower than reference (wording differs)	High score (meaning preserved)	Natural courtroom interpreting, reduced surface similarity
Interpreting E (more informal spoken)	Le témoin a dit que le suspect est parti des lieux à 22 h 45 avant l'incident.	Lower score (lexical/syntactic deviation)	High score (meaning still preserved)	More spontaneous spoken reformulation

# LLM-based evaluation

- » LLMs (e.g. GPT-style models) can generate human-like (though not necessarily reliable) annotations of problems in interpreting output
- » Wang & Fantinuoli (2024): GPT-3.5 shows strong alignment with human judgment for semantic similarity in simultaneous speech translation evaluation

» Automated metrics vs. LLMs:

	<i>Metrics</i>	<i>LLMs</i>
Stability	High	Variable
Reproducibility	High	Lower
Flexibility	Fixed (exception: xCOMET)	High; can generate explanations

# User-based evaluation

» Focus: **end-user perspective**

» Methods:

- Surveys
- Validated scales (e.g. LASI in healthcare; Pathak et al., 2021)
- Comprehension tests

» Foundation: Kurz (1993), Moser (1996) showed that user expectations vary by professional background and event type

✓ captures **real-world usefulness**

✗ **users may not detect errors**; may judge on fluency rather than accuracy

➤ Gap: Reception studies for AI interpreting - critical research need

# AI interpreting challenges

- » **Fundamental differences between human and AI interpreting: AI systems have limited situational/cultural awareness, no intentionality, professional agency, accountability**
- » How should AI interpreting be evaluated:
  - using same standards as human interpreting?
  - using different, context-specific standards?
- » Should the evaluation take into account the context in which AI interpreting is used:
  - role of AI (replacement vs support tool)
  - availability of alternatives (interpreter available vs resource/time constraints, e.g. emergencies)
  - communicative setting (high-risk vs low-risk)
  - user expectations (professional vs “good enough”)
- » Do existing evaluation categories adequately capture AI behaviour?
- **As with human interpreting, AI quality cannot be defined universally; it is contingent on context, purpose, and acceptable risk.**
- **However, its evaluation must capture AI-specific behaviours and limitations – methodological development needed (as part of research)**

# The artificial court interpreter: Machine interpreting and fairness of justice

## Research Gap

### Untested high-stakes integration

Little research on **user perceptions**

Digital gap exacerbates **social inequalities**

Lack of **human-centric, communicative evaluation**

**Why? Fairness of access to justice may be jeopardised as a result of AI solutionism**

## Research Question:

To what extent can MI be a viable alternative to HI in legal settings?

## Aim

Assess viability of MI in legal settings

Develop a **human-centred evaluation approach**

## Methodology

Controlled simulations (**MI, HI and CAI**)

**Human-Centric Quality Assessments + Automated Scoring**

**Users' comprehension (lay users and legal professionals)**

**Stakeholders' perceptions**

**Defining safe practices**

# Simulation Design & Materials

Moot court simulations replicating **authentic legal settings**

Inclusion of **natural speech features** and **interpreting challenges** (terminology, register)

English <> Romanian scenarios covering **common offences** and **key courtroom interactions**

**Law students and native speakers**



## Materials

Part 1: Opening

Part 2: (Cross)examination

Part 3: Closing speeches

Script 1: 34'95"

WPM 128

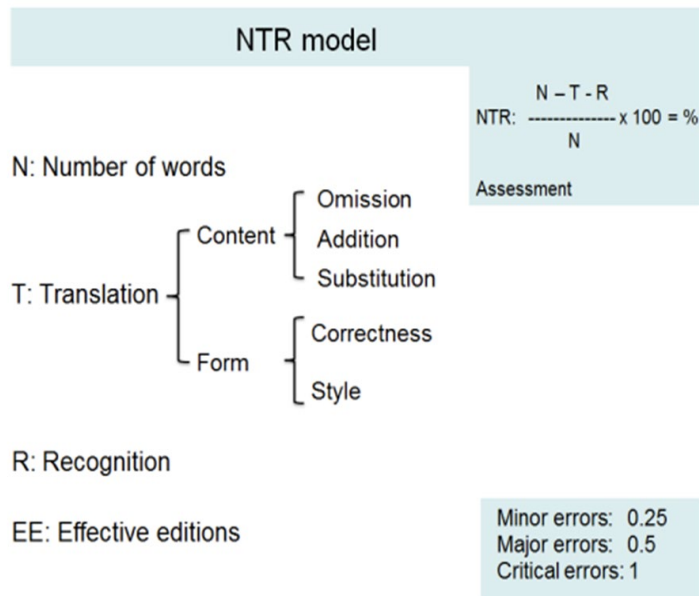
Flesch Reading Score 66.25

### MI Test & Data Collection

**Domain-level customisation only** (general legal terminology; no case-specific tuning)  
Limited vocabulary input (**400 terms** + proper names)

**Minimal expected impact** of generic terminology **on system performance**  
Recordings **submitted immediately prior to testing** to ensure experimental control

# Human-Centric Quality Assessments & Automated Scoring



Romero-Fresco & Pöchhacker, F. (2017)

Metric	Type	Compares	What it measures	Trained on
BLEU	Surface	MT vs ref	Word overlap	No training (rule-based)
ChrF	Surface	MT vs ref	Character overlap	No training (rule-based)
TER	Surface	MT vs ref	Edit distance	No training (rule-based)
BERTScore	Neural (no training)	MT vs ref	Embedding similarity	Pretrained BERT on large text (self-supervised language modeling)
BLEURT	Neural (trained)	MT vs ref	Learned quality scoring	Synthetic errors + human-rated translations
COMET	Neural (trained)	Source + MT	Human-like quality prediction	Human evaluation scores of translations (supervised learning)

Vanroy, B., Tezcan, A., & Macken, L. (2023)

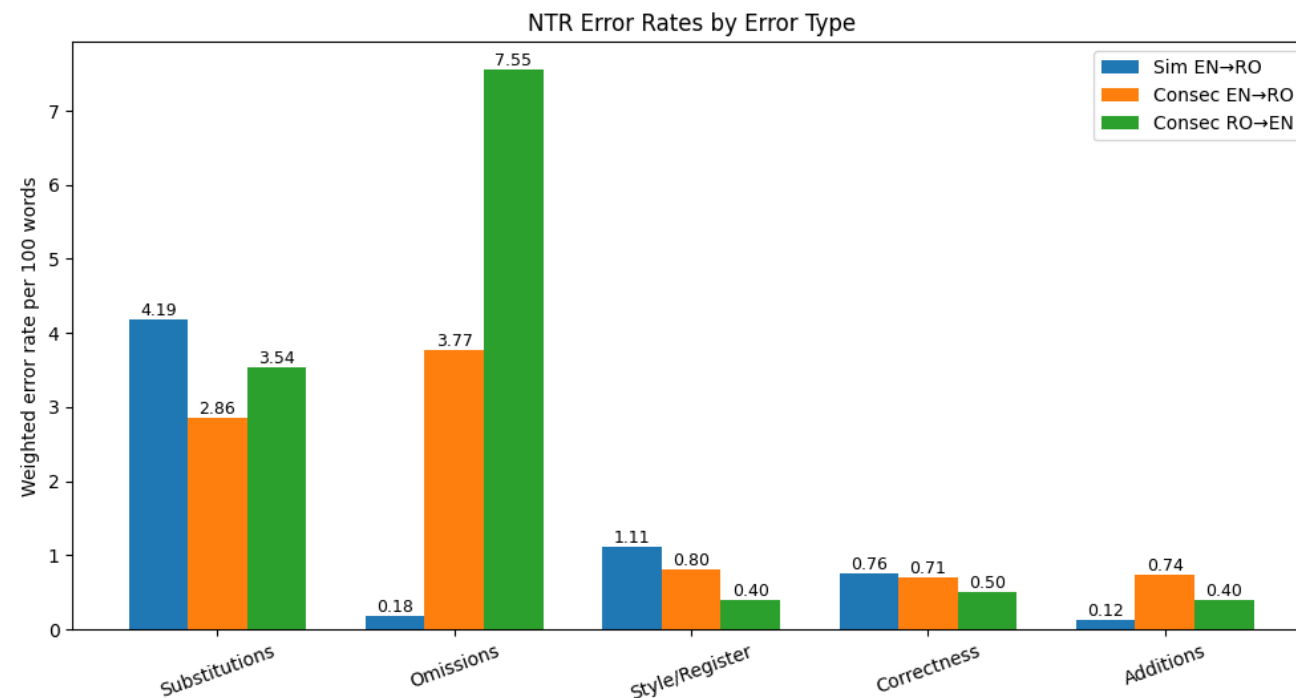
# NTR - Human-Centric Quality Assessment

Source Speech (Sim) EN	Target Speech MI RO	NTR assessment	NTR scoring
<p>The defendant is presumed innocent unless and until the evidence satisfies you, so that you are sure, that he dishonestly appropriated property belonging to another, intending permanently to deprive the owner of it.</p>	<p>Inculpatul este considerat nevinovat decât și până când probele te mulțumesc astfel încât să fii sigur că a însușirit în mod necinstit proprietăți aparținând altcuiva, intenționând să-l priveze permanent pe proprietar.</p> <p>The accused is considered <u>not guilty than and until</u> the evidence <u>pleases you</u> such that you are sure that he has appropriated in a dishonest way <u>property</u> belonging to someone else, intending to deprive permanently the owner.</p>	<p>Content - Subs. Serious - The intended legal meaning of being presumed innocent until proven guilty is not reliably recoverable for a lay listener and may imply that the defendant is innocent as a matter of fact.</p> <p>Form – Correctness (Standard): „decât și până când” - Ungrammatical / non-idiomatic connective; disrupts the logical condition (unless and until), affecting clarity of the legal test.</p> <p>Content - Subs. Serious: „probele te mulțumesc”. This substitutes the legal test (evidence satisfies you) with a notion of personal satisfaction.</p> <p>Content - Subs. Serious: The English source addresses the court as decision-maker (“ so that you are sure”), i.e. the judge. The Romanian rendering shifts this to an informal, second-person singular addressee, with no explicit institutional referent (instanța / dumneavoastră).</p> <p>Form – Correctness Minor: form of the verb ‘a însuși’ is incorrect (non-existent / malformed form)</p> <p>Content - Subs. Standard: ‘bunuri’ is the established term for property capable of appropriation, whereas „proprietăți” risks suggesting land or title rather than movable property.</p> <p>Content – Omission. Minor: ‘of it’ is missing in TT. The RO sentence ends abruptly.</p>	<p>3x Substitution Serious = 3</p> <p>1 Substitution Standard = 0.5</p> <p>1 Omission minor = 0.25</p> <p>Form error – standard = 0.5 deduction</p> <p>Form error minor = 0.25</p> <p>Total deduction: 4.50</p>

# NTR Quality Assessment of MI output

## NTR weighted error rates per 100 words by dataset

Error category	Sim EN→RO	Consec EN→RO	Consec RO→EN
Substitutions	4.19	2.86	3.54
Omissions	0.18	3.77	7.55
Style/Register	1.11	0.80	0.40
Correctness	0.76	0.71	0.50
Additions	0.12	0.74	0.40
<b>Total error rate</b>	<b>6.37</b>	<b>8.86</b>	<b>12.38</b>



# Comparative NTR Quality Assessment: MI Performance vs. Human Professional Standards

	<b>Sim EN→RO</b>	<b>Consec EN→RO</b>	<b>Consec RO→EN</b>
<b>Total error rate</b>	6.37	8.86	12.38
<b>Accuracy (%)</b>	93.63	91.14	87.62

<b>Language Professionals</b>	<b>NTR Score</b>	<b>Source</b>
<b>Professional Conference Interpreters</b>	97.93% – 98.24%	Rodríguez González (2024) CI/RSI (fast / dense speech)
<b>Interlingual Respeakers</b>	96.3% (high performer) 94.4% (low performer)	Davitti & Sandrelli (2020) SMART Project
<b>Machine Interpreting</b>	93.63% - 87.62%	Current dataset

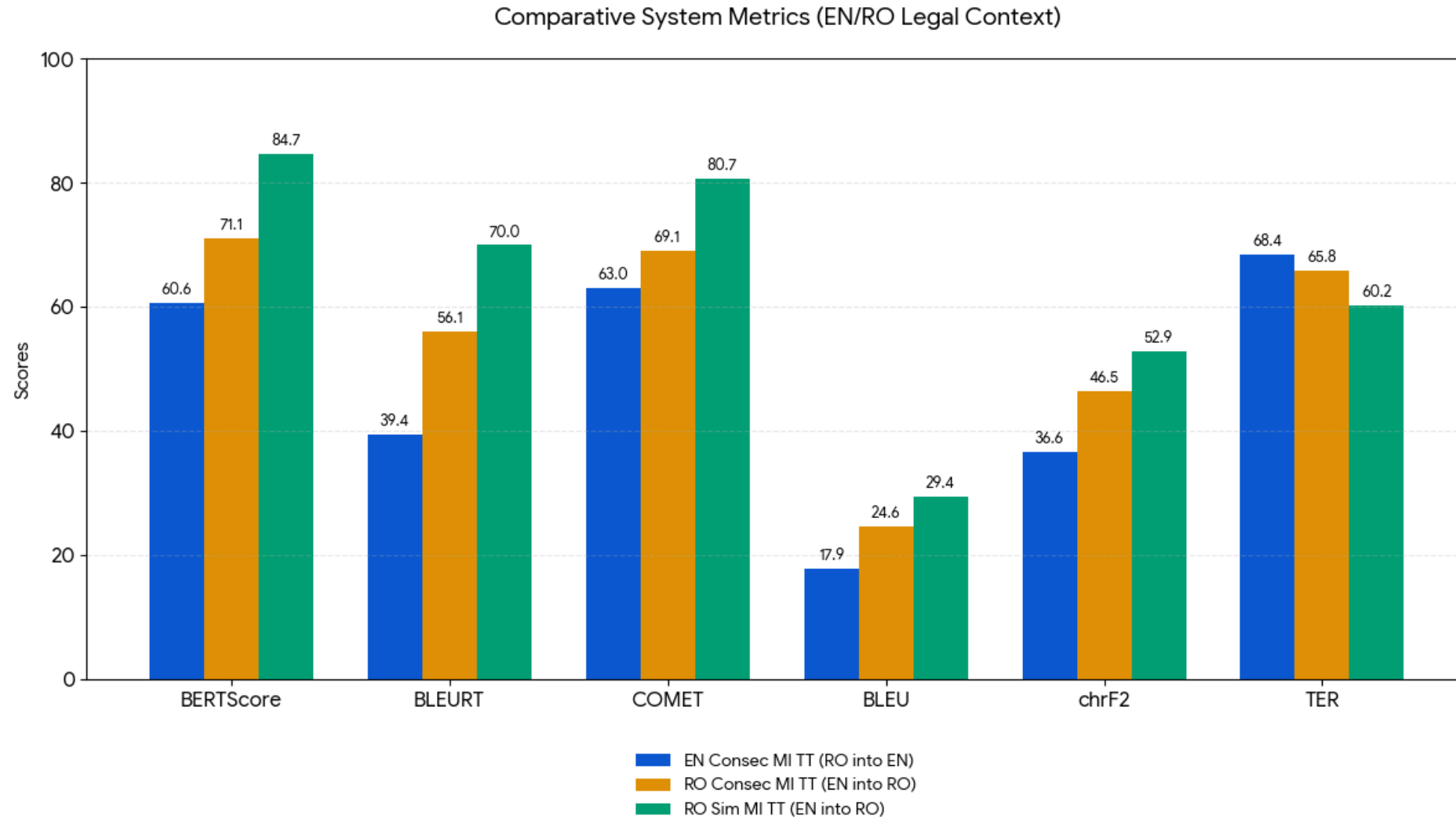
# NTR - Human-Centric Quality Assessment

Source Speech (Sim) EN	Target Speech MI RO	NTR assessment	NTR scoring
<p>The defendant is presumed innocent unless and until the evidence satisfies you, so that you are sure, that he dishonestly appropriated property belonging to another, intending permanently to deprive the owner of it.</p>	<p>Inculpatul este considerat nevinovat decât și până când probele te mulțumesc astfel încât să fii sigur că a însușirit în mod necinstit proprietăți aparținând altcuiva, intenționând să-l priveze permanent pe proprietar.</p> <p>The accused is considered <u>not guilty than and until</u> the evidence <u>pleases you</u> such that you are sure that he has appropriated in a dishonest way <u>property</u> belonging to someone else, intending to deprive permanently the owner.</p> <p>•BERTScore: 81 •BLEURT: 62 •COMET: 60</p>	<p>Content - Subs. Serious - The intended legal meaning of being presumed innocent until proven guilty is not reliably recoverable for a lay listener and may imply that the defendant is innocent as a matter of fact.</p> <p>Form – Correctness (Standard): „decât și până când” - Ungrammatical / non-idiomatic connective; disrupts the logical condition (unless and until), affecting clarity of the legal test.</p> <p>Content - Subs. Serious: „probele te mulțumesc”. This substitutes the legal test (evidence satisfies you) with a notion of personal satisfaction.</p> <p>Content - Subs. Serious: The English source addresses the court as decision-maker (“ so that you are sure”), i.e. the judge. The Romanian rendering shifts this to an informal, second-person singular addressee, with no explicit institutional referent (instanța / dumneavoastră).</p> <p>Form – Correctness Minor: form of the verb ‘a însuși’ is incorrect (non-existent / malformed form)</p> <p>Content - Subs. Standard: ‘bunuri’ is the established term for property capable of appropriation, whereas „proprietăți” risks suggesting land or title rather than movable property.</p> <p>Content – Omission. Minor: ‘of it’ is missing in TT. The RO sentence ends abruptly.</p>	<p>3x Subs. Serious = 3</p> <p>1 Subs. Standard = 0.5</p> <p>1 Omission minor = 0.25</p> <p>Form error – standard = 0.5 deduction</p> <p>Form error minor = 0.25</p> <p>Total deduction: 4.50</p>

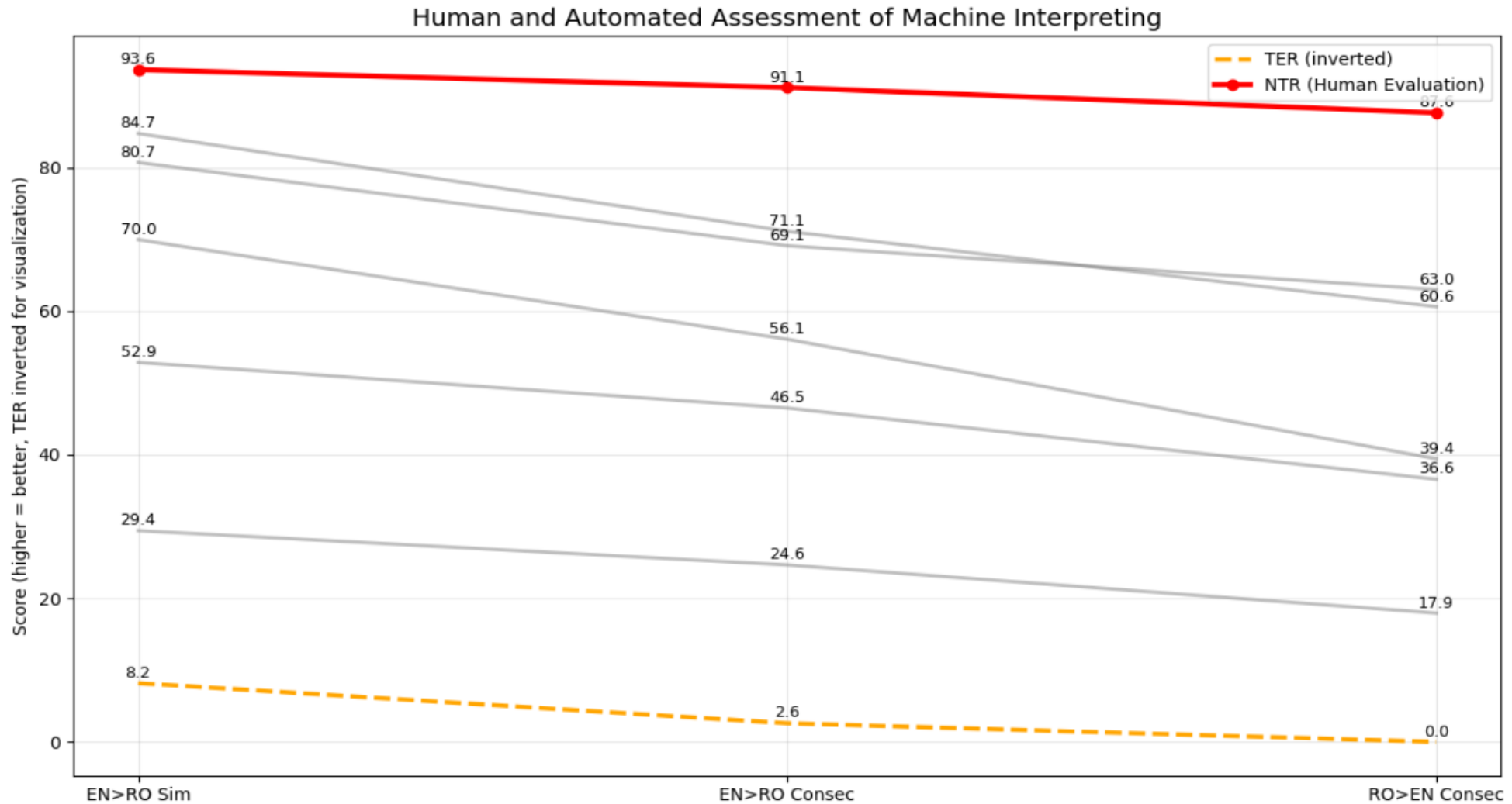
# Semantic Similarity ≠ Legal Usability

Failure Type (Meaning-Based Classification)	Source (English)	System Output (RO + Back-translation)	Automated metrics evaluation
Meaning Deletion (Omission)	<b>“Where had you been earlier that day?”</b>	(no output)	Bertscore 0; Bluert 7; Comet 36
Propositional Collapse (Residual Lexical Output)	<b>“Can you tell the court whether you have children?”</b>	Copii. <b>[Children]</b>	Bertscore 57; Bluert 16; Comet 40
Meaning Distortion (Substitution)	<b>Thank you. You may read it.</b>	Mulțumesc. Poate o citești. <b>[Thank you. Maybe you read it.]</b>	Bertscore 84; Bleurt 80; Comet 92
Meaning Fragmentation (Loss of Relational Structure)	<b>“What did the officer say or do when you woke up?”</b>	Când te-ai trezit? <b>[When did you wake up?]</b>	Bertscore 77; Bleurt 63; Comet 76;
Meaning Breakdown (Hallucination)	<b>“How did you come to be in possession of the bag?”</b>	Probabil va veni intraria, posesia genții, eu. <b>[Probably will come the in-between-thing, the possession of the bag, me.]</b>	Bertscore 69; Bleurt 36; Comet 43;
	<b>“What did you do before picking up the clothes?”</b>	O parca de recopiere a lui hini. <b>[A kind-of of re-copying of hini.]</b>	Bertscore 64; Bleurt 24; Comet 28;

# Automated Metrics Scores Across Modes and Language Direction



# Trend consistency across evaluations



# Stability and Consistency across modes and directions

Mode & Direction	BERTScore	BLEURT	COMET
	It checks whether the meaning of words in the output matches the reference. It can overestimate similarity when words share a root or are semantically close.	It's trained on human judgments and tends to reward outputs that resemble natural language usage. Largely trained on general-domain data.	It uses a trained model that considers the source sentence, the translation, and a reference to judge adequacy and fluency but may overlook nuance.
Sim (EN > RO)	84.7 (± 1.1)	69.9 (± 2.8)	80.7 (± 2.2)
Consec (EN > RO)	71.1 (± 5.5)	56.2 (± 5.7)	69.1 (± 4.1)
Consec (RO > EN)	60.6 (± 7.8)	39.4 (± 5.5)	63.0 (± 3.4)

# Rethinking Evaluation

Evaluation models were developed for **human interpreting**

**Machine interpreting** introduces **different error patterns**:

- Omissions vs. Hallucinations: MI often struggles with critical information loss or generates confident, yet entirely invented, content.
- Fluent but Inaccurate Content: High linguistic fluency masks significant semantic inaccuracies.
- Instability vs. Consistency: Automated output often lacks the consistent quality of human professionals, showing high variance.

**Limitations:**

Current frameworks may:

- misrepresent MI performance
- fail to capture communicative breakdown

# Discussion & conclusion

## » **Which approach?**

- Human, automated, and user-based approaches - all relevant for **human, AI-assisted, AI interpreting**

## » **Layered, hybrid, integrated approaches useful**, esp. for large datasets, e.g.:

- Automated (initial screening) | human (depth/safety) | user-based (professional perspective)
- Currently, lack of integration; need for multidimensional, multi-method approaches, covering human interpreting, AI systems and hybrid, human-AI collaborative workflows (Han, 2025)

## » **Emerging use cases of automated methods:**

- screening large datasets | trainee self-assessment | continuous monitoring
- However, coverage spoken-language features, interaction, embodied behaviour remains problematic

## » **Human evaluation remains indispensable, esp. when AI outputs are involved (!)**

## » **Relevance for PSIs**

- Knowledge about quality evaluation – part of continuous professional development
- Informs negotiations with clients (which language solution), client consulting (QA, monitoring)
- Supports emerging professional profiles – management of multilingual communication workflows

# ADAPTING TO CHANGE



**CENTRE FOR  
TRANSLATION  
STUDIES**

UNIVERSITY OF SURREY



## CPD course Translation and AI, next start 24 April

- Research-led, critical understanding of AI in translation
- Practical strategies to manage the impact of AI on professional practice
- Key questions about the effective use of AI tools
- Confidence in evaluating and using AI tools independently



## CPD course Interpreting and AI; next start date 23 April

- Research-led understanding of key AI concepts for interpreters
- AI support for interpreters in preparation and live performance
- Strategies for effective integration of AI tools into interpreting practice
- Informed engagement with AI tools to manage cognitive load and stress

**Overview:** <https://www.surrey.ac.uk/centre-translation-studies/continuing-professional-development>

**Translation and AI:** <https://www.iti.org.uk/training/events-calendar/translation-ai-spring2026.html>

**Interpreting and AI:** <https://www.iti.org.uk/training/events-calendar/interpreting-ai-spring2026.html>



# ADAPTING TO CHANGE



**CENTRE FOR  
TRANSLATION  
STUDIES**

UNIVERSITY OF SURREY



**MA Translation and AI**  
(incl. Chinese Pathway)



**MA Interpreting, Technology  
and AI** (incl. Chinese Pathway)



**MA Translation, Interpreting  
and AI** (incl. Chinese option)



**MSc AI for Translation and  
Interpreting** (onsite, online)

## Specialised Translation

- Language-pair specific practice towards professional level
- Economic, legal, medical, environmental translation
- Other specialisms according to special interests

## Transcreation

- Translation and writing for the creative industries
- Audiovisual translation, e.g. subtitling, game localisation;
- Advertising and the cultural sector, e.g. museum texts

## Interpreting

- Conference, media, business and public service settings
- Remote interpreting via video link
- Hybrid practices: speech-to-text, re-speaking

## Technology in translation and interpreting

- Computer-assisted and corpus-assisted translation/interpreting
- Neural Machine Translation, Speech Technologies
- Large Language Models and Generative AI

<https://www.surrey.ac.uk/centre-translation-studies/study>

<https://www.surrey.ac.uk/centre-translation-studies/study/postgraduate-courses>

## References

- Barik, H. C. (1969).** A study of simultaneous interpretation (Unpublished doctoral dissertation). University of North Carolina at Chapel Hill.
- Braun, S. (2013).** Keep your distance? Remote interpreting in legal proceedings. In C. Schäffner, K. Kredens, & Y. Fowler (Eds.), *Interpreting in a changing landscape* (pp. 67–91). John Benjamins.
- Davitti, E., & Sandrelli, A. (2020).** Embracing the Complexity: A Pilot Study on Interlingual Respeaking. *Journal of Audiovisual Translation*, 3(2), 181–205.
- Davitti, E., Sandrelli, A., Romero-Fresco, P., Korybski, T., Moores, Z., & Wallinheimo, A. S. (2023).** Interlingual respeaking - SMART project [Poster]. University of Surrey. <https://smartproject.surrey.ac.uk/wp-content/uploads/2023/05/SMART-project-poster-WEBSITE.pdf>
- Grbić, N. (2008).** Constructing interpreting quality. *Interpreting*, 10(2), 232–257.
- Han, C. (2025).** Quality assessment in multilingual, multimodal, and multiagent translation and interpreting: Proposing a unifying framework for research. *Interpreting and Society: An Interdisciplinary Journal*, 5(1), 27–55. <https://doi.org/10.1177/27523810251322645>
- Kurz, I. (1993).** Conference interpreting: Expectations of different user groups. *The Interpreters' Newsletter*, 5, 13–21.
- Lin C. (2026).** A hybrid intelligent assessment model for English translation education with improved BERT and SVM. *Scientific reports*, 16(1), 5466. <https://doi.org/10.1038/s41598-026-35042-2>
- Lommel, A., Uszkoreit, H., & Burchardt, A. (2014).** Multidimensional Quality Metrics (MQM): A framework for declaring and describing quality metrics used in quality assurance of translation deliveries. *Tradumàtica: technologies de la traducció*, (12), 455–463.
- Macháček, D., Bojar, O., & Dabre, R. (2023).** MT metrics correlate with human ratings of simultaneous speech translation. *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)* (pp. 169–179). Association for Computational Linguistics.
- Moser, P. (1996).** Expectations of users of conference interpretation. *Interpreting*, 1(2), 145–178.
- Pathak, S., Gregorich, S. E., Diamond, L. C., Mutha, S., Seto, E., Livaudais-Toman, J., & Karliner, L. (2021).** Patient Perspectives on the Quality of Professional Interpretation: Results from LASI Study. *Journal of general internal medicine*, 36(8), 2386–2391. <https://doi.org/10.1007/s11606-020-06491-w>
- Pöchhacker, F. (2001).** Quality assessment in conference interpreting: Methodological issues. *The Interpreters' Newsletter*, 11, 95–101.
- Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020).** COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.213>
- Romero-Fresco, P. (2011).** *Subtitling through speech recognition: Respeaking*. St. Jerome Publishing.
- Romero-Fresco, P., & Pöchhacker, F. (2017).** Quality assessment in respeaking: The NTR model. *Perspectives*, 25(1), 149–167.
- Vanroy, B., Tezcan, A., & Macken, L. (2023).** MATEO: MACHine Translation Evaluation Online. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, ... H. Moniz (Eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (pp. 499–500). Tampere, Finland: European Association for Machine Translation (EAMT).
- Wang, H., & Fantinuoli, C. (2024).** Exploring the Correlation between Human and Machine Evaluation of Simultaneous Speech Translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp 327–336). Sheffield, UK. European Association for Machine Translation (EAMT). <https://aclanthology.org/2024.eamt-1.28.pdf>
- World Health Organization. (2025).** WHO Evaluation of machine interpreting. [https://www.linkedin.com/posts/ghada-chadarevian-444966230\\_who-report-on-ai-interpretation-activity...](https://www.linkedin.com/posts/ghada-chadarevian-444966230_who-report-on-ai-interpretation-activity...)