Beyond Post-Editing: A Project-Based Module on MT and LLM Integration for Trainee Translators

Alina Karakanta
Leiden University Centre for Linguistics, Leiden University
a.karakanta@leidenuniv.nl

Abstract

With the rapid technologisation of translation, skills beyond post-editing (PE), such as data literacy, technology evaluation, and critical engagement with AI-based tools are becoming essential competencies for trainee translators. This paper presents a syllabus for a translation technology module that equips MA Translation students with end-to-end MT evaluation skills, from engine selection and domain adaptation to automatic and human evaluation, post-editing, and results reporting. Large language models (LLMs) are integrated throughout, as translation engines, as a basis for prompting and domain adaptation strategies, and as a source of explainable quality estimation signals. Grounded in project-based learning, students apply these skills in a simulated client scenario. Trends on students' engine selection and customisation preferences observed across three years suggest a gradual shift towards LLM-based tools, but traditional engines and built-in options remain a firm presence in student practice.

1 Introduction

Translation is more technological than ever. The widespread adoption of neural machine translation (NMT) and, more recently, large language models (LLMs) has fundamentally altered how translation is produced, reviewed, and delivered. Translators increasingly operate within complex, multi-step workflows where MT output passes through quality estimation filters, automated post-editing suggestions, and human review before reaching its final form. CAT tools have evolved accordingly, now embedding quality signals, adaptive suggestions, and LLM-powered features directly into the translation interface.

This rapid shift has significant implications for translator training. The professional landscape is diversifying: while many translators continue to work with MT as end users, there is growing demand for adjacent roles such as MT specialist and language technology consultant. Even for those who remain primarily in translation practice, critical awareness of how MT systems work, how they are evaluated, and how they can be adapted is increasingly expected. In this age, critical skills are more relevant than ever. Skills such as data literacy, technology assessment, and evidence-based reasoning about MT quality are no longer the exclusive domain of computational linguists but are becoming relevant across a range of profiles that translation graduates may occupy.

Yet translator training has been slow to reflect this breadth. Existing syllabi have made significant progress in covering the theoretical underpinnings of MT, pre-editing, human and automatic evaluation, project management, and post-editing as a core professional skill (Doherty and Kenny, 2014; Moorkens, 2018; Guerberof Arenas and Moorkens, 2019; Mellinger, 2017; Pym and Hao, 2024). However, the pace of technological development and particularly the rise of LLMs means that the field is still figuring out how to incorporate these tools meaningfully into translator education. The role of LLMs within the MT pipeline, whether as translation engines, customisation tools, or sources of explainable quality signals, remains under-represented in published ped-

agogical proposals.

This paper addresses that gap by presenting a translation technology module grounded in project-based learning, in which students develop end-to-end MT evaluation competencies across seven sessions. The module combines commercial and open-source tools incorporating state-of-the-art MT engines and LLM-based tools. The approach is flexible in that its underlying framework is adaptable as new technologies emerge.

2 Research on teaching translation technologies

2.1 Skills

Early work on teaching translation technologies established a template that has proven durable. O'Brien (2002) presented one of the first systematic proposals for teaching MT, combining theoretical knowledge of MT, terminology, and pre-editing with practical elements such as PE with different engines and programming macros. Subsequent course designs followed this integrated spirit with topics rangin from MT evaluation, controlled language, PE of MT output (Doherty et al., 2012; Kenny and Doherty, 2014), ethics, payment, collaboration, the role of the human translator (Moorkens, 2022), and translation workflows, to PE effort and process research (Koponen, 2015). Doherty and Kenny (2014) offered a full course design incorporating a range of technology-related competencies alongside translation practice. Guerberof Arenas and Moorkens (2019) designed a machine translation and post-editing course, that includes two modules: one on PE and one on MT project management. TAUS PE guidelines and course.

Compared to PE skills, evaluation has only recently received increasing attention in translator training. With the advent of neural MT, Moorkens (2018) presented a proposal for a practical in-class translation evaluation exercise where students carried out evaluations of neural and statistical MT using translation quality assessment: adequacy, post-editing effort (time), and a simple error taxonomy (word order errors, mistranslations, omissions, and additions). Macken et al. (2023) proposed incorporating evaluation with automatic metrics into translator training, by developing an open-source, accessible evaluation platform (Vanroy et al., 2023).

Hands-on work with MT systems has also been supported by dedicated resources. Two seminal coursebooks on statistical and neural MT (Koehn, 2010; Koehn, 2020) have provided theoretical and practical grounding for courses engaging with engine-level knowledge, however these are often inaccessible to translators due to their lack of technical and programming knowledge. To address this this technical barrier, the MultiTraiNMT project[1] developed a coursebook (Kenny, 2022) and an online platform allowing students to train, fine-tune, and evaluate MT systems without programming skills. Krüger (2022) approached the same challenge through prepared Jupyter notebooks, enabling students to work with MT data and evaluation pipelines hands-on without requiring a programming background.

The concept of MT literacy (Bowker and Ciro, 2019), encompassing not just the ability to use MT tools but to understand, evaluate, and critically engage with them, has been influential in framing what translators need to know. Hackenbuchner and Krüger (2023) extend this to data literacy by proposing DataLit^{MT}. which includes a Professional MT Literacy Framework, an MT-specific data literacy framework and a competence framework for Machine Translation-Oriented Data Literacy Teaching. They argue that professional machine-translation literacy encompasses various skills and knowledge areas, including technical, linguistic, economic, societal, and cognitive aspects.

2.2 Pedagogical approaches

Research broadly agrees on the value of combining a theoretical component with hands-on practice. Lecture-based methods are typically used to convey background knowledge such as MT history and system types, while practical skills are developed through student-centred activities ranging from task-based group work to large-scale simulated projects (Hao and Pym, 2023). (Mellinger, 2017) making the case for embedding technology throughout the translation curriculum rather than treating it as a standalone topic. Project-based and authentic assessment approaches have also been advocated, with students working on realistic scenarios that mirror professional practice (Kiraly, 2005; Li et al., 2015; Mitchell-Schuitevoerder, 2020).

[1] https://www.multitrainmt.eu/

2.3 GenAI and LLMs

The integration of generative AI and LLMs into translation education is still emerging. Pym and Hao (2024) offers a guide to using generative AI for language learning activities, while Ayvazyan et al. (2024) provides a practical inventory of activities involving generative AI for translators. These contributions signal growing interest in the area, but systematic pedagogical proposals for integrating LLMs into translator training, particularly within an end-to-end MT evaluation framework, remain scarce.

3 Context

The module described in this paper forms the second block of Translator's Tools, a first-year course in the MA Translation programme at Leiden University. The first block introduces students to CAT tools, terminological databases, quality assurance, and cloud-based collaboration interfaces, assessed through an in-class CAT tool assignment. Students therefore enter the second block with a working knowledge of translation workflows, translation memories, terminology management, and cloud and collaborative tools. The second block, the focus of this paper, covers language technologies, machine translation, customisation, evaluation, post-editing, and quality estimation.

The module consists of seven weekly two-hour sessions, combining a theoretical and a practical component. Sessions take place in a translation lab with access to high-performing laptops. Licences for commercial MT engines and CAT tool integrations are obtained through the institution's resource infrastructure or through educational agreements with providers for the duration of the block. The module has been taught over three consecutive years to cohorts of approximately 25 students per year, allowing for iterative refinement based on student feedback and developments in the field.

4 Approach

The module is grounded in project-based learning, an approach that seeks to foster independent thinking, teamwork, and problem-solving in unfamiliar situations, with a view to acquiring transferable skills that remain relevant regardless of technological change (Kiraly, 2005). Rather than strictly following any single pedagogical model, the module draws on the spirit of Kiraly's empowerment-oriented approach (Kiraly, 2005), in which students take ownership of a complex, realistic task and develop competencies through doing. This is complemented by formative assessment: students submit weekly individual exercises based on the practical activities of each session, receiving written feedback or participating in class discussions that address common tendencies across the cohort. The summative assessment is a group project in which students propose and evaluate an MT solution for a simulated client scenario, described in detail in Section 4.3.

While the weekly exercises develop discrete skills progressively, the project brings these together into an end-to-end MT evaluation pipeline: students select engines, perform customisation, conduct automatic and human evaluation, carry out post-editing, and report their findings and recommendations. This integrated pipeline is the central pedagogical contribution of the module. Rather than treating each skill in isolation, students experience how the components of MT evaluation relate to and inform one another in practice.

Large language models are integrated throughout the module, serving three distinct functions. First, LLMs are included alongside conventional MT engines in the engine selection and comparison activities, giving students direct experience of their translation behaviour and characteristics. Second, LLMs are used as a basis for prompting strategies in domain adaptation, including retrieval-augmented generation approaches, allowing students to explore customisation methods that go beyond the translation memory and glossary integration available in traditional MT systems. Third, in the quality estimation session, LLMs provide explainable quality signals (word- and segment-level error highlights and automatic post-editing suggestions) that students learn to interpret and use to support their post-editing decisions. This integration reflects the current professional landscape, in which LLMs are increasingly embedded in MT workflows, and prepares students to engage with them critically rather than as passive end users.

The module uses a combination of commercial and open-source tools throughout, exposing students to the range of options available in professional practice. The underlying framework is designed to be flexible: as new engines, metrics, and LLM-based features emerge, individual components can be updated without disrupting the overall

structure.

4.1 Learning objectives

By the end of the module, students will have developed a comprehensive set of competencies spanning the full MT pipeline. They will understand how MT systems and LLMs work and be able to identify their role in contemporary translation workflows, as well as prepare and manage the data resources these systems rely on, including corpora, translation memories, and glossaries. Students will be able to customise MT systems using different types of data and prompting strategies, and assess the quality of MT output through both automatic metrics and human evaluation protocols. They will practise post-editing in accordance with professional guidelines, and learn to interpret quality estimation signals, such as explainable LLM-based outputs, to support post-editing decisions. Finally, students will gain experience in conducting, analysing, and reporting on an end-to-end MT evaluation project situated in a realistic professional scenario.

4.2 Session descriptions

The module is organised in 7 sessions. Each session includes a theoretical part, explaining basic notions and clarifying the readings, and a practical session with hands-on activities.

Session 1. Introduction to MT and language technologies

This session introduces machine translation by first eliciting students' own conceptualisations of the technology, before covering its evolution from rule-based and statistical approaches to neural MT and large language models. The theoretical component explains how NMT and LLMs work at a conceptual level, surveys language technologies relevant to the translator, such as optical character recognition and automatic speech recognition, and clarifies the distinction between translation memories and MT engines.

Practical activities:

- Explore a range of commercial and open-source MT engines, including DeepL[2], Google Translate[3], Lara[4], Widn[5], Apertium[6], and No Language Left Behind (Costa-Jussà et al., 2022), mapping each to its underlying architecture and noting available functionalities such as image-based translation, speech input, glossary integration, and quality or explanation features.
- Translate two short texts from different domains (institutional and audiovisual) using multiple engines and compare outputs based on translation theory knowledge, without formal evaluation criteria.
- Connect an MT engine to a CAT tool (e.g. Trados Studio) and practise working with MT suggestions within a translation interface.

Formative exercise: Students submit a short comparative reflection on the engine outputs for the two texts, discussing differences in translation behaviour across engines and domains based on their observations.

Session 2. Corpora and glossaries

This session covers the data that is used to build MT systems. The theoretical component explains the role of parallel corpora and glossaries in MT training and customisation, introducing students to the concept of in-domain and out-of-domain data and the principle that data quality directly affects translation quality. Students are introduced to existing parallel corpus resources such as OPUS (Tiedemann, 2009), and learn about common data quality issues including noise, misalignments, and domain mismatch. The session also covers how glossaries can be prepared and formatted for use in MT customisation.

Practical activities:

- Explore the OPUS parallel corpus repository and discuss the types of data available, their origins, and their suitability for different translation domains.
- Perform translation memory cleaning using a TM maintenance tool or Okapi Olifant,[7] identifying and removing noisy or misaligned segments.
- Prepare a domain-specific glossary in the format required for MT customisation, including tab-separated term pairs and any required metadata.

Formative exercise: Students submit a cleaned TM and a prepared glossary based on provided data, with a short reflection on the quality issues identified and the decisions made during cleaning.

[2]https://www.deepl.com/
[3]https://translate.google.com/
[4]https://laratranslate.com/
[5]https://www.widn.ai/
[6]https://apertium.org/
[7]https://okapiframework.org/wiki/index.php?title=Olifant

Session 3. **Machine Translation customisation**
This session covers the difference between generic and domain-adapted MT systems. The theoretical component explains what a domain is, why domain adaptation matters for translation quality, and the main strategies available for customising MT engines. Students learn about three approaches to customisation: fine-tuning with in-domain parallel data, glossary integration, and prompt-based adaptation. The latter is introduced through a demonstration of an API call to a large language model with few-shot translation examples and glossary terms integrated into the prompt, illustrating how LLMs can be steered towards domain-specific translation behaviour without retraining.

Practical activities:

- Customise a MT engine using a domain-specific glossary (e.g. via the DeepL glossary function), comparing output with and without customisation.
- Use an MT engine (e.g. Lara) to translate with an integrated translation memory, observing how the system incorporates existing translations into its output.
- Apply prompt-based customisation by constructing prompts with domain instructions, few-shot examples, and glossary terms, and evaluating the effect on translation output.

Formative exercise: Students submit a comparison of the three customisation approaches applied to a short text, discussing the effect of each strategy on translation quality based on their observations.

Session 4. **Human evaluation of MT**
This session introduces the methods used to evaluate MT quality through human judgement. The theoretical component covers the main human evaluation protocols: adequacy-fluency ratings, direct assessment (Graham et al., 2013), ranking, and error annotation based on Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). The session features a collaborative in-class evaluation activity designed to make the evaluation process concrete and tangible. Large printed sheets are posted around the room, each displaying a source sentence, a human reference translation, and the output of three MT systems. Working together, students collect adequacy-fluency ratings, perform direct assessment, rank the outputs, and conduct MQM annotation, discussing their judgements and disagreements as a group. Students also learn how to select sentences for human evaluation, including strategies for sampling representative and informative segments from a test set.

Practical activities:

- Perform in-class human evaluation using adequacy-fluency ratings, direct assessment, and ranking across multiple MT outputs.
- Conduct MQM annotation, identifying and classifying errors in MT output using an MQM scorecard.
- Discuss inter-annotator agreement, the subjectivity of human judgements, and the implications for evaluation reliability.

Formative exercise: Students submit a completed MQM annotation of a short MT output, including error classification and a brief reflection on the challenges encountered and decisions made during annotation.

Session 5. **Automatic evaluation of MT**
This session introduces automatic MT evaluation as a complement to human judgement. The theoretical component motivates the need for automatic metrics: human evaluation is costly, time-consuming, and difficult to scale. It also covers the main families of metrics in use today. Students learn about surface-level metrics based on word and phrase overlap, i.e., BLEU (Papineni et al., 2002), chrF (Popović, 2015), and TER (Snover et al., 2006), embedding-based metrics that capture semantic similarity, i.e., BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020), and learned metrics trained on human judgement data, i.e., COMET (Rei et al., 2020). The strengths and limitations of each family are discussed, including their varying degrees of correlation with human assessments.

Practical activities:

- Prepare a test set for automatic evaluation, ensuring correct formatting and alignment between source, hypothesis, and reference files.
- Evaluate the output of multiple MT systems using the MATEO platform (Vanroy et al., 2023), applying a range of metrics and comparing results across systems.
- Interpret metric scores, analyse agreements and disagreements across metrics, and discuss the use of statistical significance testing to validate comparisons.

- Reflect on the relationship between automatic scores and the human judgements collected in the previous session, identifying cases of agreement and divergence.

Formative exercise: Students submit an automatic evaluation of two MT systems using MATEO, including a written interpretation of the results across metrics and a discussion of the relationship of the scores obtained with the human evaluation scores from Session 4.

Session 6. **Post-editing**
This session introduces post-editing as a professional practice. The theoretical component covers what post-editing is and how it differs from traditional editing and reviewing, the distinction between light and full post-editing, and professional guidelines and standards. The session also addresses how professional translators experience and perceive post-editing, questions of cognitive effort, productivity, and the implications for professional practice and payment.

Practical activities:

- Discuss the differences between light and full post-editing and when each is appropriate in a professional context.
- Study and apply post-editing guidelines, including the TAUS post-editing guidelines, to a sample MT output.
- Conduct a timed productivity exercise in which students post-edit two short texts from two different MT engines, following a counterbalanced design with two groups. Students record their completion times, which are then aggregated and discussed as a class. The exercise illustrates variability across post-editors, the relationship between engine quality and post-editing effort, and the methodological challenges of designing valid productivity studies.
- Reflect on the types of errors encountered and the decisions made during post-editing, connecting observations to the error typologies covered in Session 4.

Formative exercise: Students submit a post-edited text with tracked changes, accompanied by a short reflection on the types and frequency of errors encountered, the effort required, and their experience of working with MT output in a CAT tool environment.

Session 7. **Quality estimation and human-in-the-loop**
This session introduces quality estimation (QE) as a method for predicting MT quality without a reference translation. The theoretical component explains the difference between quality evaluation, which requires human judgements or reference translations, and quality estimation, which operates on source and MT output alone. Students learn why QE is used in production workflows, how it can be used to triage segments for post-editing, and how recent developments in explainable QE provide word- and segment-level error signals that go beyond a single quality score. Multi-agent workflows are introduced, including LLMs for generating explainable QE outputs, automatic post-editing (APE) suggestions and for quality assurance.

Practical activities:

- Compare quality estimation and quality evaluation, discussing the scenarios in which each is appropriate and their respective limitations.
- Examine the output of state-of-the-art QE models, including xCOMET (Guerreiro et al., 2024) and xTower (Treviso et al., 2024), focusing on word-level error highlights and segment-level quality scores.
- Use SmartPE,[8] an error highlighting and post-editing tool, to practice post-editing with automatic error highlights and translation correction suggestions generated through xTower's automatic post-editing.
- Discuss the explainability functionalities enabled by LLMs in QE tools, critically evaluating the reliability and usefulness of the signals provided.

Formative exercise: Students submit a post-edited text completed using SmartPE, with a reflection on how the quality estimation signals influenced their post-editing decisions, which suggestions they accepted or rejected, and why.

4.3 Machine Translation project

The summative assessment for the module is a group project in which students of 3–4 propose and evaluate an MT solution for a simulated client scenario. The project is designed to bring together the skills developed across the seven sessions into an end-to-end evaluation pipeline, requiring students

[8] https://github.com/fatalinha/SmartPE

to make and justify decisions at each stage rather than simply applying procedures mechanically.

Students are provided with a dataset one month before the deadline, containing a test set of aligned source and target sentences, a translation memory in both .tmx and Trados formats, and a domain-specific glossary. The goal is to answer the research question: which is the best MT engine and setting for this content? Working from this question they proceed through four stages. First, they translate the test set using at least three MT engines of their choice. Second, they perform automatic evaluation using a range of metrics via the MATEO platform (Vanroy et al., 2023), selecting the best-performing baseline engine. Third, they customise one engine using the provided translation memory and glossary, and evaluate the effect of customisation through automatic metrics. Fourth, they select the two best-performing systems and conduct human evaluation on a sample of fourty sentences, choosing from adequacy-fluency rating, ranking, MQM annotation, and perform post-editing.

Project report Students report their findings in a research paper of 3000–4000 words, excluding references and appendices. The introduction situates the project within the relevant literature on MT, domain adaptation, and evaluation; the methods section describes the data, scenario, and evaluation procedures; the results section presents and interprets findings across automatic evaluation, customisation, and human evaluation; and the discussion reflects on the best MT solution for the given content, the reliability of evaluation methods used, and relevant ethical and professional considerations. Human evaluation data are included as an appendix. The project is completed collaboratively, with students dividing tasks and contributing jointly to the written report. Group-based assessment reflects the collaborative nature of MT evaluation work in professional settings, where tasks are typically distributed across team members with different areas of responsibility (Kiraly, 2005).

Assessment criteria Projects are assessed on five dimensions: i) academic content, including the quality of the literature review and theoretical grounding; ii) technical soundness, covering the suitability and correct application of evaluation methods and tools; iii) quality of analysis, including the interpretation and critical discussion of results; iv) structure and argumentation; and v) language quality. Together these criteria reward both technical competence and the ability to communicate findings clearly and critically in an academic register.

Scenario The simulated scenario changes each year, providing students with a realistic context that motivates the evaluation choices they make. In one iteration of the module, the scenario involved translating institutional health reports from Dutch into English for online publication, a domain that combines technical terminology with a requirement for high output quality. The scenario is designed to raise questions that go beyond metrics: is MT suitable for this content at all? Is the raw output publishable, or are further steps needed?

5 MT usage and customisation trends

How do student practices evolve alongside developments in the broader technology landscape? As LLMs are becoming dominant in life and market, are students adopting them more for translation, abandoning traditional NMT engines? This section examines student preferences in engine selection and customisation strategies, drawing on the submitted project reports from all three years of the module.

Engine	2023	2024	2025
Google Translate	6 (32%)	4 (13%)	6 (17%)
DeepL	5 (26%)	7 (23%)	5 (14%)
ModernMT	4 (21%)	5 (17%)	3 (9%)
NLLB	4 (21%)	1 (3%)	1 (3%)
ChatGPT/OpenAI	-	4 (13%)	6 (17%)
Widn	-	1 (3.%)	2 (6%)
Lara	-	-	3 (9%)
Others	-	2 (7%)	-
Total	19	30	35

Table 1: Distribution of MT engine selections by year

Table 1 shows the distribution of generic MT engine selections across the three years of the study. It should be noted that the available engine pool expanded over time. Widn and Lara, for instance, were only introduced in class from 2024 onwards. This means that the trends reflect both changes in the offering and shifts in student preferences. Google Translate and DeepL remained the most consistently chosen engines, with 6, 4, and 6 selections, and 5, 7, and 5 selections respectively. ModernMT was the next popular selection in 2023 and 2024, but declined in 2025. NLLB, the only open-

source NMT engine in the selection, was chosen in 4 projects in 2023 but only in one in subsequent years. This may be due to a better offering of engines for this high-resource language pair. The most notable trend is the rapid adoption of ChatGPT/OpenAI, which was absent in 2023 but rose to 4 selections in 2024 and 6 in 2025, matching Google Translate as the most selected engine. The initial absence is partly explained by a practical limitation: pasting the entire text in the ChatGPT interface did not preserve sentence boundaries, requiring manual realignment before automatic evaluation. This issue was subsequently resolved by accessing the model via the OpenAI API. Newer engines such as Widn and Lara also gained traction in the last year. These trends suggest that while traditional NMT providers are not being abandoned, students are increasingly experimenting with LLM-based solutions, reflecting the shifts in the professional translation technology market.

Method	2023	2024	2025
Glossary	4 (57%)	3 (50%)	7 (64%)
TM	3 (43%)	2 (33%)	2 (18%)
Prompting	-	1 (17%)	2 (18%)
Total	7	6	11

Table 2: Customisation options selected per year

Table 2 shows the customisation methods selected by students across the three years. Glossary-based customisation was the most popular method throughout, accounting for the majority of selections in all three years. Most providers have the option to directly upload a glossary, which students reported to be the easiest method for customisation. TM-based customisation was the second method preferred with 2 to 3 projects selecting it. Prompting is gradually emerging as a customisation strategy, with students including context information about the use case and/or the glossary entries in the translation prompt. This reflects the growing familiarity of students with LLM-based tools and their affordances for domain adaptation. Nevertheless, traditional customisation methods built into the engine interface, such as glossary upload and TM integration, remain preferred, likely because they are more straightforward to apply and easier to evaluate systematically than a custom prompt whose effect on output quality is harder to isolate.

6 Student evaluation and impressions

Student feedback was collected through anonymous voluntary surveys administered at the end of each iteration. The following reflections draw on three years of feedback and the instructor's own observations across the cohorts.

What worked well Students consistently recognised the relevance of the module to their professional development. The technological dimension was valued precisely because of its currency, with students recognising the importance of gaining experience with MT and AI. The variety of tools explored across the module (commercial, open-source, and LLM-based) was also appreciated. The combination of theory and hands-on practice was frequently highlighted as effective, with students appreciating that concepts introduced in the theoretical part were immediately applied in practical activities. The project-based assessment was particularly well received, with students valuing the experience of working on a realistic scenario with assignments for clients in the real world.

Points for improvement The most consistent criticism across all three years concerned the balance between theory and practice. Students frequently reported feeling overwhelmed by the theoretical content, particularly in the early sessions, and requested more time for hands-on activities. In response, the format was adjusted from the second year onwards: slide presentations were shortened, readings were assigned before class rather than presented in full during sessions, and contact time was reserved for practical work and discussion of questions arising from the readings.

The integration of LLM-based engines introduced a practical challenge around costs. Since students required access to LLM APIs for translation tasks, token usage had to be monitored and sometimes limited to avoid excessive costs. In practice, this was managed by pre-generating translations and sharing them with students, avoiding repeated generation of the same output. While functional, this solution reduced the immediacy of the hands-on experience. This remains an open challenge, particularly as LLM-based tools become more central to the module. Two tools were introduced incrementally and only reached their current form in the most recent iteration: translation-specific LLMs (Widn, Lara) were added gradually in the second and third years

as they became available, and the enhanced post-editing tool with explainable QE signals was piloted in the final iteration. Both additions were positively received but have not yet been evaluated across a full cohort in their current form.

Recommendations Based on three years of experience, several recommendations emerge for instructors designing similar modules. First, limiting in-class theory in favour of flipped classroom elements, for example assigning readings and explanatory materials before sessions, frees contact time for the practical work students find most valuable. Second, managing access to commercial and LLM-based tools requires planning: educational licences, pre-generated outputs, and token budgets should be arranged well in advance and are not always available. This is why investing on open-source tools is a valuable alternative. Finally, students expressed interest in hearing from working professionals: a guest lecture by an MT specialist or language technology consultant would complement the academic content and reinforce the professional relevance of the skills developed.

7 Conclusion

This paper has presented a project-based translation technology module that equips MA Translation students with end-to-end MT evaluation skills, integrating LLMs throughout as translation engines, customisation tools, and sources of explainable quality signals. The module has been taught over three consecutive years, allowing for iterative refinement in response to student feedback and developments in the field.

Analysis of student project reports over three consecutive years reveals a gradual shift in engine preferences, with LLM-based tools gaining ground while traditional engines remain in use. In terms of customisation, glossary-based methods continue to dominate due to their ease of implementation and perceived efficiency, though prompting is emerging as an alternative strategy.

Some limitations should be acknowledged. The module has been taught at a single institution with a single language pair (English-Dutch), and no formal measurement of learning outcomes has been conducted beyond student feedback. The evaluation of recently introduced components, such as translation-specific LLMs and the explainable QE tool, remains incomplete. Formal evaluation of the module, including self-efficacy measures and more detailed qualitative and quantitative feedback, will be conducted in future iterations. Furthermore, the module relies heavily on commercial engines and tools, a deliberate choice to maximise relevance to real-world professional practice, made possible through institutional resources and academic agreements with technology providers. However, it is recognised that not all institutions have access to such resources. A related challenge concerns instructor expertise: setting up open-source engines, hosting evaluation tools, or running models such as xCOMET requires technical knowledge that many translation technology instructors may not have. While commercial alternatives for enhanced post-editing exist, e.g. platforms such as Phrase and Bureauworks offer comparable functionality, the development of accessible, open-source post-editing and QE tools remains important for broadening access to this type of instruction (Lefer et al., 2024; Mutal et al., 2020).

More broadly, the challenge facing translator training programmes is not simply to incorporate specific tools, but to develop frameworks flexible enough to absorb continuous technological change without requiring wholesale curriculum redesign. The module presented here attempts to address this by grounding instruction in a stable evaluation pipeline that can accommodate new engines, metrics, and LLM-based features as they emerge. Future work includes the development of an online component to deliver theoretical content outside contact hours, freeing in-class time for the hands-on practice and critical discussion that students find most valuable.

References

Ayvazyan, Nune, Yu Hao, and Anthony Pym. 2024. Things to do in the translation class when technologies change: The case of generative AI. In Peng, Yuhong, Huihui Huang, and Defeng Li, editors, *New Advances in Translation Technology: Applications and Pedagogy*, pages 219–238. Springer, Singapore.

Bowker, Lynne and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Group Publishing, Bingley.

Costa-Jussà, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Doherty, Stephen and Dorothy Kenny. 2014. The design and evaluation of a statistical machine translation syllabus for translation students. *The Interpreter and Translator Trainer*, 8(2):295–315.

Doherty, Stephen, Dorothy Kenny, and Andy Way. 2012. Taking statistical machine translation to the student translator. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Commercial MT User Program*, San Diego. Association for Machine Translation in the Americas.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In Pareja-Lora, Antonio, Maria Liakata, and Stefanie Dipper, editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia. Association for Computational Linguistics.

Guerberof Arenas, Ana and Joss Moorkens. 2019. Machine translation and post-editing training as part of a master's programme. *The Journal of Specialized Translation*, 31:217–238.

Guerreiro, Nuno M, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Hackenbuchner, Janiça and Ralph Krüger. 2023. DataLitMT–Teaching data literacy in the context of machine translation literacy. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 285–293, Tampere. European Association for Machine Translation.

Hao, Yu and Anthony Pym. 2023. Choosing effective teaching methods for translation technology classrooms: Teachers' perspectives. *Forum*, 21(2):190–212.

Kenny, Dorothy and Stephen Doherty. 2014. Statistical machine translation in the translation curriculum: Overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer*, 8(2):276–294.

Kenny, Dorothy, editor. 2022. *Machine Translation for Everyone*. Language Science Press, Berlin.

Kiraly, Don. 2005. Project-based learning: A case for situated translation. *Meta*, 50(4):1098–1111.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge University Press, Cambridge.

Koponen, Maarit. 2015. How to teach machine translation post-editing? Experiences from a post-editing course. In *Proceedings of the 4th Workshop on Post-editing Technology and Practice*.

Krüger, Ralph. 2022. Using Jupyter notebooks as didactic instruments in translation technology teaching. *The Interpreter and Translator Trainer*, 16(4):503–523.

Lefer, Marie-Aude, Romane Bodart, Justine Piette, and Adam Obrusník. 2024. MTPE quality evaluation in translator education: the postedit.me app. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Mikel Forcada, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 23–24, Sheffield. European Association for Machine Translation.

Li, Defeng, Chunling Zhang, and Yuanjian He. 2015. Project-based learning in teaching translation: Students' perceptions. *The Interpreter and Translator Trainer*, 9(1):1–19.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):455–463.

Macken, Lieve, Bram Vanroy, and Arda Tezcan. 2023. Adapting machine translation education to the neural era: A case study of MT quality assessment. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 305–314, Tampere. European Association for Machine Translation.

Mellinger, Christopher D. 2017. Translators and machine translation: Knowledge and skills gaps in translator pedagogy. *The Interpreter and Translator Trainer*, 11(4):280–293.

Mitchell-Schuitevoerder, Rosemary. 2020. *A project-based approach to translation technology*. Routledge, London/New York.

Moorkens, Joss. 2018. What to expect from neural machine translation: A practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4):375–387.

Moorkens, Joss. 2022. Ethics and machine translation. In Kenny, Dorothy, editor, *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence*, pages 121–140. Language Science Press, Berlin.

Mutal, Jonathan, Pierrette Bouillon, Perrine Schumacher, and Johanna Gerlach. 2020. COPECO: a collaborative post-editing corpus in pedagogical context. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 61–78, Virtual. Association for Machine Translation in the Americas.

O'Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. In *Proceedings of the 6th EAMT workshop: Teaching Machine Translation*, Manchester. European Association for Machine Translation.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon. Association for Computational Linguistics.

Pym, Anthony and Yu Hao. 2024. *How to Augment Language Skills: Generative AI and Machine Translation in Language Learning and Translator Training*. Routledge, London/New York.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts. Association for Machine Translation in the Americas.

Tiedemann, Jörg. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007*, pages 237–248. Benjamins, Amsterdam.

Treviso, Marcos V, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and André FT Martins. 2024. xTower: A multilingual LLM for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami. Association for Computational Linguistics.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere. European Association for Machine Translation.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR 2020)*.