A comparison of translation performance between DeepL and Supertext

Alex Flückiger, Chantal Amrhein, Tim Graf, Frédéric Odermatt, Martin Pömsl, Philippe Schläpfer, Florian Schottmann, Samuel Läubli

Supertext

{firstname.lastname}@supertext.com

Abstract

As strong machine translation (MT) systems are increasingly based on large language models (LLMs), reliable quality benchmarking requires methods that capture their ability to leverage extended context. This study compares two commercial MT systems – DeepL and Supertext – by assessing their performance on unsegmented texts. We evaluate translation quality across four language directions with professional translators assessing segments with full document-level context. While segment-level assessments indicate no strong preference between the systems in most cases, document-level analysis reveals a preference for Supertext in three out of four language directions, suggesting superior consistency across longer texts. We advocate for more context-sensitive evaluation methodologies to ensure that MT quality assessments reflect real-world usability.[1]

1 Introduction

After the transition from statistical to neural modelling roughly a decade ago (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015), the field of MT is undergoing another paradigm shift towards leveraging LLMs (Xu et al., 2024; Wu et al., 2024b; Kocmi et al., 2024). LLM-based translation offers the potential for significantly improved translation quality, especially with respect to consistent translation of documents. Unlike neural machine translation (NMT) systems, which typically process documents as isolated sentences or paragraphs (Post and Junczys-Dowmunt, 2023), many LLMs operate with context windows that can span thousands of words, allowing them to maintain consistency throughout a document – for instance, by ensuring that a word's translation in the final sentence matches its previous forms (Wu et al., 2024b).

In the most recent shared task at the Conference on Machine Translation (WMT24) that focuses on evaluating the state of the art in general-domain translation quality, the majority of the 28 system submissions were already based on LLMs (Kocmi et al., 2024). Although without statistical significance and for the language direction English to German only, one system even outranked the human reference translations as evaluated by professional human annotators.

Despite this impressive achievement, findings of human-machine parity should be approached with caution. Similar claims already emerged with pre-LLM technology (Hassan et al., 2018), yet have subsequently been refuted due to limitations in the evaluation design focusing on single segments in isolation (Läubli et al., 2018; Toral et al., 2018; Freitag et al., 2021). The WMT24 shared task also highlights that evaluations based on automatic metrics (rather than human evaluation) can lead to wrong conclusions when comparing strong MT systems (Kocmi et al., 2024).

However, these insights are often overlooked in evaluations of commercial MT systems. For example, Intento's The State of Machine Translation 2024 report,[2] which assesses 52 providers across 11 language pairs, serves as a valuable resource for potential users in real-world settings, but its benchmarking methodology relies on automatic scoring of sentence-level data, and the authors acknowledge that 'you may need a human linguist' to ensure greater reliability.

In this paper, we evaluate two commercial translation systems (Section 2) under conditions that allow for leveraging the full-text capabilities of

[1] We release all evaluation data and scripts for further analysis and reproduction at https://github.com/supertext/evaluation_deepl_supertext

[2] https://inten.to/machine-translation-report-2024

Proceedings of Machine Translation Summit XX Volume 2, pages 52–57, June 23-27, 2025

LLMs. The segmentation of the source text is handled by the translation systems alone without any prior splitting (Section 3), and the resulting translations are rated by professional translators considering the full document context (Section 4). We find that while both systems translate a similar number of segments better than the other, the difference is more pronounced on the document level (Section 5), which we attribute to differences in how much context the systems consider during translation (Section 6). Our findings suggest that the adoption of LLMs creates opportunities for smaller players to challenge dominant industry leaders (Section 7).

2 Systems

We compare the free online offering of two commercial MT providers:

DeepL DeepL[3] is a widely used MT provider boasting 'unrivalled translations that set the standard'.[4] In the latest Intento report, it scores best among nine 'real-time engines' and, together with GPT-4, is found to 'consistently outperform other models'.[2] Due to its closed source, the technology behind DeepL's translation system is not publicly known.

Supertext Supertext[5] builds on an open, general-purpose LLM that has been specialised for the task of translation with proprietary methods and data. While the system can be adapted to specific domains, we use the freely available generic version. For the purpose of the evaluation described in this paper, we use both systems with default settings. For example, we do not specify politeness (formal/informal) although supported by both systems in some language combinations.

While both DeepL and Supertext provide target language variants for English (British and American), Supertext also provides target language variants for German (Austria, Germany, Switzerland), French (France, Switzerland), and Italian (Italy, Switzerland). As our use case is machine translation for people in Switzerland, we use the Swiss target language variant whenever available (Section 3.2).[6]

[3] https://www.deepl.com

[4] https://www.deepl.com/en/quality, see also Appendix A.

[5] https://www.supertext.com

[6] Compared to English variants, the Swiss variants of other languages differ minimally.

Language Direction	Texts	Segments	Words
de → en-GB	20	281	3336
de → fr-CH	20	276	3336
de → it-CH	20	265	3336
en → de-CH	20	211	3483
Total	80	1033	13491

Table 1: Evaluation data by language direction.

3 Data

3.1 Source Texts

We collect 20 texts each in two source languages: English (en) and German (de). All texts stem from news websites: New York Times[7] for English and Neue Zürcher Zeitung[8] for German, respectively. We select 10 FAQ pages and 10 recent news articles in the economy section from each website. Notably, these texts are only available in a single language; they are unlikely to be contained in the training data of either system we evaluate. To balance the distribution of text lengths, we trim the end of some texts by omitting their final paragraphs.

3.2 Target Texts

We create translations in four language directions (Table 1) directly in the respective online translation interface of each system as a regular user would.[9] We do not modify the texts before translation and paste them in their original formatting, including newlines. The translation systems may segment the text into smaller chunks internally.

After translation, we manually split and align the source texts and translations into sentences. If one of the systems merges two or more sentences into a single sentence, we ensure that the same content is merged for the other system, such that the raters are presented with parallel segments. Table 1 shows the resulting number of segments per language pair. The texts per source language are identical, differing only in how they were manually segmented for the A/B test after translation. Across the language pairs, the median number of segments per document is 13.

[7] https://www.nytimes.com

[8] https://www.nzz.ch

[9] All translations were produced on 27 January 2025.

4 Evaluation Setup

We conduct a blind A/B test in which professional translators rate DeepL and Supertext outputs with full document-level context.

4.1 Raters

We enrol 8 professional translators with experience in evaluating machine translation output, 1 to 3 per language direction. All raters have between 2 and 19 years of professional experience (average=8.6 years) in the language combination they are assigned to and are native in the respective target language.

4.2 Materials

We arrange all segments of a source document with their corresponding translations by both systems in a spreadsheet. The segments are presented in original document order, including formatting such as newlines, such that raters see the full source text and both translations side-by-side. We randomly assign the system outputs to columns labelled Translation A and Translation B for each text such that raters do not have any information about which translation stems from which system (a blind A/B test setting). System assignments are kept consistent within a text such that the document context remains natural.

4.3 Procedure

Documents are assigned to single raters. For each segment in each document, the assigned rater is asked to choose whether Translation A is better, Translation B is better, or whether both translations are of equal quality.

Our instructions explicitly state that 'equal' can mean that two translations are equally good or equally bad. Moreover, the raters were asked to focus on the content rather than punctuation to avoid that the results get biased because of specifics of a language variant.

5 Evaluation Results

Segment-level and text-level preference ratings are shown in Figures 1 and 2, respectively.

5.1 Segment-level

Across all language pairs, 9.5% of the segments generated by DeepL and Supertext are identical. The overlap is highest in de → en-GB, particularly in the FAQ texts where 26.1% of the segments were translated identically.

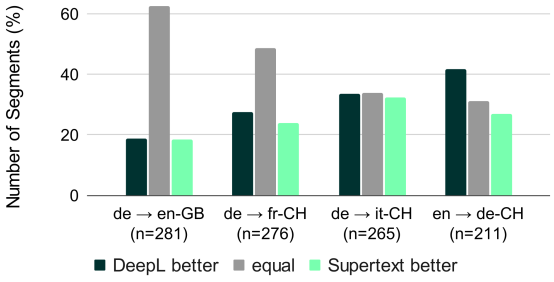

Figure 1: Segment-level ratings.

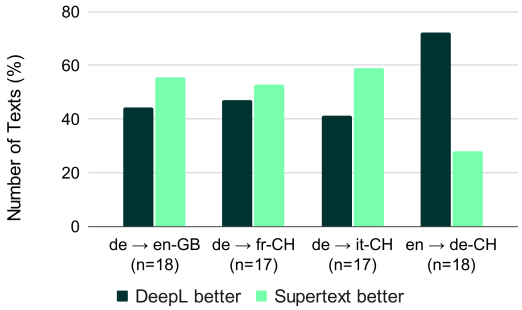

Figure 2: Aggregated segment-level ratings per text. Texts with the same number of preferred segments for both systems are excluded.

Participants rate most segments as equal in terms of translation quality in three out of four language directions. While the number of segments where one system is preferred over the other is similar for DeepL and Supertext in these language directions, raters show a preference for DeepL in en → de-CH (88 DeepL, 66 equal, 57 Supertext).

5.2 Document-level

We derive document-level preferences by aggregating the segment-level ratings of each evaluated document. For example, a text is counted as 'DeepL better' if the rater preferred DeepL's translations for more segments than Supertext's translations in that document.

In contrast to the pooled segment-level ratings (Section 5.1), raters show a preference for documents translated by Supertext in three out of four language directions, most notably in de → it-CH (7 DeepL, 3 equal, 10 Supertext). In en → de-CH, however, raters show a clear preference for doc-

uments translated by DeepL (13 DeepL, 2 equal, 5 Supertext).

6 Discussion

Our evaluation highlights that conclusions drawn from MT quality assessments may vary significantly depending on the unit of measurement. While raters in our study preferred a similar number of translated segments by DeepL and Supertext overall, the difference becomes more pronounced at the document level. This discrepancy suggests that segment-level assessments alone may not fully capture translation quality as perceived in real-world usage, where coherence and consistency across entire documents play a critical role.

Notably, while segment-level ratings indicate no strong preference between the two systems in most language directions, document-level aggregation reveals a more distinct pattern. Raters favour Supertext's translations at the document level in three out of four language directions, with the most pronounced difference observed in de → it-CH. This suggests that Supertext may provide better consistency or fluency across longer texts in these language directions. While we have yet to conduct a systematic qualitative comparison of system outputs, we find texts where the same word is translated differently by DeepL and consistently by Supertext across paragraphs. An example is shown in Table 2, where DeepL translates the German word *Startseite* as either *start page*, *home page*, or *Home page*.

In contrast, for en → de-CH, raters show a clear preference for DeepL at both the segment and document levels, indicating a potential strength of DeepL in handling this specific language combination. Our preliminary analysis is inconclusive at this point, but the Supertext outputs seem to contain a higher number of within-sentence errors such as wrong choices for individual words or omissions. Another hypothesis is that Supertext, which supports three different German target language variants, may introduce inconsistencies by mixing region-specific elements in translation outputs.

7 Conclusion

Our study highlights the growing significance of document-level evaluations in MT quality benchmarking, especially as LLM-based systems leverage broader context windows to enhance translation consistency. While segment-level assessments suggest no clear preference between DeepL and Supertext in most of the language directions we examined, document-level aggregation reveals notable differences. Supertext is preferred in three out of four language pairs, where its translations exhibit greater consistency. In contrast, en → de-CH shows a clear preference for DeepL, possibly due to fewer within-sentence errors or differences in regional language handling.

As LLM-based MT systems continue to evolve, future studies should further investigate the impact of context length on commercial MT benchmarking campaigns. Insights into how different systems leverage context and resolve ambiguities will be essential for advancing evaluation methodologies and ensuring that translation systems meet real-world user expectations.

Limitations

While A/B tests are commonly used for comparing two systems and a reliable basis for incrementally improving MT systems (Tang et al., 2010; Wu et al., 2024a), they provide no insight into the severity of errors within a translation or across different systems compared to MQM ratings (Freitag et al., 2021). Absent the use of more time-intensive evaluation frameworks, such limitations persist irrespective of whether preferences are aggregated at the system level or pooled by document.

During real-world usage, some mistakes may be harder to spot than others when not being shown contrastively against an alternative translation. Similarly, the preference in an A/B test may not correlate with the effort needed for post-editing the translation. To address these questions, we plan to extend our evaluation efforts.

The evaluation was carried out by professional translators working for Supertext. Since the A/B assignments were randomized and anonymized, we do not assume any bias. Additionally, in the interest of transparency, we publicly share the complete dataset, including the source text, translations from each system, and the corresponding ratings.

Finally, the scope of this study is not exhaustive but is limited to a subset of language pairs, two domains, and a limited number of documents. Yet, we are providing details that go beyond what DeepL is sharing publicly on their website.[4]

SID	Source text (de)	DeepL (en-GB)	Supertext (en-GB)
1	Wie kann ich die NZZ als Startseite festlegen?	How can I set the NZZ as my start page?	**How can I set NZZ as my homepage?**
2	Öffnen Sie Ihren Browser:	Open your browser:	Open your browser:
3	- Stellen Sie sicher, dass der Browser geöffnet ist, den Sie verwenden möchten (z.B. Google Chrome, Mozilla Firefox, Microsoft Edge, Safari).	- Make sure the browser you want to use is open (e.g. Google Chrome, Mozilla Firefox, Microsoft Edge, Safari).	- Make sure the browser you want to use is open (e.g., Google Chrome, Mozilla Firefox, Microsoft Edge, Safari).
4	Gehen Sie zu den Einstellungen:	Go to the settings:	Go to settings:
5	- In den meisten Browsern finden Sie die Einstellungen oder Optionen im Menü oben rechts, oft dargestellt durch drei Punkte (?) oder Striche (≡).	- In most browsers, you will find the settings or options in the menu at the top right, often represented by three dots (?) or dashes (≡).	- In most browsers, you'll find the settings or options in the menu in the top right corner, often represented by three dots (?) or lines (≡).
6	- Klicken Sie auf diese Punkte oder Striche, um das Menü zu öffnen, und wählen Sie dann Einstellungen.	- Click on these dots or dashes to open the menu and then select Settings.	- Click on these dots or lines to open the menu, and then select Settings.
7	Finden Sie die Option für die Startseite:	Find the option for the home page:	Find the option for the homepage:
8	- In den Einstellungen gibt es einen Abschnitt für die Startseite oder Beim Start.	- In Settings, there is a section for Home page or On startup.	**- In the settings, there is a section for the homepage or start page.**
9	- Suchen Sie nach einer Option wie „Startseite festlegen“ oder „Beim Start“ (je nach Browser kann der genaue Wortlaut unterschiedlich sein).	- Look for an option such as 'Set home page' or 'On startup' (the exact wording may vary depending on your browser).	**- Look for an option like "Set homepage" or "Start page" (the exact wording may vary depending on the browser).**
10	NZZ.de als Startseite festlegen:	Set NZZ.de as your start page:	**Set NZZ.de as your homepage:**
11	- Wenn Sie die Option „Bestimmte Seite öffnen“ sehen, wählen Sie diese aus.	- If you see the option 'Open specific page', select it.	- If you see an option called "Open a specific page," select it.
12	- Geben Sie in das Eingabefeld https://www.nzz.de ein.	- Enter https://www.nzz.de in the input field.	- Enter https://www.nzz.de in the input field.
13	Einstellungen speichern:	Save settings:	Save settings:
14	- Die meisten Browser speichern die Änderung automatisch, sobald Sie das Feld verlassen.	- Most browsers save the change automatically as soon as you exit the field.	- Most browsers will save the change automatically as soon as you leave the field.
15	Sie können das Einstellungsfenster nun schließen.	You can now close the settings window.	You can now close the settings window.

Table 2: Example of a rated de → en-GB document. Better-rated translations are highlighted in bold; segments without any bold translation were rated as equal. System names are not shown during evaluation (Section 4).

Acknowledgments

We thank all the professional translators involved for their support with this evaluation.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, San Diego, CA.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 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.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of EMNLP*, pages 1700–1709, Seattle, WA.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *ArXiv*, abs/2304.12959.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS*, pages 3104–3112, Montreal, Canada.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: more, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 17–26, New York, NY, USA. Association for Computing Machinery.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Guojun Wu, Shay B Cohen, and Rico Sennrich. 2024a. Evaluating automatic metrics with incremental machine translation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2994–3005, Miami, Florida, USA. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024b. Adapting large language models for document-level machine translation. *Preprint*, arXiv:2401.06468.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

A Appendix

For the sake of persistency, we share the archived link as well: `https://web.archive.org/web/20250215011944/https://www.deepl.com/en/quality`