معهد الدراسات اللغوية والترجمة
Institute of Applied Linguistics and Translation

# Automatic Evaluation vs. Human Verification of English-Arabic Machine Translation

Presented by

Iman Magdy Refaat Hassan Mohamed

A Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy in Arts

Translation, English

Institute of Applied Linguistics and Translation

Faculty of Arts, Alexandria University

May 2025

Under the Supervision of

Prof. Sameh Alansary
Professor of Computational guistics and Head of Phonetics Department
Faculty of Arts, Alexandria University

Prof. Abeer M.Refky M.Seddeek
Professor of English Language and Literature and Dean of College of Language and Communication
Arab Academy for Science, Technology, and Maritime Transport

Prof. Nevine Sarwat
Professor of Linguistics and Translation and Manager of Institute of Applied Linguistics and Translation
Faculty of Arts, Alexandria University

## Abstract

This study investigates the reliability of automatic machine translation (MT) evaluation and estimation metrics, both reference-based and reference-free, in comparison with human evaluation methods, namely rating and error typology, for assessing English-Arabic MT output. Combining quantitative and qualitative methodologies, the research begins with a survey and semi-structured interviews involving MT developers, researchers, localizers, and more to explore evaluation practices across the industry in order to decide the tools and domains needed in the research. It proceeds with an experimental phase comparing human and automatic evaluations of translations generated by Google API and ChatGPT across three domains: Business-HR, Life Science-Pharmaceutical, and Gaming, totaling 34,696 words and 3,643 segments. Human evaluation employed Scalar Quality Metric (SQM) for rating and the harmonized version of Data Quality Framework and Multidimensional Quality Metric (DQF-MQM) for error typology, while automatic metrics included MT quality evaluation metrics (reference-based), BLEU, BERTScore, BLEURT, COMET, and MetricX-23-XL, and MT quality estimation metric (reference-free), CometKiwi. Findings reveal that top-performing automatic metrics offer reliable assessments, especially in technical domains; however, they struggle with idiomatic or creative content, fail to detect untranslated segments, and misevaluate good translations by providing low scores thereto. Human evaluation, particularly DQF-MQM, remains essential for identifying nuanced errors and optimizing both MT systems and automatic metrics. While rating methods like SQM align closely with DQF-MQM, they lack diagnostic depth. The study concludes that automatic metrics cannot fully replace human evaluation, and reference-based metrics, especially COMET, generally outperform reference-free alternative; though CometKiwi shows strength in specific technical domains.



**Keywords:**

Machine Translation, Machine Translation Quality Metrics, Automatic Machine Translation Quality Evaluation, Automatic Machine Translation Quality Estimation, Human Verification, Rating, Error Typology